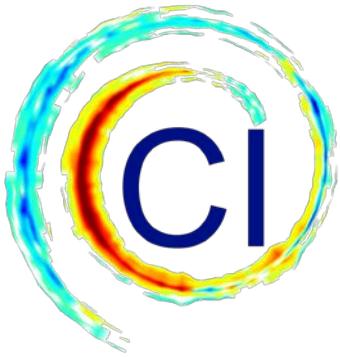


Proceedings of the 9th International Workshop on **Climate Informatics: CI 2019**

Volume editors
Julien Brajard
Anastase Charantonis
Chen Chen
Jakob Runge

Series editors
Imme Ebert-Uphoff
Claire Monteleoni
Doug Nychka
Eniko Szekely

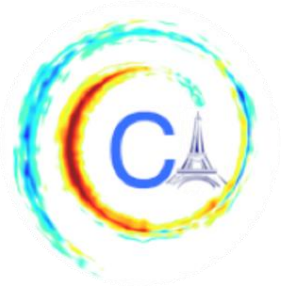
NCAR Technical Notes
NCAR/TN-561+PROC

National Center for Atmospheric Research
P. O. Box 3000
Boulder, Colorado
80307-3000
www.ucar.edu

NCAR IS SPONSORED BY THE NSF

How to Cite this Document:

Brajard, J., Charantonis, A., Chen, C., & Runge, J. (Eds.). (2019). Proceedings of the 9th International Workshop on Climate Informatics: CI 2019 (No. NCAR/TN-561+PROC). doi:10.5065/y82j-f154



The CI2019 logo on the cover page is courtesy of Michael Tippett modified by Julien Brajard. Colors show deviations of sea-surface temperatures from their climatological values in the equatorial Pacific from January 1997 to April 2000 with time going counter-clockwise.

Information about future workshops and other CI news can be found on our website, http://www.climateinformatics.org

The website of the 2019 workshop is https://sites.google.com/view/climateinformatics2019

To be added to the workshop mailing list, please send an email to climate.informatics.workshop@gmail.com.

9th International Workshop on Climate Informatics

Table of Contents

Foreword by the CI 2019 Workshop Chairs

The Climate Informatics (CI) workshop series was founded in 2011 by machine learning scientist Claire Monteleoni and NASA director Gavin A. Schmidt. In its 9th year now, the workshops aim to stimulate the discussion of new ideas, foster new collaborations, grow the climate informatics community, and thus accelerate discovery across disciplinary boundaries between researchers from statistics, mathematics, machine learning and data mining and researchers in climate science. The format of the workshop seeks to overcome cross-disciplinary language barriers and to emphasize communication between participants by featuring tutorials, hackathons, invited talks, panel discussions, posters and break-out sessions.

The 2019 Climate informatics Workshop was held in Paris France, hosted by the Ecole Normale Supérieure (ENS), between the 2nd and 4th of October 2019. The workshop had 166 attendees with 81 submitted papers, of which 68 were accepted for poster presentations. The participants came from 23 countries covering all continents (except Antarctica). While conference proceedings are common in many computer science and informatics disciplines, our aim was to encourage more participation from both climate science and statistics where journals are a more common venue of publication. We are pleased that these proceedings represent 54 of these papers.

The Best Paper award was given to Jussi Leinonen, Tianle Yuan and Alexis Berne for their paper "Generative Adversarial Network for Climate Data Field Generation" and the Best Young Scientist Paper award to Bram Kraaijeveld, Sem Vijverberg and Dim Coumou for the paper titled "Forecasting Eastern United States Heatwaves: Combining Climate Simulations and Machine Learning".

The first day of the workshop was an optional Hackathon led by Julien Le Sommer, Anna Sommer and Redouane Lguensat. The hackathon platform https://codalab.org/ was used to host a team-based prediction challenge, wherein groups of attendees were tasked with making a data-driven approach to predict CO2 fluxes between the oceans and the atmosphere. The results of this hackathon will be published in a forthcoming collaborative publication in 2020.

The main workshop (October 3rd to 4th) featured six invited speakers, two poster sessions of submitted papers and short spotlight talks by several early-career scientists whose papers were judged to be outstanding.

Invited speakers covered many topics from across the spectrum of climate informatics: Pascale Braconnot opened the workshop with an overview of long-term climate variability followed by Michael Ghil as a pioneer of nonlinear dynamics in climate sciences. The afternoon of the first day moved from a causal discovery lecture by early-career scientist Marlene Kretschmer to novel hybrid machine learning approaches for climate modelling by Pierre Gentine. On the second day, Sebastian Engelke introduced extreme value theory and the workshop was closed by Turing award winner Yoshua Bengio with a talk on AI and the Climate Crisis.

We would like to thank many people on the organizing committee, at the Institute Pierre Simon Laplace (IPSL) and at École Normale Supérieure (ENS), whose hard work was crucial for the success of the workshop.

First and foremost, we would like to thank the steering committee for years of leadership and funding efforts that have ensured continuity of the workshop: Imme Ebert-Uphoff, Claire Monteleoni, Doug Nychka and Eniko Szekely.

The chairs of the program committee, Chen Chen and Anastase Charantonis, did a fantastic job of not only orchestrating the paper reviews but also organizing these proceedings and countless other tasks.

Communications chair Soukayna Mouatadid did a wonderful job advertising the workshop and communicating with attendees and registrants.

Travel chair Sophie Giffard-Roisin was instrumental in securing and coordinating travel funding for the young researchers and we were able to offer 10 travel awards this year.

We thank Julien Le Sommer, Anna Sommer and Redouane Lguensat for their work on the Hackathon.

The 58 members of the program committee provided thorough and rapid reviews of the submitted papers, and we thank them for volunteering so much of their time to do so.

We want to thank our local chair Olivier Talagrand for being instrumental in securing the prestigious location and help up organise locally. We also deeply thank Christina Auguste-Charlery and Catherine Michaut at IPSL as well as Hélène Rouby and Ouissem Trabelsi at ENS for their logistical help and administrative support.

We thank Dorit Hammerling for being our science contact at NCAR and allowing for the publication of this NCAR proceedings.

Finally, we would like to thank our sponsors. Firstly, ENS provided the salle Jean Jaurès, which often hosts major cultural and scientific events. IPSL's administrative support was indispensable and travel grants were sponsored by the Michael Stifel Center Jena, NCAR, and the Artificial Intelligence Journal. Finally, Microsoft generously sponsored the hackathon and workshop and also provided prizes for hackathon winners.

The move to hold the conference in Europe was, we feel, an important step since CI 2019 was the first time the event was hosted outside the United States, emphasizing the growth of the international community.

The other important takeaway is, based on the feedback we received, that the community has grown to the point that the length and scope of the Climate Informatics meetings warrant its transition to a longer conference.

Next year's CI2020 will be in Oxford, hosted in the historical university's examination schools.

We are excited about the future of the CI workshop and CI community and we look forward to many coming years.

CI2019 Workshop Co-Chairs,
Jakob Runge and Julien Brajard

Organizing committee

Workshop chairs

Jakob Runge (German Aerospace Center, Jena)
Julien Brajard (Nansen Center, Bergen and Sorbonne Université, Paris)

Local chair

Olivier Talagrand (École Normale Supérieure, Paris)

Program Committee chairs

Anastase Charantonis (École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise, Evry)
Chen Chen (University of Chicago)

Communications chair

Soukayna Mouatadid (University of Toronto)

Travel grant chair

Sophie Giffard (University of Colorado)

Hackathon chairs

Julien Le Sommer (IGE)
Redouane Lguensat (IGE, CNES)
Anna Denvil-Sommer (LSCE)

Steering committee

Imme Ebert-Uphoff (Colorado State University)
Claire Monteleoni (University of Colorado, Boulder)
Doug Nychka (Colorado School of Mines, Golden)
Eniko Szekely (EPFL Lausanne / ETH Zurich)

Program Committee Members

Conrad Albrecht, TJ Watson Research Center, IBM Research

Ibrahim Ayed, Sorbonne Université, LIP6, Thales

Venkatramani Balaji, Princeton University / IPSL

Laurent Bertino, Nansen Center

Marc Bocquet, Ecole des Ponts ParisTech

Julien Brajard, Université Pierre et Marie Curie, LOCEAN

Gustau Camps-Valls, Image Processing Laboratory (IPL), Universitat de Valencia

Alberto Carrassi, NERSC and Un. of Bergen

Won Chang, University of Cincinnati

Anastase - Alexander Charantonis, École Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise, Evry

Chen Chen, University of Chicago

Dan Cooley, Colorado State University

Raphael de Fondeville, Ecole Polytechnique Fédérale de Lausanne

Anna Denvil-Sommer, LSCE/CEA

Joachim Denzler, University Jena

Mohamed Djallel Dilmi, LATMOS/CNRS/UVSQ/Université Paris-Saclay

Andre Richard Erler, University of Toronto

Ronan Fablet, Telecom Institute; Telecom Bretagne

David Gagne, National Center for Atmospheric Research

Sophie Giffard-Roisin, CNRS

Lukas Gudmundsson, ETH Zurich

Whitney Huang, University of Victoria

Nikola Jajcay, Institute of Computer Science, Academy of Sciences of the Czech Republic

Cédric Jamet, LOG/ULCO

Karthik Kashinath, Lawrence Berkeley Lab

Myriam Khodri, LOCEAN/IPSL

Vipin Kumar, University of Minnesota

Mikael Kuusela, SAMSI / UNC Chapel Hill

Redouane Lguensat, Université Grenoble Alpes

Bo Li, University of Illinois at Urbana-Champaign

Stefan Liess, University of Minnesota

Karen McKinnon, University of California, Los Angeles

Carlos Mejia, IPSL/Locean/CNRS

Claire Monteleoni, University of Colorado Boulder

Philippe Naveau, CNRS-LSCE-IPSL

Peer Nowack, Imperial College London

Nikunj Oza, NASA

Andrew Poppick, Carleton College

Jakob Runge, German Aerospace Center

Moumita Saha, University of Colorado Boulder

Francine Schevenhoven, University of Bergen

Olivier Schwander, Université de Genève

Sebastian Sippel, ETH Zurich

Joanna Slawinska, University of Wisconsin-Milwaukee

Jebb Stewart, NOAA/ESRL

Eniko Szekely, Swiss Data Science Center, EPFL/ETH

Pierre Tandeo, IMT-Atlantique

Sylvie Thiria, Université Pierre et Marie Curie

Peter Jan van Leeuwen, Colorado state university

Michele Volpi, Swiss Data Science Center

Mathieu Vrac, LSCE/IPSL/CNRS

Rebecca Willett, University of Chicago

Acknowledgements of Sponsors

We gratefully acknowledge the generous contributions of the following sponsors who have helped make CI 2019 possible:

Workshop Location

The workshop was held for the first time outside the United States in Paris at École Normale Supérieure, located directly in the center of Paris in the former premises of the Cinémathèque de Paris.

Workshop Group Photo

Workshop Agenda

Wednesday 2nd October 2019 — Hackathon day

09:30 - 10:00 : Welcome coffee

10:00 - 10:45 : Schedule of the day and introduction to the challenge

10:45 - 11:00 : Team finalization - technical question

11:00 - 12:30 : Teamwork 1

12:30 - 13:30 : Snack

13:30 - 15:30 : Teamwork 2

15:30 - 16:00 : Pooling of progress

16:00 - 16:30 : Coffee break

16:30 - 18:00 : Teamwork 3

18:00 - 18:30 : Conclusions - next steps

Thursday 3rd October 2019 — Workshop Day 1

08:30 - 09:00 : Registration and welcome coffee

09:00 - 09:15 : Welcome and intro

09:15 - 10:15 : Invited talk - Pascale Braconnot: Investigating long term climate variability and changes with Earth System models

10:15 - 11:15 : Invited talk - Michael Ghil: Low-Frequency Climate Variability: Markov Chains and Nonlinear Oscillations

11:15 - 13:15 : Lunch break

13:15 - 13:30 : Sponsor talk - Jennifer Marsman: Microsoft AI for Earth talk

13:30 - 14:30 : Invited talk - Marlene Kretschmer: Assessing teleconnection pathways with causal inference techniques

14:30 - 15:30 : Poster session #1 - Coffee served (see here the repartition of the posters)

15:30 - 16:30 : Invited talk - Pierre Gentine: Hybrid modeling: best of both worlds?

16:30 - 18:30 : Free time + optional labeling session for the ClimateNet project

18:30 - 21:00 : Reception

Friday 4th October 2019 Workshop Day 2

08:30 - 08:45 : Registration

08:45 - 09:45 : Highlight talks

08:45 - 09:05 : Eniko Szekely - A direct approach to detection and attribution of climate change

09:05 - 09:25 : Ibrahim Ayed - Learning the hidden dynamics of ocean temperature with Neural Networks

09:25 - 09:45 : Jussi Leinonen - Generative Adversarial Network for Climate Data Field Generation

09:45 - 10:15 : Coffee break

10:15 - 11:15 : Invited talk - Sebastian Engelke: Graphical models and causality for extreme events

11:15 - 13:15 : Lunch break

13:15 - 14:15 : Invited talk - Yoshua Bengio: AI and the Climate Crisis

14:15 - 15:15 : Poster session #2 - coffee served (see here the repartition of the posters)

15:15 - 16:15 : Highlight talks

15:15 - 15:35 : Christian Reimers - Using causal inference to globally understand black box predictors beyond saliency maps

15:35 - 15:55 : Pierrick Bruneau - Computing flood probabilities using Twitter: application to the Houston urban area during Harvey

15:55 - 16:15 : Jorge Baño-Medina - The importance of inductive bias in convolutional models for statistical downscaling

16:15 - 16:35 : Hackathon feedbacks

16:35 - 16:40 : Conclusions and final announcements, group photo

18:00 - 18:30 : Meeting point for the boat tour

18:30 - 19:45 : Boat tour

Invited talks

Pascale Braconnot
LSCE-IPSL, unite mixte CEA-CNRS-UVSQ, Université Paris-Saclay, France

Investigating long term climate variability and changes with Earth System models

Climate models, called Earth System Models, coupling ocean, atmosphere, land surface and sea-ice components though the energy, the water and biogeochemical cycles, have become key resources to understand how the climate works and evolved in response to natural or anthropogenic forcing. The sciences questions to be addressed now are not limited to the mean climate changes. They require being able to explore climate trends, variability or extremes at the global or regional scales, as well as the interactions between climate and the environment. In this presentation I will first provide an overview of some of the questions and the need for simulations with increased model complexity, improved resolution, longer integration and with several members. Then, using Holocene snap shot or transient (the last 6000 years) climate simulations I will illustrate some of the challenges behind long term simulations designed to understand climate feedbacks, the linkages between changes in the climate mean state and variability, or the ability to represent climate variations outside the modern range. This will also include some thoughts on model development and tuning, and the use of new methodologies, based on entropy and graph theory to guide the analyses.

Yoshua Bengio
Mila, Department of Computer Science and Operations Research, Université de Montréal

AI and the Climate Crisis

AI is moving out of universities and into society, which gives a new social responsibility to AI researchers. Climate change is the biggest crisis that humanity is currently facing. Machine learning is not guaranteed to help tackle this crisis – but it can help. From optimizing energy forecasting to synthesizing new molecules for batteries and improving crisis response, there is a plethora of ways in which ML can be used for both mitigation of and adaptation to the climate crisis. The presentation will discuss in more detail climate-related projects at Mila and elsewhere, including work on synthesizing new materials and on visualizing the effects of climate change.

Pierre Gentine
Department of Earth & Environmental Engineering, Earth Institute, Data Science Institute, Columbia University

Hybrid modeling: best of both worlds?

In recent years, we have witnessed an explosion in the applications of machine learning, especially for environmental problems. Yet for broader utilization, those algorithms may need to respect exactly some physical constraints such as the conservation of mass and energy. In addition, environmental applications (e.g. drought impact) are typically focusing on extremes and basically on out-of-sample generalization. This can be a problem for typical algorithms, which typically interpolate very well. I will here show how a hybridization of machine learning algorithms, imposing physical constraints within them, can help tackle those different issues and offer a promising avenue for environmental applications and process understanding.

Marlene Kretschmer
Potsdam Institute for Climate Impact Research, University of Reading

Assessing teleconnection pathways with causal inference techniques

Teleconnections refer to recurrent large-scale pressure patterns with low-frequency variability, connecting far-away geographical regions. They reflect modifications of atmospheric circulation affecting e.g. the position of the jet stream, storm track intensity or Monsoon strength and thus have a strong impact on our weather. However, extracting the physically relevant teleconnection pathways from observation or model data remains challenging. One major issue is to separate the signal from the noise given large internal atmospheric variability. This is compounded by varying dimensions in space and time and competing effects of different processes. Here, we discuss how novel data-driven causal methods beyond the commonly adopted correlation techniques can help to overcome some of these current limitations. We give an overview of causal inference frameworks and identify promising application cases common in climate science.

Michael Ghil
Ecole Normale Supérieure, Paris, and University of California, Los Angeles

Low-Frequency Climate Variability: Markov Chains and Nonlinear Oscillations

Two complementary ways of describing, understanding and predicting intraseasonal atmospheric variability have been proposed, episodic and oscillatory (Ghil & Robertson, PNAS, 2002). Recent progress in the methodology and results of these two approaches will be presented for subseasonal-to-seasonal (S2S) variability (Ghil et al., in Robertson & Vitart, Eds., Elsevier, 2018).

Sebastian Engelke
Research Center for Statistics, University of Geneva

Graphical models and causality for extreme events

Climate extremes such as heat waves, heavy rainfall or flooding attract an increasing attention by researchers and the public. The accurate statistical assessment of the small occurrence probabilities of such rare scenarios is based on extreme value theory. We will first give a short introduction to this theory and the well-established tools used for univariate data. Many practical questions however concern many variables at the same time. Climate scientists observe an increasing risk of compound events due to a combination of different variables, such as wildfires caused by low precipitation and extreme heat. Similarly, the flood risk of a river catchment depends highly upon the network structure and whether floods at different locations occur simultaneously or independently of each other. Recent advances in extreme value theory therefore concentrate on the dependence between rare events in complex multivariate or spatial systems.

Graphical models have recently been seen to be powerful tools for the analysis of such complex extreme events. We will present several methods to estimate underlying graph structures in a data driven way. This provides sparse and interpretable statistical models even in higher dimensions which can be easily communicated to practitioners. Directed graphical models are also the basis for causal inference. Causality for extremes is a hot topic at the moment with many applications, including the detection of causes for climate extremes. In the framework of linear structural equation models with heavy-tailed noise variables, we will present a computationally efficient algorithm to discover causal mechanisms that manifest in the extreme values of the data.

Jennifer Marsman
Microsoft, Microsoft AI for Earth

Microsoft AI for Earth

In this brief session, you will learn about Microsoft's $50 million USD investment in AI for Earth grant funding, as well as an example of how a grant recipient is using this program to combat climate change.

CI2019 Peer-Reviewed Papers

AUTOMATISING CONSTRUCTION AND EVALUATION OF AGE-DEPTH MODELS FOR HUNDREDS OF SPELEOTHEMS

Carla Roesch[1], Kira Rehfeld[1,2]

***Abstract*—Speleothems, mineral cave deposits, provide high resolution information of water stable isotope ratios for the reconstruction and understanding of past climates. Crucial to the use of these proxy records is the associated time, which is obtained via absolute dating and age-depth modeling. Age-depth modelling approaches of varying complexity have been established in the literature, but often no information on techniques and software used, or estimates of age uncertainty, are given. Case-by-case manual adjustment, as is customarily done, would be forbiddingly time consuming for the target number of 600+ records. Therefore, we aim to automate the process of constructing, comparing and evaluating several age-depth modelling approaches to present transparent, reproducible and reliable Monte Carlo ensembles of age-depth models with age uncertainties. In this work we present the successful evaluation of a complex subset of 88 speleothem records with "well-behaved" constant and continuous growth, variable growth rates, age outliers as well as growth interruptions. We obtained sufficient and acceptable age-depth models for all but six records, which is promising.**

I. MOTIVATION

Major changes in Earth's climate are recorded in paleoclimate archives such as speleothems (mineral cave deposits), corals, sediments, ice cores, tree rings and pollen. This paleoclimate data provides quantitative information on the Earth system response to changes in external forcings. Proxy measurements from such archives are used to date, investigate and understand the consequences of these changes on temperature, sea levels, precipitation or atmospheric patterns. These resulting changes in paleoclimate inform about multi-centennial to millennial baseline variability, against which the recent changes can be compared to assess whether or not they are unusual [1]. Thus, paleoclimate data is fundamental for understanding future climate and evaluating climate model abilities for projections. Reliable age information is crucial and basic information for investigating past climate. However, often techniques and software for constructing age models are not documented, data provided is incomplete, and age uncertainties remain undetermined [2]. Multiple age-depth modelling approaches of varying complexity have been commonly used. However, there is no standard approach for constructing the growth vs. time relation and no universal method of estimating uncertainties. Thus, reproducibility and reliability of age estimates is often not given.

We work with a comprehensive quality-controlled global compilation of stable isotope records from speleothems spanning large parts of the late Pleistocene. Original chronologies, stored in the database, were based on a variety of, sometimes unknown, modelling approaches. For the vast majority of records there is no information on age uncertainty. Our aim is to construct, compare and classify different age-depth models for each record, to give a recommendation on which approach to use as reference. Monte Carlo ensembles than allow to integrate age uncertainties for time-series analyses.

II. DATA & METHOD

Speleothems are mineral cave deposits such as stalagmites or flowstones [3]. Absolute age estimates, so-called radiometric 'dates', can be obtained with high precision using methods based on radioactive decay measurements. In this work we focus on records dated through Uranium-series dating, based on the accumulation of Uranium-daughter products. Knowing the abundance of the naturally, within the material, occurring radioactive isotope, its decay products and it's half-life, age estimates can be obtained. To calculate the age estimates it is assumed that the carbonate

Correspondence: croesch@iup.uni-heidelberg.de, krehfeld@iup.uni-heidelberg.de [1]Institute of Environmental Physics, Heidelberg University, Germany [2]Interdisciplinary Center for Scientific Computing, Heidelberg University, Germany

remains a closed system throughout the accumulation process. However, in reality, the system doesn't remain perfectly closed and measurement uncertainties arise from debris inclusions during the speleothem's growth. These measurement errors are considered to be Gaussian distributed. For the dating process, samples are taken at discrete depths of the speleothem, where the top is generally considered the youngest part of the speleothem (see Fig. 1 left). Additionally, a speleothem is screened for changes in colour and fabric (crystal structure) along the growth axis, which could point at breaks in the speleothem's growth (hiatuses).
Closely spaced measurements of stable oxygen and carbon isotopes in the calcite or aragonite matrix, which to first order represent changes in the isotopic composition of the above-cave precipitation, make speleothems good paleoclimate archives.
Advances in mass spectrometry allow for more frequent dating, but the temporal resolution of proxy samples remains generally higher. Thus, age-depth modelling is required to infer age estimates between adjacent dates (see Fig. 1 right).

The SISAL (Speleothem Isotope Synthesis and Analysis) database v1b [5], [6] is a compilation of 438 carefully screened, globally distributed speleothem stable isotope records covering much of the late Pleistocene. The SISAL working group is a international working group under the auspices of the Past Global Changes programme[1]. 350/438 of the dates are based on U-series dates. The SISAL database [5], [6] provides the data used for constructing the original chronology of the stable isotopes records, including dates that were not used in the original age models. Approaches, which were adapted to the only physical constraint, namely positive semi-definite growth rates, of varying complexity (e.g. linear interpolation, linear regression, etc.) were used for these original chronologies. 20/438 records had no published age model and only 81/438 of all data sets contain estimates of uncertainty of the age model.
Here, we apply six different age-depth modelling techniques to a subset of records with some complexity. We compare and evaluate their suitability for version 2 of the database, where they should provide a reproducible, transparent median age-depth model with confidence intervals derived from Monte Carlo ensembles and (ideally) also the associated Monte Carlo based ensembles of possible age-depth relationships.

[1]PAGES: http://pastglobalchanges.org

We use the open source statistical software `R` [7] for evaluating and implementing the models. The methods included are:
Linear regression (further denoted `lR`), linear interpolation (further denoted `lI`) (e.g. as tested in [8]), `COPRA` [9] and a fitting approach available as `R` function compilation `StalAge` [10]. These four methods generate Monte Carlo ensembles from a Gaussian distribution and interpolate between adjacent dates. However, `StalAge` interpolates by fitting error weighted straight lines through sub-sets consisting of three adjacent data points (see [10]).
Two Bayesian approaches available as `R` packages are tested: a gamma autoregressive semiparametric model `rbacon` [11], [12] and `Bchron` [13], [14], a method based on a continuous Markov monotone stochastic process.
All methods determine uncertainties and median ages based on Monte Carlo ensembles initially generated when running the age-depth model.
`Bchron`, `lI` and `StalAge` cannot deal with hiatuses (However: about 20 % of the records have one or multiple growth interruptions), therefore artificial hiatus dates are added to the input data, with an uncertainty spanning from the date prior to the hiatus to that below the hiatus.
`lI` and `StalAge` require the identification and resolution of non-tractable reversals; we do not delete the dates (that would require a choice of which date to remove, as there are often multiple ways to resolve reversals), but increase uncertainties at the inconsistent ages.
`COPRA` is Matlab-based with a graphic user interface. In standard implementation it cannot be automated. Therefore, too few `COPRA` chronologies are currently available for a comparison.

For testing the algorithm and methods we divided the database in subsets based on the difficulty and complexity. Major challenges arise through hiatuses and reversals (reversed dates of the speleothem which would, if correct, imply physically impossible backwards growth; see Fig. 1). A tractable reversal is present if the 2-sigma dating uncertainties of the two involved dates overlap; else the reversal is considered non-tractable [9]. We distinguished four classes:

(i) Records that run without any additional manipulation: at most tractable reversals, no hiatuses, U-series dates
(ii) Records that at most consist of tractable reversals, only include U-series dates but have at least one

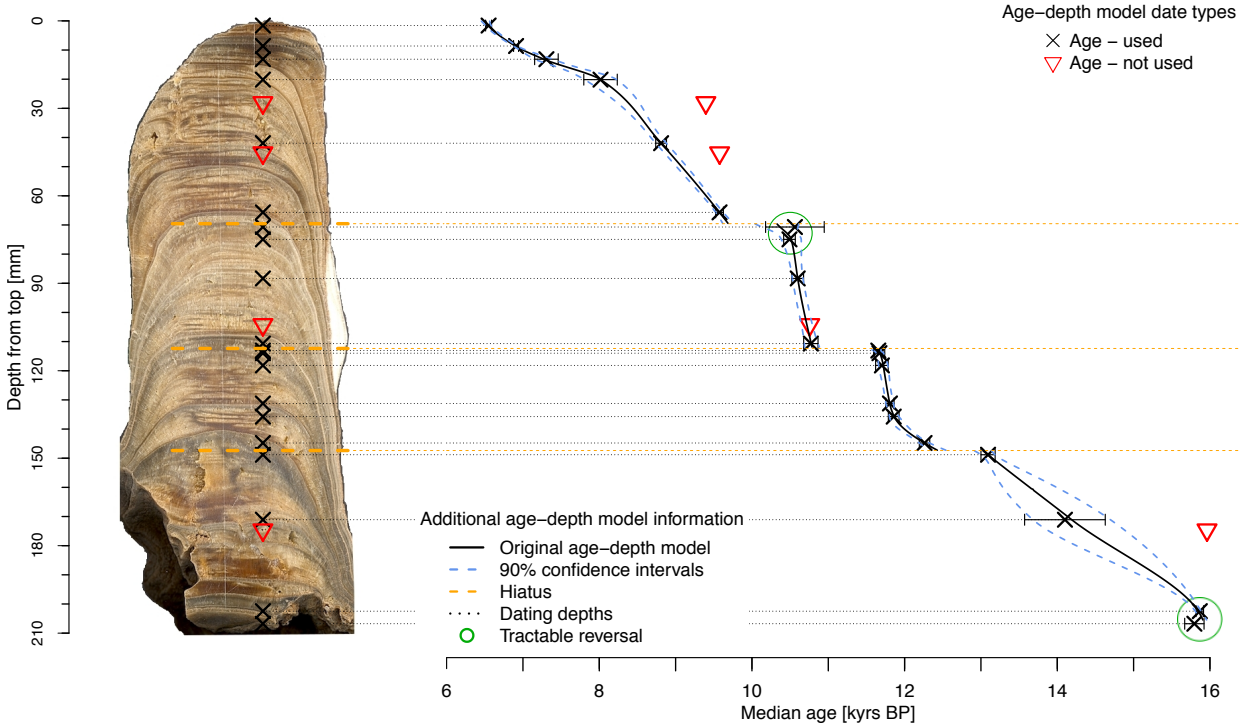


Fig. 1: Illustration of speleothem dating and age modelling for entity ID 63. **Left:** The top of the speleothem is the youngest part. At discrete depths samples are taken and dated using a radiometric method (U-series dating). Changes in colour and/or fabric (crystal structure) of the speleothem can hint at possible hiatuses (growth breaks/gaps). **Right:** An age-depth relationship is constructed using reliable dates (crosses). Triangles represent additional dates that were not used. The age-depth model (black solid line) provides age estimates between radiometric dates for the proxy measurements. Green circles mark tractable reversals (reversed speleothem growth). Image from ref. [4].

hiatus

(iii) Any record that is not U-series dated

(iv) Records that cannot be evaluated, since they have too few or poor inconsistent dates (i.e. non-tractable reversals)

The algorithm used to evaluate the database and construct the age-depth models has to flexibly handle inaccuracies in the underlying data, varying inputs for different methods and test compatibility of method and data:

Step 0: Screen database for all records containing at most tractable reversals, U-series dates and at least 3 dates.

Step 1: For each record selected in step 0:

1) Read in and cast data. Check correction mask if any modifications of the database version are required (missing events: i.e. actively forming event, hiatus, duplicated depths).
2) For each method:
   a) Scan the input data for tractable reversals and hiatuses.
   b) Adjust the input data: for `lI`, `Bchron` and `StalAge` add artificial hiatus ages; for `lI` increase uncertainties at the inconsistent dates (following [10], [9]).
   c) Save the modified input data.
   d) Execute the age-depth model:
      If the age depth model was successfully executed:
      (i) Save the Monte Carlo ensembles (if available).
      (ii) Determine the median age and quantiles;

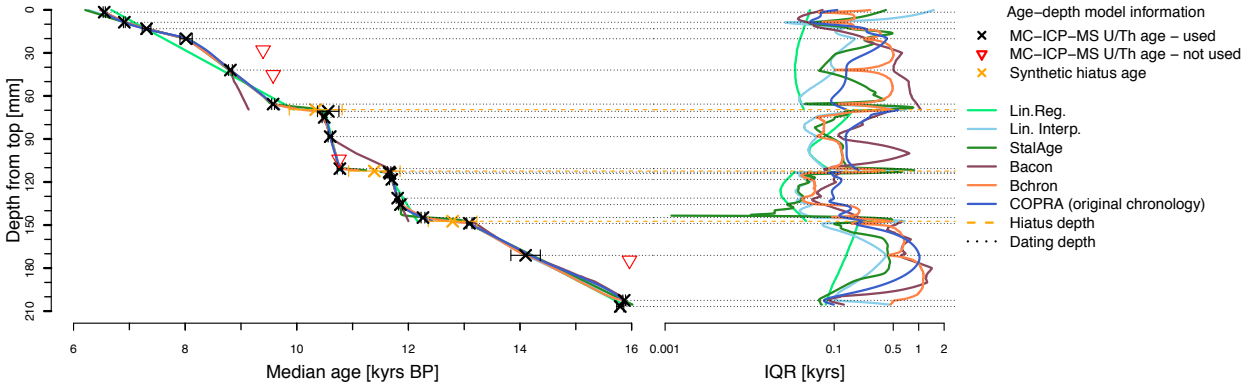


Fig. 2: **Left:** Age-depth models of entity ID 63 for six methods including the original chronology (here: `COPRA`). Used dates are marked by black crosses, orange crosses denote the artificial hiatus dates for `lI`, `Bchron` and `StalAge`. **Right:** Interquartile ranges (IQR) for the age model uncertainty. Dotted black lines representing dating depths, and dashed orange lines marking hiatuses.

save the chronology.

3) Plot an overview plot showing the full available age information for each record: age-depth models including all available data (Fig. 2 left) and age uncertainties (Fig. 2 right).
4) Catch errors thrown during the execution of step 1, save to an error file.
5) Age models that did not run for individual records were marked as FALSE in a separate output file to simplify the final evaluation and comparison of the different methods (step 3).

Step 2: Update the chronology table with the successfully executed age-depth models.

Step 3: Evaluate the chronology table (see Sect. 3); accept or reject (set chronology to NA) the age-depth model.

III. Evaluation

We devised four checks to evaluate the age models:

Check 1 (necessary): No reversals - growth rates are to be positive or zero – speleothems don't grow backwards.
Check 2 (sufficient): Flexibility - the age model is to follow clear growth rate changes.
Check 3 (sufficient): Uncertainty increases between dates and at hiatuses - in the absence of information, uncertainty about the true depth relation should not decrease.
Check 4 (sufficient): Increasing absolute uncertainty with age - the older the sample the larger the age uncertainty should be.

Passing the first criterion needs to be fulfilled for any working age model. The remaining criteria determine the rank of the age model, they are sufficient but not necessary.
Our test subset comprised 88 records, which included speleothems with 'well-behaved' constant and continuous growth, some with variable growth rates, with at least one age outlier as well as at least one hiatus.
For cases where the age model did not run, it was set to a score of -1. An age-depth model that is not reversal free (i.e. fails check 1) scores 0 (i.e., it fails). An age model that had positive semidefinite growth-rates, but failed the remaining criteria, scored 1. Each additional criterion passed gives an additional point.
For most (82/88) entities, at least one acceptable age model could be obtained (Fig. 3). For `lR` reversals appeared for short hiatuses and insufficient input data around them.
`lI`, `Bchron`, `StalAge` and `rbacon` each failed to run for a few records, but did not produce reversals. The failed runs are due to too many tractable reversals or too few dates available between hiatuses. For a large number of entities the evaluation plot shows that they met multiple criteria (dark blue), and acceptable age models can be obtained.

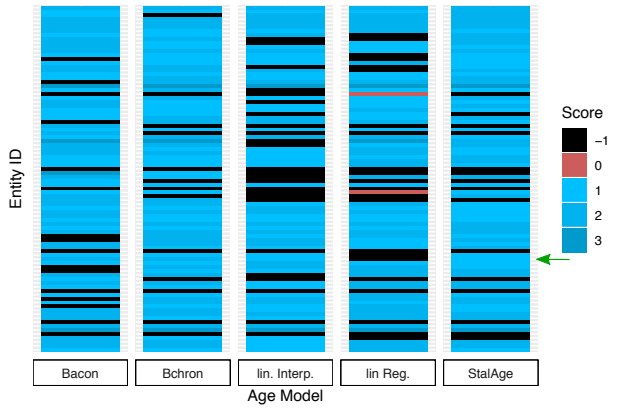


Fig. 3: Quantification of the evaluated age-depth model approaches (x-axis) for 88 entities (y-axis) with hiatuses and outliers. Black (-1) marks cases where the model did not run. Age-depth models that failed with reversals in the median output were marked as red (0). Shades of blue (1-4) represent age models that at least passed this necessary criterion. The darker the shade of blue the more sufficient criteria were met. The green arrow marks the sample core entity ID 63.

IV. Discussion, Conclusion & Outlook

Investigating the behaviour of each method regarding single sufficient criteria, it became apparent that `StalAge` is not flexible enough to properly react to changes in the speleothem growth rate. Further, absolute uncertainties do increase with depth, but not between dates and only at few hiatuses, even with introduced artificial hiatus uncertainty. `rBacon`, `Bchron` and `lI` appeared most robust regarding age uncertainties: they generally increased with depth and in the absence of information (between dates and at hiatuses). Our solution of adding artificial hiatus dates to the dating file appeared to work well, as the IQRs at hiatuses increased as desired. However, the current workflow around `lI` failed for many tractable reversals, as only for 60 % of the subset an age-depth model was obtained. Therefore, a method to filter and consequently delete individual dates from the modelling process will be required.

For a difficult subset of 88 speleothems from the SISAL database [5], [6] we were able to construct at least one working age-depth model for all but six records. This suggests that it will be possible to construct working age-depth models for most of the records. Our extension of `lI`, `Bchron` and `StalAge` for dealing with hiatuses showed promising results, as uncertainties around the growth interruptions increased, but no reversals were produced. We found that even small inconsistencies and inaccuracies in the present version of the database could lead to failed age-depth models, and had to be caught, verified with the original reference, and fixed. Additional screening and expert evaluation of the final chronologies will therefore be conducted, which will also allow the cross-checking of our automated approach. This underlines that dealing with large palaeoclimate datasets requires careful and flexible approaches.

Apart from U-series dated ages, radiocarbon dated records are also included in the SISAL database. These records require special treatment, since in a first step the radiocarbon ages have to be calibrated using different calibration curves depending on the extraction region and the time period it covers. For some records, with high scientific value and complex age models that could not be considered in the automated routine, tailored age models will be added.

Acknowledgments

We thank the SISAL working group for the continued development of the database and Sahar Amirnezhad-Mozhdehi, Dennis Rupprecht, Laia Comas-Bru, Franziska Lechleitner and Nils Weitzel for discussions. We acknowledge PAGES support for SISAL. We appreciate greatly Sebastian Breitenbach's ideas regarding the synthetic hiatus ages, Maarten Blaauw's continuous improvements of the `rBacon` package, and Denis Scholz' input for the adjustment of `StalAge`. KR acknowledges the German Research Foundation for funding (code RE3994-2/1). The code used for constructing the age models is available at github.com/palaeovar/SISAL.AM.

References

[1] V. Masson-Delmotte, M. Schulz, A. Abe-Ouchi, J. Beer, A. Ganopolski, J. González Rouco, E. Jansen, K. Lambeck, J. Luterbacher, T. Naish, T. Osborn, B. Otto-Bliesner, T. Quinn, R. Ramesh, M. Rojas, X. Shao, and A. Timmermann, "Information from paleoclimate archives," in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Stocker, T.F., D. Qin, G. Plattner, M. Tignor, S. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. Midgley, eds.), Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013.

[2] M. Blaauw, "Methods and code for 'classical' age-modelling of radiocarbon sequences," *Quaternary Geochronology*, vol. 5, no. 5, pp. 512–518, 2010.

[3] I. J. Fairchild and A. Baker, *Speleothem Science*. Wiley Online Library, 2012.

[4] F. A. Lechleitner, S. F. Breitenbach, H. Cheng, B. Plessen, K. Rehfeld, B. Goswami, N. Marwan, D. Eroglu, J. Adkins, and G. Haug, "Climatic and in-cave influences on δ 18 o and δ 13 c in a stalagmite from northeastern india through the last deglaciation," *Quaternary Research*, vol. 88, no. 3, pp. 458–471, 2017.

[5] K. Atsawawaranunt, L. Comas-Bru, S. Amirnezhad Mozhdehi, M. Deininger, S. P. Harrison, A. Baker, M. Boyd, N. Kaushal, S. M. Ahmad, Y. Ait Brahim, and SISAL working group members, "The SISAL database: A global resource to document oxygen and carbon isotope records from speleothems," *Earth System Science Data*, 2018.

[6] L. Comas-Bru, S. P. Harrison, M. Werner, K. Rehfeld, N. Scroxton, C. Veiga-Pires, and S. working group members, "Evaluating model outputs using integrated global speleothem records of climate change since the last glacial," *Climate of the Past*, vol. 15, no. 4, pp. 1557–1579, 2019.

[7] R core team, "R: a language and environment for statistical computing [online]. R foundation for statistical computing," 2016.

[8] R. J. Telford, E. Heegaard, and H. J. B. Birks, "All age–depth models are wrong: but how badly?," *Quaternary Science Reviews*, vol. 23, no. 1-2, pp. 1–5, 2004.

[9] S. F. M. Breitenbach, K. Rehfeld, B. Goswami, J. Baldini, H. Ridley, D. Kennett, K. Prufer, V. Aquino, Y. Asmerom, V. Polyak, *et al.*, "Constructing proxy-record age models (COPRA).," *Climate of the Past*, vol. 8, no. 5, pp. 1765–1779, 2012.

[10] D. Scholz and D. L. Hoffmann, "Stalage – an algorithm designed for construction of speleothem age models," *Quaternary Geochronology*, vol. 6, no. 3-4, pp. 369–382, 2011.

[11] M. Blaauw, J. A. Christen, *et al.*, "Flexible paleoclimate age-depth models using an autoregressive gamma process," *Bayesian Analysis*, vol. 6, no. 3, pp. 457–474, 2011.

[12] M. Blaauw, J. A. Christen, J. E. Vazquez, T. Belding, J. Theiler, B. Gough, and C. Karney, *rbacon: Age-Depth Modelling using Bayesian Statistics*, 2019. R package version 2.3.9.1.

[13] J. Haslett and A. Parnell, "A simple monotone process with application to radiocarbon-dated depth chronologies," *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, vol. 57, no. 4, pp. 399–418, 2008.

[14] A. Parnell, *Bchron: Radiocarbon Dating, Age-Depth Modelling, Relative Sea Level Rate Estimation, and Non-Parametric Phase Modelling*, 2018. R package version 4.3.0.

RELIABLE PATTERN EXTRACTION FOR CLIMATE DATA

James Fulton[1], Gabriele Hegerl[1]

***Abstract*—We propose a method to quantitatively test the ability of latent variable models to recover underlying climate modes by using climate-like synthetic data. We develop methods to generate the synthetic data and find that the most common way of extracting modes, empirical orthogonal functions, also known as principal component analysis, performs worse than two more modern methods in 181 out of 182 cases tested. Interpretations of the modes found by the two more modern methods would be more correct both qualitatively and quantitatively.**

I. MOTIVATION

Our climate is a highly variable and chaotic system. However, out of the chaos we observe emergent patterns that repeat. These are referred to as climate modes.

Our understanding of climate modes is dependent on our ability to extract these physically relevant patterns from observations and climate simulations. These modes can have global effects, such as the El Niño Southern Oscillation (ENSO) which has a worldwide impact, or more localised effects such as the Indian Ocean Dipole (IOD). Once discovered, these modes are used further in teleconnection studies [1], in attribution studies [2], and in future forecasts [3], [4].

One of the major tools which is used to extract patterns is empirical orthogonal functions (EOFs) [5], also known as principal component analysis (PCA). The spatial vectors, principal components (PCs), returned by this method are often interpreted as if they represent a single physical process in the climate, such as ENSO, which is sometimes described by two PCs [6]. Modes such as ENSO are well established physical processes and the PCs are often used merely to back them up or visualise them. However, different modes, with less understood mechanisms, such as the Pacific Decadal Oscillation (PDO) [7], often rely on their PC to validate their existence and direct the search for their mechanism.

Corresponding author: J. Fulton, james.fulton@ed.ac.uk
[1]School of GeoSciences, University of Edinburgh

Along with the spatial vector, there is a time series associated with each of these modes, known as its index. The PDO is often defined spatially by a PC and in time by an index.

Although PCA as a tool is maximally efficient in compressing data and capturing as much variance as possible with any reduced number of dimensions, interpreting these modes or using them as evidence to support a climate mechanism can be misleading [8].

One of the major issues with PCA, often discussed, is that the modes returned are orthogonal, whilst modes of the climate system are most certainly not. Often this is assumed to mean that the first few modes returned by PCA are correct, but later modes, as they are constrained to be orthogonal to the earlier ones, don't represent physically coherent patterns. This interpretation is misleading, and in fact even the first PC can't capture the true underlying modes if the system modes are non-orthogonal.

This is a significant problem, which is shared by other methods of mode discovery in climate, such as extended EOFs. PCA can easily find the optimal plane for a lower dimensional subspace, but it cannot correctly identify the direction of modes on this plane. This is to be expected as PCA does not take into account any of the time correlation information in the data. In fact, the same PCs are returned if a series of climate measurements are permuted in time than if they are not. This means the PCA algorithm essentially discards this information. It is interesting to note that when interpreting the PCs recovered, we use physical mechanisms and reasoning which rely on time ordering, and yet the modes we try to interpret are discovered using a method which disregards this.

There are many latent variable models (LVMs) which take data and, with a various sets of constraints and optimisation objectives, often including the use of time order information, project the data onto a reduced dimension. In this paper we suggest alternative, more promising methods to PCA, propose a framework to create climate-like synthetic data and test the methods

on generated datasets.

II. METHOD

Synthetic data. Generating realistic data should be an integral part of testing any LVM. When analysing climate measurements we may only have a few periods of the lowest frequency modes, this further complicated by having missing observations and inhomogeneous data.

Even when using the output of climate simulations, where we have complete information on the state of the system at any time step, there exists no ground truth of climate modes to which we can compare. Therefore, we should be sure that the methods we employ to extract these modes perform reliably in synthetic cases where there truly are underlying coherent modes which are maximally similar to our beliefs about climate observations and simulations.

We generate modes such as those in the first column of figure 1 using a draw from a squared exponential (SE) covariance kernel function [9] with distance quantified by great-arc length. This allows us to encode the assumption that any climate modes, such as surface temperature or sea level pressure, have spatial correlation structures based on globe surface distance. This also gives us the ability to choose the spatial scale of correlations. We chose to use a length-scale of 30°. By controlling the draw from this kernel we can also generate patterns which are spatially sparse, as in figure 1 row (c) , and spatially dense, as in rows (a) and (b).

We can generate an index time series by simulating a SARMA process. The SARMA time series model is a stochastic model often used to describe climate data [5]. Together these methods give us modes which are linear, but spatially and time correlated, and having correlation statistics which are easily tuned. We project the spatial modes $v^{(i)}(x, y)$ along with their index $z^{(i)}(t)$ to give the projection of each individual mode $D^{(i)}(x, y, t) = v^{(i)}(x, y)\, z^{(i)}(t)$ and sum many of these projections to create a mixed mode dataset $X(x, y, t) = \sum_i D^{(i)}(x, y, t)$. We add space-time correlated noise to the data in a similar way.

We can also generate non-linear modes by extending the kernel function into the time dimension. By taking the product of the original spatial kernel with an SE time covariance function we can create modes which move back and forth along a non-linear path between two extremes. If we use a periodic time covariance function, then we can simulate processes which are non-linear and loop, such as a hysteresis mode. This is very appropriate for curves which grow and shrink along different pathways, which is likely the case for most climate modes. We can also generate moving wave modes, like the the Madden–Julian oscillation [10], by creating a linear spatial mode and rotating it on the surface of the sphere with time.

If the mode is non-linear in one of these ways then the space and time components cannot be factored out, therefore we express the projection of the mode as $D^{(i)}(x, y, t) = v^{(i)}(x, y, z^{(i)}(t))$

Quantifying matches. We have chosen the metric described in equation (1) to quantify the match of the discovered modes to the generated modes. Here we use an overline, $\overline{A}$ of any matrix, A, to indicate its time average. We also use $\widehat{A}$ to indicate a quantity A already described but discovered by an LVM. For example $\widehat{v}^{\,(i)}$ $\left(\widehat{z}^{\,(i)}\right)$ would be a spatial mode (index) discovered by an LVM.

$$m_{ij} = \frac{1}{N} \frac{(D^{\,(i)} - \overline{D}^{\,(i)}) \cdot (\widehat{D}^{\,(j)} - \overline{\widehat{D}}^{\,(j)})}{(\frac{1}{N} D^{(i)} \cdot D^{(i)} - \overline{D}^{\,(i)} \cdot \overline{D}^{\,(i)})^{1/2} (\frac{1}{N} \widehat{D}^{\,(j)} \cdot \widehat{D}^{\,(j)} - \overline{\widehat{D}}^{\,(j)} \cdot \overline{\widehat{D}}^{\,(j)})^{1/2}} \tag{1}$$

where $A \cdot B$ denotes the Frobenius inner product between A and B.

This metric is generalised such that it can be used whether the generated mode and the discovered mode are linear or non-linear or a mixture. It simply relies on the ability to perform an independent projection of each mode.

In the case where both the generated mode and the discovered mode are linear then this metric is identical to

$$m_{ij} = \left(\frac{1}{N} \frac{(z^{(i)} - \overline{z}^{\,(i)}) \cdot (\widehat{z}^{\,(j)} - \overline{\widehat{z}}^{\,(j)})}{\sigma^{(i)} \widehat{\sigma}^{\,(j)}} \right) \left(v^{(i)} \cdot \widehat{v}^{\,(j)} \right), \tag{2}$$

where $\sigma^{(i)}$ is the variance of $z^{(i)}$ and because the spatial vectors are normalised this is also equal to the variance of the projection $D^{(i)}$.

This metric is simply the correlation coefficient of the two index time series multiplied by the dot product of the spatial vectors. This makes the metric of choice

highly interpretable in the simple case and highly generalised for more complex, non-linear cases. This metric gives a maximum value of unity when the projections are perfect and a minimum value of zero when the projections are orthogonal.

Following this, testing any method of mode discovery is straight forward. We generate modes and their associated time series using the methods described above, mix them together additively, fit the LVM and then quantify the matches. We will have multiple generated modes and multiple discovered modes which require matching. This is accomplished via greedy one-to-one matching using the match metric matrix (equation (1)) to create a set of matches $\{(i_n\,, j_n)\}$.

Test cases. Under the described testing procedure we test different realisations of synthetic climate to see how each LVM deals with different features. We can consider how each method performs when all of the generated modes are linear, or when some of them are non-linear. We can consider how sensitive each method is to different levels of noise added to the data or whether it is confounded by having variance dominated by a single mode.

We devised the following test cases, where we generated 30 $(32\,|\,60)$ datasets for test cases 1–3 $(4\,|\,5)$ each with multiple modes, and each mode and index pair consisting of 7008 spatial points and 14400 time measurements.

Test case descriptions:

1) 8 dense linear generated modes each with equal variance. Low level of noise.
2) 4 dense and 4 sparse linear generated modes each with equal variance. Low level of noise.
3) 8 dense linear generated modes, one with variance multiple times larger than the rest. Low level of noise.
4) 8 dense linear generated modes each with equal variance. Various high levels of noise.
5) Number of modes chosen from a Poisson distribution with peak at 9 and an enforced minimum of 5. Mixture of linear and non-linear modes where type is randomly drawn. Variance of each mode drawn from a wide beta distribution. Level of noise also drawn from beta distribution.

To allow for further summarisation we devise the two summary metrics for the overall performance of an LVM on a single dataset from one of the above test cases. These are M the mean of the match metrics m_{ij} across the set of one-to-one matches, and M_σ which is the variance weighted mean of one-to-one matches given below.

$$M_\sigma = \frac{1}{\sum_{k=1}^{N} \sigma^{(k)^2}} \sum_{n=1}^{N} m_{i_n j_n} \sigma^{(i_n)^2} \tag{3}$$

Latent variable models. The methods we have chosen to test are PCA, slow feature analysis (SFA) [11], dynamic mode decomposition (DMD) [12] and a baseline model.

We chose SFA and DMD because they both utilise the time correlations in the data and can both be used to return linear modes. They therefore have a much better chance at capturing the correct modes and also present easy alternatives to fit into the standard climate analysis workflow. SFA aims to find the slowest varying modes and DMD aims to find modes which evolve independently.

The baseline model is essentially projection onto a blind guess of the modes. To generate the modes of this model we take dense draws from an SE spatial kernel function with length-scale of 30°. To find the time index for these 'discovered' modes we project the generated mixed dataset X onto each spatial vector $\widehat{v}^{(i)}$.

III. EVALUATION

SFA and DMD significantly outperformed PCA (EOFs) in finding the correct generative modes under all of our test cases.

Overall, out of 182 datasets, SFA scored highest under M_σ in 162 cases, DMD in 19, and PCA in 1. Out of 1456 modes across all datasets, SFA had the highest one-to-one match metric for 1177 modes, DMD for 274, and PCA for 5.

SFA and DMD scored higher than the baseline model under M_σ in 100% of datasets, and PCA beat the baseline in 96% of datasets. SFA found better one-to-one matches than the baseline for 100% of modes, DMD beat the baseline for 99% of modes and PCA beat the baseline for only 36% of modes.

SFA and DMD performed better whether the modes were all linear and dense, or a mixture of dense and sparse; whether the set of modes forming a dataset was complicated by having non-linear modes, or high amounts of noise, or having a mixture of all of these features. This is shown in the median summary metrics across test cases in table I.

An example of generated modes and the modes discovered by PCA, SFA and DMD is shown in figure 1. These three modes are taken from a typical dataset from test case 2. Row (a) of the plot shows a mode where PCA has failed to capture the positive area over South America and the negative areas in the pacific. It

TABLE I: The median non-weighted, M, and variance weighted M_σ, match metrics across all generated datasets for PCA, SFA and DMD across the five test cases. In the test cases where the variances are all equal $M = M_\sigma$.

	test case 1	test case 2	test case 3		test case 4	test case 5	
	$M = M_\sigma$	$M = M_\sigma$	M	M_σ	$M = M_\sigma$	M	M_σ
Base	0.248	0.247	–	–	–	0.202	0.225
PCA	0.318	0.320	0.395	0.651	0.270	0.342	0.379
DMD	0.682	0.702	0.791	0.878	0.613	0.551	0.623
SFA	0.814	0.821	0.848	0.915	0.639	0.602	0.664

has also exaggerated the positive areas over Australia and the Middle East. If this behaviour were to occur whilst analysing measurements of the real climate then they could be misinterpreted.

Even in row (b) of the figure, the mode for which PCA achieved its best match in this dataset, both SFA and DMD still scored higher. In this case PCA narrows the highly positive region over North America and misses the positive area over Australia.

The final row of the figure shows a false mode discovered by PCA which is characteristic of the method. In this case PCA has made a dipole out of what is really a monopole. This is a feature we saw often in our results and is similar to the example used in [8], where they question whether the IOD is in fact a dipole. This concern could be overcome by using SFA or DMD which, as shown in this case, can correctly capture simple monopoles.

IV. CONCLUSION

We presented methods to create climate-like synthetic datasets on which we can test latent variable models for climate mode discovery. We show, by experiment on these newly created datasets, that the most commonly used method of extracting climate modes, principal component analysis, also known as empirical orthogonal functions, is the least reliable of the methods tested at uncovering the underlying modes. It was beaten by the other two methods tested in 181 out of 182 datasets and for 1451 out of 1456 total modes. We showed that this method, in both a quantitative and qualitative sense, could lead to the discovery of false climate modes, which following more analysis could lead to false proposed climate mechanisms.

We believe that the climate community should adopt either of the other methods tested, either dynamic mode decomposition or slow feature analysis, for the purposes of extracting linear climate modes and perhaps use both so that discovered patterns can be verified.

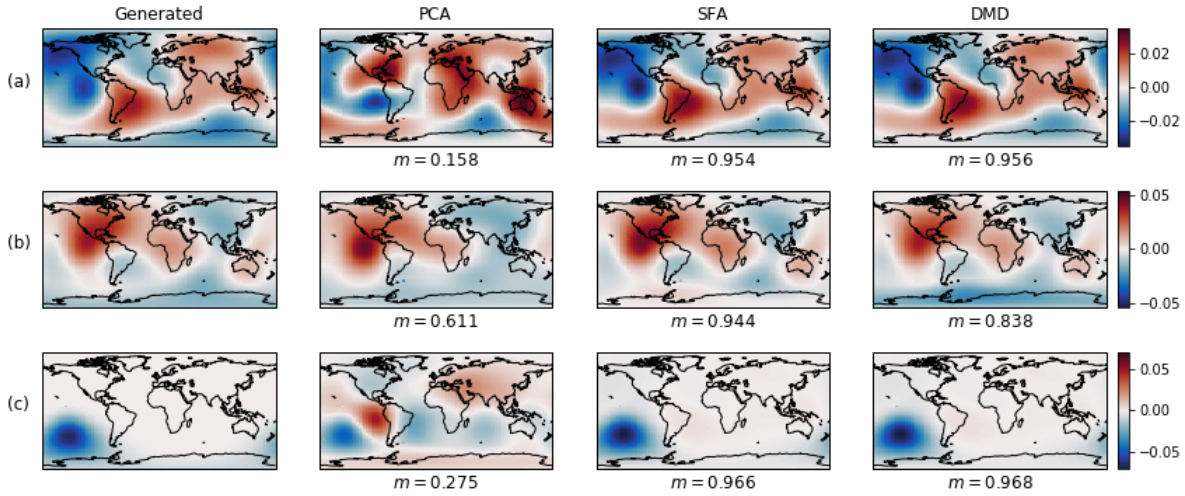


Fig. 1: A comparison between three synthetic generated modes and those found using three different methods. The left column shows the three of the eight modes generated for this dataset. The PCA, SFA and DMD columns show the mode discovered by each method which best matches the generated mode in the same row. The match metric as per equation (1) for each match is stated below each subplot. Note that the coastlines are fictitious but included to act as an eye guide and highlight differences which in real climate data we might try to interpret.

Acknowledgments. Funding for the authors was provided by NERC through the E3 Doctoral Training Partnership.

REFERENCES

[1] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, p. 8502, 2015.

[2] A. Weisheimer, N. Schaller, C. O'Reilly, D. A. MacLeod, and T. Palmer, "Atmospheric seasonal forecasts of the twentieth century: multi-decadal variability in predictive skill of the winter north atlantic oscillation (nao) and their potential value for extreme event attribution," *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 703, pp. 917–926, 2017.

[3] Z. Wu, B. Wang, J. Li, and F.-F. Jin, "An empirical seasonal prediction model of the east asian summer monsoon using enso and nao," *Journal of Geophysical Research: Atmospheres*, vol. 114, no. D18, 2009.

[4] N. C. Johnson, D. C. Collins, S. B. Feldstein, M. L. L'Heureux, and E. E. Riddle, "Skillful wintertime north american temperature forecasts out to 4 weeks based on the state of enso and the mjo," *Weather and Forecasting*, vol. 29, no. 1, pp. 23–38, 2014.

[5] H. Von Storch and F. W. Zwiers, *Statistical analysis in climate research*. Cambridge university press, 2001.

[6] A. Timmermann, S.-I. An, J.-S. Kug, F.-F. Jin, W. Cai, A. Capotondi, K. Cobb, M. Lengaigne, M. J. McPhaden, M. F. Stuecker, *et al.*, "El niño–southern oscillation complexity," *Nature*, vol. 559, no. 7715, p. 535, 2018.

[7] N. J. Mantua, S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, "A pacific interdecadal climate oscillation with impacts on salmon production," *Bulletin of the american Meteorological Society*, vol. 78, no. 6, pp. 1069–1080, 1997.

[8] D. Dommenget and M. Latif, "A cautionary note on the interpretation of eofs," *Journal of climate*, vol. 15, no. 2, pp. 216–225, 2002.

[9] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*, vol. 2. MIT Press Cambridge, MA, 2006.

[10] C. Zhang, "Madden-julian oscillation," *Reviews of Geophysics*, vol. 43, no. 2, 2005.

[11] L. Wiskott and T. J. Sejnowski, "Slow feature analysis: Unsupervised learning of invariances," *Neural computation*, vol. 14, no. 4, pp. 715–770, 2002.

[12] J. N. Kutz, X. Fu, and S. L. Brunton, "Multiresolution dynamic mode decomposition," *SIAM Journal on Applied Dynamical Systems*, vol. 15, no. 2, pp. 713–735, 2016.

PREDICTING 3D RADIATIVE HEATING RATE FIELDS FROM SYNERGISTIC A-TRAIN OBSERVATIONS COMBINED WITH DEEP LEARNING TECHNIQUES

Friederike Hemmer[1], Claudia J. Stubenrauch[1], Sofia E. Protopapadaki[2]

***Abstract*—Upper tropospheric clouds strongly influence the energy budget of the Earth, but the structure of their vertical heating rate profiles is still poorly known. This is due to the fact that global observations of these heating rates are sparse. The active lidar and radar measurements from CALIPSO and CloudSat as part of the A-Train satellite constellation provide such heating rate profiles, but only on narrow nadir tracks separated by about 2500 km between successive orbits. The Atmospheric Infrared Sounder (AIRS) on the other hand provides cloud properties with a large instantaneous horizontal coverage, but not their vertical structure. In this study, we train deep learning neural networks with four years of collocated data, including meteorological reanalyses, to develop optimized non-linear regression models which predict these heating rates as a function of the most suitable cloud and atmospheric properties. These models are then applied to the full statistics of more than 15 years of AIRS observations in order to construct complete 3D radiative heating rate fields which can be related to the different parts of tropical convective systems for process and climate studies.**

I. MOTIVATION

Clouds play an important role in the global climate system by altering the net surface radiation and influencing the diabatic heat budget of the atmosphere through radiative heating/cooling as well as latent heat release (e.g. [1], [2]). In particular, Upper Tropospheric (UT) clouds, which are most frequently observed in the tropics and represent about 40% of the Earth's total cloud cover [3], often form as cirrus anvils from convective outflow, building mesoscale systems. In a warming climate, tropical convection will intensify, leading to colder convective systems which may include a larger fraction of thin cirrus within and around the anvils. The radiative heating of these thinner cirrus may be critical to cloud climate feedback. However, the horizontal and vertical structure of the radiative heating rates is still poorly known which is partly due to a lack of observation. This study aims to fill some of these gaps by constructing complete 3D radiative heating rate fields obtained from combining several satellite observations and meteorological reanalyses with deep learning which is an extremely active research area [4] with a constantly increasing number of applications in climate science.

The satellite observations used here originate from the A-Train constellation [5] composed of several satellites equipped with different instruments in a sun-synchronous polar orbit with local overpass times around 1:30 AM and 1:30 PM. We will focus on AIRS aboard the Aqua satellite as well as the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) aboard CALIPSO and the Cloud Profiling Radar (CPR) aboard CloudSat. The good spectral resolution of AIRS leads to reliable cloud properties, even for thin cirrus. As a passive cross-tracking instrument it provides a large horizontal coverage but does not give information on the vertical structure of the clouds. The latter can be obtained from the active CloudSat radar and CALIPSO lidar measurements which are performed only on a narrow nadir path, so the coincidence with the passive measurements is limited to narrow nadir tracks separated by about 2500 km (see upper panel of Fig. 1 for illustration). To fill the gaps between the orbits and expand the vertical information over complete cloudy scenes, we use supervised deep learning based on artificial neural networks (ANN) to relate the cloud properties retrieved from AIRS, together with coincident atmospheric and surface properties from the meteorological reanalyses

Corresponding author: C. Stubenrauch, stubenrauch@lmd.polytechnique.fr. [1]Laboratoire de Météorologie Dynamique/Institut Pierre-Simon Laplace (LMD/IPSL), Sorbonne Université, Ecole Polytechnique, CNRS, Paris, France. [2]COOPETIC.

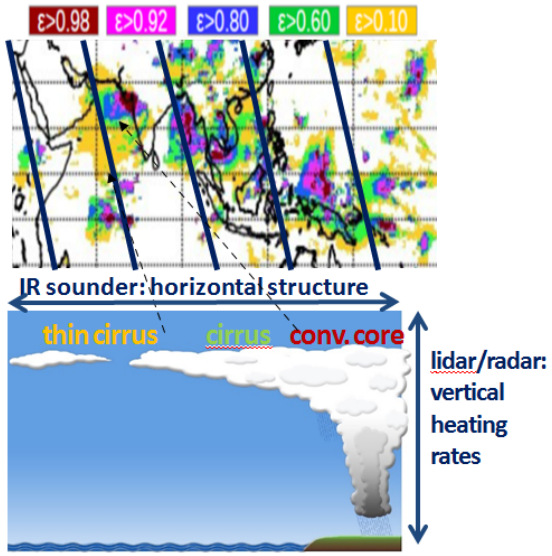


Fig. 1. Upper panel: emissivity distribution of UT cloud systems from AIRS, overlaid with the nadir tracks of the CALIPSO lidar/CloudSat radar. Lower panel: illustration of the cloud system approach.

ERA-Interim [6], to the vertical radiative heating rate profiles obtained from CloudSat/CALIPSO.

II. METHOD

A. Cloud System Concept

As mentioned above, we are particularly interested in UT clouds which occur most frequently in the tropics. Thus, we will focus on the tropical latitude band (30 °N to 30 °S) in this paper. UT clouds often form as cirrus anvils from convective outflow and build mesoscale systems. To study their properties in dependence of the convective strength, a cloud system concept has been developed which is based on two independent variables retrieved from AIRS measurements: emissivity and height [7]. As a first step, cloud systems are built from adjacent elements of similar cloud height represented by a cloud pressure $p_{cloud} < 440$ hPa. Secondly, the horizontal emissivity structure allows to distinguish between convective cores (Cb defined by an emissivity $\epsilon > 0.98$), thick cirrus ($0.5 < \epsilon < 0.98$) and thin cirrus anvil ($\epsilon < 0.5$). The lower panel of Fig. 1 illustrates such a cloud system with its three different cloud types where the convective core only represents a small portion of the system. The upper panel presents an example of the horizontal emissivity structure of these cloud systems.

B. Data

The Aqua satellite carrying the AIRS instrument has been launched in 2003, hence a 15-year time series of cloud properties is available (2003-2018)[3]. The retrieved cloud properties are cloud emissivity, pressure, temperature and height, together with their uncertainties, derived from eight radiances along the wings of the CO2 absorption band around 15 μm. The retrieval is based on a weighted χ^2-method [8]. To relate adjacent pixels within 2° x 2° grid boxes, 16 weather states have been determined based on a k-means method applied to histograms of cloud emissivity and pressure as described by [9]. ERA-Interim meteorological reanalyses [6] are used to obtain surface and atmospheric properties, including temperature and water vapor profiles from which the relative humidity (RH) for ten atmospheric layers is calculated in a similar way as in [10]. All variables are summarized in Table I.

The shortwave (SW) and longwave (LW) radiative heating rate profiles retrieved from the active instruments originate from the CloudSat 2B-FLXHR-LIDAR product which is provided by the National Aeronautics and Space Administration (NASA) and has been described by [11] and [12]. CloudSat and CALIPSO have been launched in 2006. The collocated AIRS-CloudSat-CALIPSO-ERA-Interim dataset used in this study comprises four years of data from 2007 to 2010.

C. Algorithm and Experiments

To extend the vertical heating rate structure throughout entire cloud systems, we develop optimized non-linear regression models by using supervised deep learning. The models are trained and tested along the nadir tracks of the active instruments using the four years of collocated AIRS-CloudSat-CALIPSO-ERA-Interim data. We apply the TensorFlow framework and the Keras program library for python. Our ANN consists of three fully connected layers with relu activation. The training data set is randomly separated in three portions as follows: 80% are used for training, 10% for validation and 10% for testing, stratified by cloud type and by a day/night flag. We use the mean absolute error (MAE) between the prediction and the true value along the track as loss function. The training is performed by the Adam optimizer.

Two different kinds of sensitivity studies are conducted. On one hand, the effect of varying input variables is tested. On the other hand, we investigate how many models have to be developed to optimally extend the radiative heating rate profiles. This second set of sensitivity studies is performed to examine if it is advantageous to separate the training for land and ocean as well as for different cloud types, since the

TABLE I
LIST OF VARIABLES.

Clouds	
CIRS-AIRS cloud properties	ϵ_{cloud}, p_{cloud}, T_{cloud}
Cloud retrieval uncertainties	$d\epsilon_{cloud}$, dp_{cloud}, dT_{cloud}, χ^2_{min}
Cloud spectral emissivity diff.	($\epsilon_{cloud}(12\,\mu m)$ - $\epsilon_{cloud}(9\,\mu m)$)
CIRS weather state at 2° x 2°	WS (1-16), kernel distance
Atmosphere	
Brightness temperatures	$Tb_{11.85}$, $\sigma(Tb_{11.85})$, $Tb_{7.18}$
ERA-Interim atmos. properties	total precip. water, $p_{tropopause}$
Atmospheric classification	TIGR atmosphere [13]
Relative humidity profile	RH profile over 10 layers
Temperature profile	T profile over 10 layers
Surface	
ERA-Interim surface properties	T_{surf}, p_{surf}, nb of atm. layers

cloud properties vary strongly between the cloud types and over land/ocean.

III. RESULTS

To analyse the influence of the different input variables, ANN models have been developed without cloud type separation including data of all cloud types over ocean. Table II presents the MAE of the predicted cloud LW heating rates, depending on the set of input parameters. As a first step, the basic cloud properties (emissivity, spectral emissivity difference, pressure and temperature, together with their according uncertainties and the minimum from the χ^2-method as quality index), atmospheric properties (total precipitable water, tropopause pressure, classification of the atmospheric profile from TIGR [13] and brightness temperatures at 11.85 μm and 7.18 μm as well as the brightness temperature variance at 11.85 μm over 3 x 3 AIRS footprints) and surface properties (surface pressure, temperature and the number of layers of the profile) have been used. For this basic experiment, a MAE of 0.84 is obtained. Adding firstly the weather states, secondly the RH profile and finally the temperature profile, leads to slight improvements. However, using all available parameters the maximum improvement is only about 6% compared to the basic experiment.

As a next step, we investigated if the training should be performed separately over land and ocean and for different cloud types. Figure 2 shows the predicted LW heating rate profiles for Cb, cirrus and thin cirrus over ocean compared to the observations (black line). The dark blue line represents the prediction from the model that has been trained with only high clouds only over ocean, the light blue line represents the prediction from the model trained with all clouds only over ocean and the red line represents the prediction

TABLE II
MEAN ABSOLUTE ERROR OF THE DIFFERENT EXPERIMENTS.

Basic	+ weather states	+ RH profile	+ T profile
0.84	0.84	0.80	0.79

from the model trained with all clouds over ocean and land together. The predicted profiles are very similar for all cases and agree well with the observations from CALIPSO-CloudSat. The cloud type Cb may be represented slightly better when applying the model developed for only high clouds because the frequency of Cb is small compared to the other cloud types (only 5% of all clouds). Figure 2 also illustrates the strong cooling above Cb (200 hPa) and a slight heating of thin cirrus (100 hPa) in the upper troposphere.

Finally, the models developed for all clouds over ocean and land together as well as over ocean and land separately have been applied to one month of data, January 2008, corresponding to a La Niña situation. During La Niña, the tropical convection is shifted towards the West Pacific as illustrated by the observed emissivity structure of the UT cloud systems for this month presented in Fig. 3. Figure 4 shows the corresponding LW radiative heating rates at the four pressure levels of 106, 200, 525 and 850 hPa from the predictions with the two different models compared to the nadir track statistics from the CALIPSO/CloudSat observations. The horizontal structure of both predictions agrees quite well with the one from CALIPSO/CloudSat. However, the structure of the laterally extended fields appears much clearer. Compared to Fig. 3, the warming in the upper troposphere (106 hPa) corresponds to thin cirrus while the cooling at 200 hPa is found above optically thick cirrus which heat the middle troposphere (525 hPa) at the same time. In the lowest layer (850 hPa), a cooling above low clouds and a heating by thick high clouds can be observed. The LW radiative heating rate fields

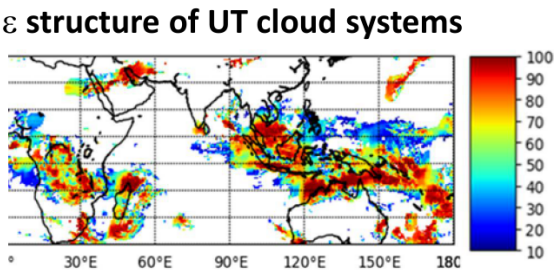


Fig. 3. Emissivity structure of UT cloud systems from AIRS for January 2008 (La Niña).

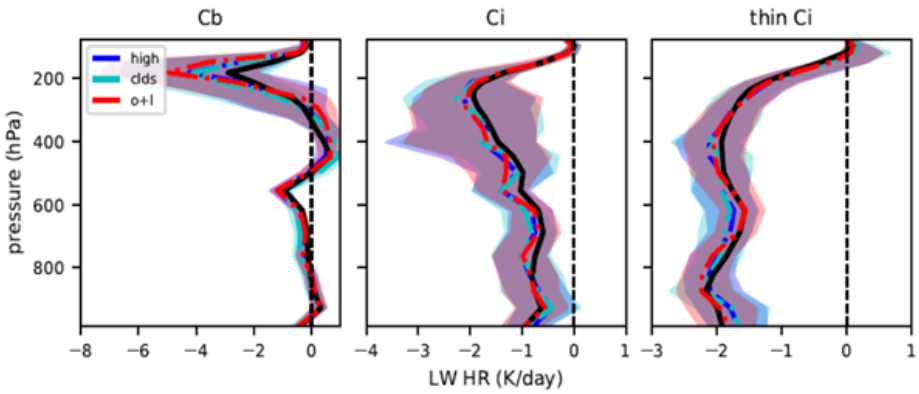


Fig. 2. Observed (black line) and predicted (dark blue line: model trained with only high clouds over ocean, light blue line: model trained with all cloud types over ocean, red line: model trained with all cloud types over ocean and land together) LW radiative heating rate profiles for Cb, cirrus and thin cirrus (from left to right).

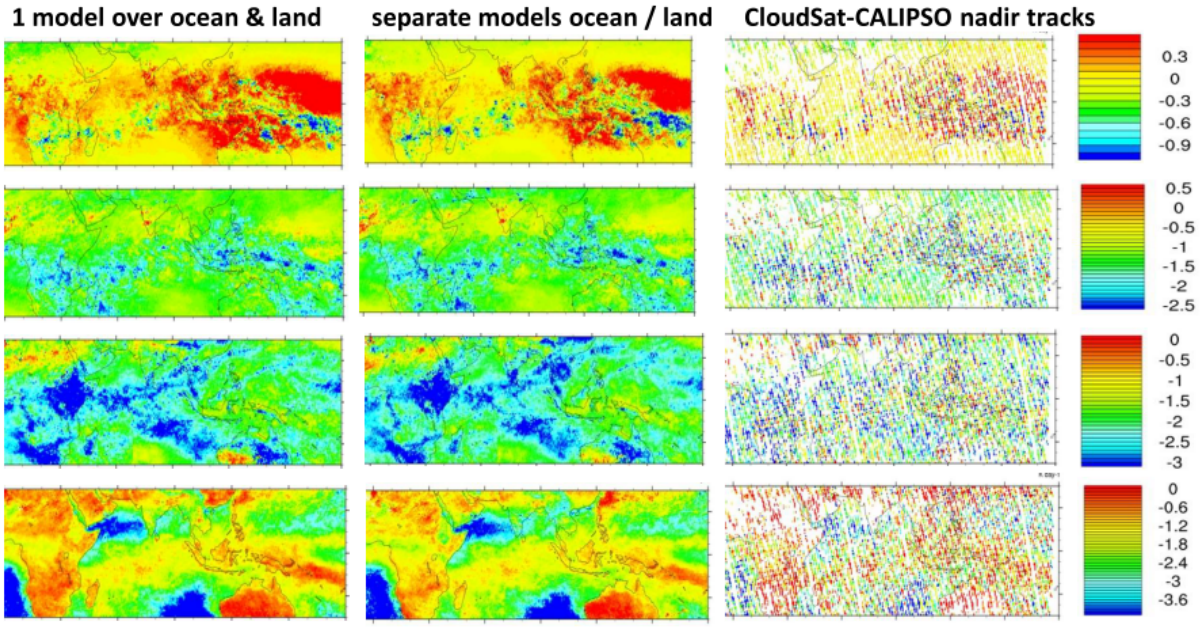


Fig. 4. Tropical map (30 °N to 30 °S) of the LW radiative heating rates at pressure levels of 106 hPa (first row), 200 hPa (second row), 525 hPa (third row) and 850 hPa (fourth row) for January 2008 (La Niña) at local time 1:30 PM. The first column shows the predictions of the model trained with all clouds over ocean and land together, the second column the predictions of the models trained with all clouds over ocean and land separately and the third column the monthly statistics obtained from the observations.

predicted by the two different models are very similar although slight differences occur, especially concerning the intensity of the heating. These differences are currently under investigation.

IV. Conclusions and Perspectives

We have shown for the first time that deep learning permits to relate the appropriate cloud and atmospheric properties from AIRS and ERA-Interim to the LW (and SW) radiative heating rate profiles, which are only given along the CALIPSO/CloudSat narrow nadir tracks, to laterally extend these heating rates. To improve our ANN models, the most suitable variable configuration has been investigated by conducting sensitivity studies on the parameters that govern the heating rates. The complete 3D radiative heating rate fields obtained for the AIRS 15-year time series together with the cloud system approach will allow detailed process and climate feedback studies. It is planned to use these fields to force a global climate model which will permit to investigate their influence on global circulation patterns.

ACKNOWLEDGMENTS

This work is supported by the Centre National de la Recherche Scientifique (CNRS) and the Centre National d'Etudes Spatiales (CNES). The authors thank the members of the AIRS, CALIPSO and CloudSat science teams for their efforts and cooperation in providing the data, as well as the engineers and space agencies who control the data quality.

REFERENCES

[1] S. Fueglistaler, A. E. Dessler, T. J. Dunkerton, I. Folkins, Q. Fu, and P. W. Mote, "Tropical tropopause layer," *Rev. Geophys.*, vol. 47, no. 1, 2009.

[2] T. S. L'Ecuyer and G. McGarragh, "A 10-year climatology of tropical radiative heating and its vertical structure from trmm observations," *J. Climate*, vol. 23, no. 3, pp. 519–541, 2010.

[3] C. J. Stubenrauch, A. G. Feofilov, S. E. Protopapadaki, and R. Armante, "Cloud climatologies from the infrared sounders AIRS and IASI: Strengths and applications," *Atmos. Chem. Phys.*, vol. 17, no. 22, pp. 13625–13644, 2017.

[4] X. Chen and X. Lin, "Big data deep learning: Challenges and perspectives," *IEEE Access*, vol. 2, pp. 514–525, 2014.

[5] G. L. Stephens, D. G. Vane, R. J. Boain, G. G. Mace, K. Sassen, Z. Wang, A. J. Illingworth, E. J. O'Connor, W. B. Rossow, S. L. Durden, S. D. Miller, R. T. Austin, A. Benedetti, C. Mitrescu, and the CloudSat Science Team, "The CloudSat mission and the A-Train: A new dimension of space-based observations of clouds and precipitation," *Bull. Amer. Meteor. Soc.*, vol. 83, pp. 1771–1790, 2002.

[6] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. K. U., Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kallberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K.Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart, "The era-interim reanalysis: configuration and performance of the data assimilation system," *Quart. J. Roy. Meteor. Soc.*, vol. 137, no. 656, pp. 553–597, 2011.

[7] S. E. Protopapadaki, C. J. Stubenrauch, and A. G. Feofilov, "Upper tropospheric cloud systems derived from IR sounders: Properties of cirrus anvils in the tropics," *Atmos. Chem. Phys.*, vol. 17, no. 6, pp. 3845–3859, 2017.

[8] C. J. Stubenrauch, A. Chédin, R. Armante, and N. A. Scott, "Clouds as seen by satellite sounders (3I) and imagers (ISCCP). Part II: A new approach for cloud parameter determination in the 3I algorithms," *J. Climate*, vol. 12, no. 8, pp. 2214–2223, 1999.

[9] W. B. Rossow, G. Tselioudis, A. Polak, and C. Jakob, "Tropical climate described as a distribution of weather states indicated by distinct mesoscale cloud property mixtures," *Geophys. Res. Lett.*, vol. 32, p. L21812, 2005.

[10] C. J. Stubenrauch and U. Schumann, "Impact of air traffic on cirrus coverage," *Geophys. Res. Lett.*, vol. 32, no. 14, p. L14813, 2005.

[11] T. S. L'Ecuyer, N. B. Wood, T. Haladay, G. L. Stephens, and P. W. Stackhouse Jr., "Impact of clouds on atmospheric heating based on the r04 cloudsat fluxes and heating rates data set," *J. Geophys. Res. Atmos.*, vol. 113, no. D8, 2008.

[12] D. S. Henderson, T. S. L'Ecuyer, G. L. Stephens, P. Partain, and M. Sekiguchi, "A multisensor perspective on the radiative impacts of clouds and aerosols," *J. Appl. Meteor. Climatol.*, vol. 52, no. 4, pp. 853–871, 2013.

[13] A. Chédin, S., Serrar, N. A. Scott, C. Crevoisier, and R. Armante, "First global measurement of midtropospheric co2 from noaa polar satellites: Tropical zone," *J. Geophys. Res. Atmos.*, vol. 108, no. D18, p. 4581, 2003.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

OPTIMAL SAMPLING OF TEMPERATURE ANOMALIES ON EARTH THROUGH SUPERVISED RECONSTRUCTION

Donà Jérémie[1], Arthur Pajot[1] , Patrick Gallinari[1,2], Sylvie Thiria [3]

***Abstract*—Sensor placement is a important problem in physics, climate and environmental science. In this paper we introduce this problem as a machine learning one, where we want to find sensor emplacements which maximize the reconstruction information of temperature map. The main challenges lie in the dimension of the considered problem and the desired properties of the sampling operation. Indeed, exploring all possible sets of points in a large dimension signal is computationally prohibitive. In order to tackle this challenge, we design a learning framework to leverage the representation and inference power of neural networks to derive a sampling and a reconstruction operator that minimizes both the number of sensors and the reconstruction loss. We show that this learning framework is helpful in climate science by associating sampling areas to climate natural variability modes. This allows us to produce an accurate reduced model of the temperature at the Earth's surface.**

I. INTRODUCTION

The task we aim to solve is the reconstruction of the temperature anomalies maps under a budgeted sensing constraint. Simply put, we want to learn two functions: 1) a sampling operator s that tells us where to sample the temperature anomalies 2) a reconstruction algorithm G that generate accurate temperature anomalies map from the measurements given by s. This problem is of crucial concern for few reasons. First, regarding the current situation towards climate change, monitoring the Earth's climate is of primary importance. In addition, a smarter sensing over the Earth's climate variables would allow a more variate use of satellites, reducing drastically the cost for satellite imaging based applications. Also, it could also encourage smarter measurements campaigns for example in meteorology. Moreover, as we will see later on, a well designed sampling scheme would also allow us to unveil physical phenomena, and reveal information and structure about the data. This double problem of sampling and reconstruction is challenging. Primarily, our problem is highly dimensional for if one has to place p sensors optimally in a $n \in \mathbb{N}$ dimensional space, it would have to explore $\binom{n}{p}$ possibilities, which is not desirable in our case of large temperature anomalies maps. Also, the temperature anomalies maps are highly structured images with complex patterns, so we want to use a deep learning prior to leverage its representation power. However, in order to solve this problem using neural network, we want to find a differentiable formulation of the sampling and reconstruction task. In this perspective, we consider that a sampling operator is a performing one if it allows a high level of reconstruction.

More formally, our task is to learn jointly a sampling operator s (a binary mask) that selects at most λ points and a reconstruction function G_ϕ, ϕ being the parameters the neural network G_ϕ, such that for all temperature anomalies maps x:

$$G_\phi(s \odot x) \approx x, \text{ such that, } ||s||_0 \leq \lambda \qquad (1)$$

Classically ℓ_0-norm is a natural idea. However, the ℓ_0-norm is non-differentiable and cannot be used as such in neural networks. Our contributions are as follows:

- We use a ℓ_1-norm formulation to relax equation (1) to make the problem differentiable.
- We propose a learning framework for the sampling operator s that uses a stochastic exploration scheme to search the space of masks.
- Our work is compared to an existing benchmark in order to validate our approach.
- We show that our algorithm is able to reconstruct the temperature anomalies from few selected measurements and prove that our reconstruction approach provides physically sound results.

Corresponding authors: Jérémie Donà, jeremie.dona@lip6.fr, arthur.pajot@lip6.fr [1] Sorbonne Universités, UMR 7606, LIP6, F-75005 Paris, France [2] Criteo AI Labs, [3] Sorbonne Universités, UMR 7159, LOCEAN, F-75005 Paris

II. METHOD

This work focuses on optimizing sampling points and the reconstruction of surface temperatures anomalies (difference between the climatology and observed temperatures). The temperatures anomalies maps denoted x results from an coupled ocean atmosphere model [1] and are treated as images.

A. Reconstruction

Let x be the temperature anomalies map we want to recover from partial observation x^{obs}. Let ϕ be the parameters of our reconstruction neural networks G_ϕ. We can write the reconstruction objective as a ℓ_2-norm minimization:

$$\min_\phi \mathcal{L}_2(\phi) = \min_\phi \mathbb{E}_{x^{obs},x} \ell_2(G_\phi(x^{obs}) - x) \quad (2)$$

The formulation that minimizes the ℓ_2 risk above is the conditional expectancy $\mathbb{E}(x|x^{obs})$. This optimum may induce blurry results if the conditioning is not informative enough, averaging among all possible sample of $x|x^{obs}$. Also, as we aim to find a sparse sampling, which makes the conditioning $x|x^{obs}$ less constrained, the drawbacks of the ℓ_2 risk minimization must be addressed. In order to do so, we take advantage of the conditional GAN formulation [2]. A GAN-like regularization tends to produce sharper and cleaner output, because at the optimum the generator samples from a distribution close to $p(x|x^{obs})$ [3].

Briefly speaking, a conditional-GAN has two components: the first one is a discriminator network with parameters ψ called D_ψ trained to differentiate "true" data labeled as 1 from data generated by G_ϕ labeled as 0. To do so, D_ψ is presented samples from the true and generated distribution (and conditional information). If it fails the classification tasks, its weights are updated based on a log loss. The second one is the generator, i.e reconstruction operator G_ϕ, that takes as input a random variable γ (and some conditional information such as a label or pixels) and aims at fooling D_ψ, making it classify the conditionally generated images as true. It leads to the cost function $\mathcal{L}_{cGAN}$:

$$\mathcal{L}_{cGAN}(\phi,\psi) = \mathbb{E}_{\gamma,x,x^{obs}} \log D_\psi(x, x^{obs}) + \mathbb{E}_{\gamma,x^{obs}} \log 1 - D_\psi(G_\phi(\gamma, x^{obs}), x^{obs}) \quad (3)$$

Moreover, if the sampling is very sparse, the observed anomalies x^{obs} might correspond to more than one full temperatures anomalies map x. We can model the superposition of possible states as a probability distribution, for which the GAN framework is well suited. Our reconstruction objective is a weighted sum of the supervision loss (2) and cGAN loss (3):

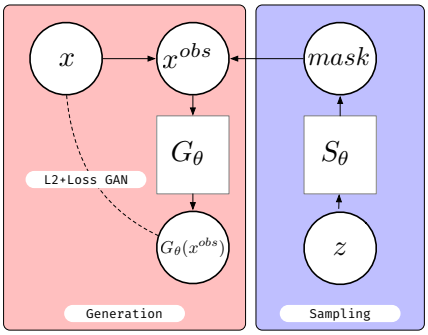


Fig. 1. Summary of the algorithmic flow, z is sampled to generate a mask through s_θ. The mask corrupts the true data. G_ϕ estimates a reconstructed map.

$$\min_\phi \max_\psi \mathcal{L}_{cGAN}(\phi,\psi) + \mu.\mathcal{L}_2(\phi) \quad (4)$$

Also, as shown in [4], adding a supervision (ℓ_1 or ℓ_2, second terms of the loss in equation 4), helps the algorithm to converge faster around the conditional expectancy (or median in case of ℓ_1-norm supervision).

B. Sparsity and masks exploration

As stated previously, the number of possible masks is very high and exploring all the states is computationally unfeasible. We tackle this issue by parametrizing the sampling operator s by a neural network s_θ taking as input a random variable z and the mask is defined as $mask = s_\theta(z)$ as described in figure (1). Indeed, we postulate that the variance of the random variable z allows the sampling operator s_θ to explore sufficiently the space of masks. Also, aiming to derive an optimal mask for the whole distribution p_x, we must ensure that $z \sim p_z$ and $x \sim p_x$ are not correlated. In order to do so, we sample our latent variable z according to a $\mathcal{N}(0, I_n)$ distribution. Regarding the cost function, we chose to constrain the ℓ_1-norm of the mask due to its thresholding behavior, enforcing sparsity. We define the cost function for the sampling operator as:

$$\min_\theta \mathbb{E}_{x,z}\ \mu.||G_\phi(s_\theta(z)\odot x) - x|| + \mathcal{L}_{cGAN} + \lambda.\ell_1(s_\theta)(z)$$

The formulation above formalizes the ambition to derive a sampling opertor that is both sparse (ℓ_1-norm term) and that leads to accurate and likely reconstruction ($\mathcal{L}_{cGAN}$ and $\mathcal{L}_2$ terms). The overall optimization problem is:

$$\min_{\phi,\theta} \max_\psi \mathcal{L}_{GAN}(\phi,\psi,\theta) + \mu\mathcal{L}_2(\phi,\theta) + \lambda\mathbb{E}_z\ell_1(s_\theta(z)) \quad (5)$$

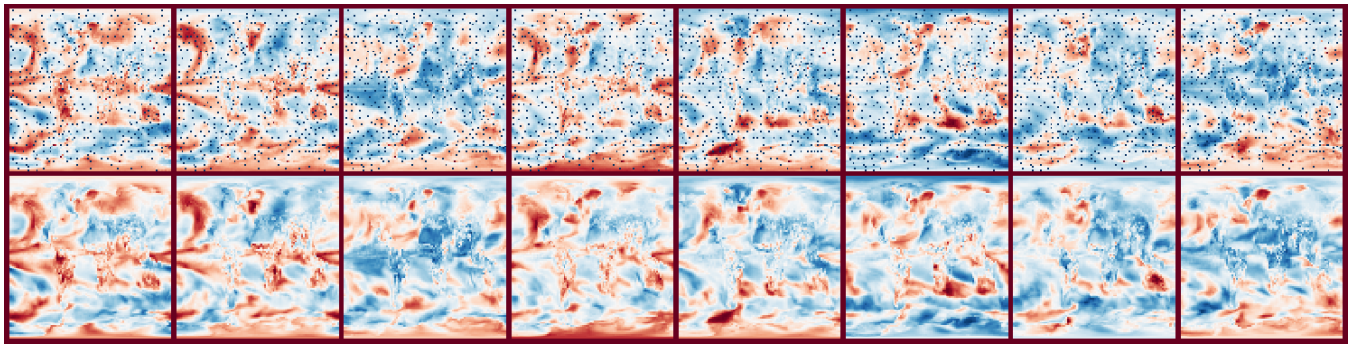


Fig. 2. First line: Reconstruction of the temperature anomalies. The darkest dots are the sampled points. Second line: Ground Truth. The hyperparameters for this simulation are: $\lambda = 1e-2$ (sparsity constraint) and $\mu = 100$ (ℓ_2 supervision constraint)

III. Optimization issues

The above formulation implies two issues that need to be dealt with: 1) calibrating the hyper parameters of the reconstruction and sparsity objectives, 2) inducing true zeros in the mask. Indeed, the major pitfall of our optimization problem is the following: let (s_θ, G_ϕ) be a solution for the optimization program (5) then $(s_\theta/2, G_\phi(2 * x))$ has the same reconstruction power with lower ℓ_1 norm. Therefore, our problem is ill posed and we need to constrain the norm of G_ϕ (we employ spectral norm introduced in [5]), and find a way to induce "true" zeros.

A. Setting the hyperparameters

One difficulty in our approach is the fact that the supervision and the sparsity constraints have opposite objectives. Let G be fixed for the moment and s be a vector, $s \in \mathbb{R}^n$ and G^i be the i^{th} function reconstructing pixel i in the image. Without the GAN loss the optimization program is:

$$\min_{s,\phi} \mathcal{L}(s;\phi) = \frac{1}{2}||G(s \odot x, \phi) - x||_2 + \lambda||s||_1 \quad (6)$$

We now investigate the direction of the gradient for the sampling operator:

$$\frac{\partial \mathcal{L}}{\partial s_j} = \frac{\partial}{\partial s_j}\frac{1}{2}\sum_{i=1}^{n}\{(G^i(s \odot x) - x^i)^2 + \lambda \sum_j |s_j|\}$$
$$= \underbrace{\sum_{i=1}^{n} x^j \frac{\partial G^i}{\partial x_j}(s \odot x)(G^i(s \odot x) - x^i)}_{\text{reconstruction gradient}} + \lambda$$

We can consider two simple cases in order to get intuition of the optimization process:

- The value of pixel j has a low impact on the reconstruction i.e. $\forall i \ \frac{\partial G^i}{\partial x^j} \approx 0$. This may occur due to redundancy of the information contained at pixel j. Then, the reconstruction gradient is zero λ term prevails and pushes s^j towards 0. Thus, λ needs to be high enough to induce sparsity.
- If the reconstruction gradient is negative, an increase on x^j decreases the error. Then the update with learning rate η would be $s^j \leftarrow s^j - \eta \ . \ (\lambda - |\text{Reconstruction gradient}|)$. Therefore, λ needs to be low enough to allow a good reconstruction.

Consequently, the magnitude of the λ must be carefully set in order to induce both a satisfying reconstruction and sparsity level. In order to do so, we initialize our sampling operator s_θ to generate masks very close to 1 and increase progressively the λ to enforce the ℓ_1 constraint. This allows the reconstruction network to learn easily in the beginning, and then discard progressively information that is not necessary or redundant for reconstruction.

B. Hard concrete trick

Despite the spectral regularization, we cannot ensure that our algorithm does not converge to a local minimum inducing a mask with uniformly low values. This is troublesome as it does neither define a probability density for sampling nor good sampling points. To induce true zeros and avoid this effect, we rely on [6] which introduces the hard concrete distribution. It is defined by: $u \sim \mathcal{U}(0,1)$, and $s = Sigmoid(\log u - \log 1 - u + \log \alpha)/\beta$, then $z = \min(1, \max(0, \lambda s + \gamma))$ is distributed in $[0,1]$ with two modes in 0 and 1 for $\lambda > 1$ and $\gamma < 0$. We apply this trick in order to obtain more binary outputs in our network. In our work, the uniform sampling on u is replaced by the output of our sampling operator s_θ (constrained in $[0,1]$).

s_θ, G_ϕ are neural networks which can be trained using back-propagation as the optimization program

is fully differentiable. We use ADAM optimization scheme with parameters $(0.7, 0.9)$ for G_ϕ and $(0.3, 0.7)$ for s_θ as it needs to explore more (reducing the first order momentum term).

IV. EVALUATION

We carried out our experiment on thanks to a coupled ocean-atmosphere model designed by IPSL [1]. We focus on the surface temperatures output of this model. De-trending, de-seasonalisation of the monthly temperature has been applied in order to obtain temperature anomalies. The temperature anomalies maps are 96×96 images. We compare our method with a baseline proposed in [7], that decomposes the signal in a POD (proper orthogonal decomposition) basis and uses a QR decomposition on the principal modes matrix to obtain the measurement matrix. The evaluation of our algorithm is twofold. We first compare the reconstruction power in terms of mean squared error. However, MSE might not be sufficient to ensure the quality and physical coherence of the reconstruction. Thus, we also use evaluation functions based on climate's intrinsic variability, inspecting physical indexes. Our reconstruction operator G_ϕ is a 64-filter 6-layers-deep convolutional Resnet like in [4], and the sampling operator s_θ is a one layer linear neural network, further architectures can be investigated in order to put a stronger prior on the mask.

A. Error Comparison

We use one major baseline described in [7]. As the baseline uses a PCA decomposition for reconstruction and sensor selection, we need to fix the number of modes i.e proper vectors for reconstruction. We refer as Linear180 Linear300 the baseline with respectively 180 and 300 modes (principal components of the PCA). We show the results for 1000 and 2000 selected sensors among the nearly 10^4 pixels of the image. In our case, the number of sensors will be a function of the value of λ (and μ).

As shown in table (I), the MSE is an order of magnitude lower for our model than for the baseline.

B. Correlation to physical indexes

	ENSO		TPI		AMO	
	CORR	**MSE**	**CORR**	**MSE**	**CORR**	**MSE**
Linear300-2	0.965	0.127	0.986	0.117	0.985	0.725
Ours-1	0.997	0.0019	0.998	0.0057	0.98	0.00094

TABLE II
CORRELATION COEFFICIENT AND MSE ON PHYSICAL INDEXES USING THE BEST BASELINE AND OUR ALGORITHM

	Parameters	**MSE**
Linear180-1	1000 points	0.0571
Linear180-2	2000 points	0.056
Linear300-1	1000 points	0.047
Linear300-2	2000 points	0.03579
Ours 1	$\lambda = 10^{-3}$, $\mu = 100$ ($\sim$ 500points)	0.00470
Ours 2	$\lambda = 10^{-2}$, $\mu = 100$ ($\sim$ 500points)	0.00536

TABLE I
MSE ON THE PROCESSED SURFACE TEMPERATURE ANOMALIES FOR VARIOUS BASELINES AND OUR ALGORITHM

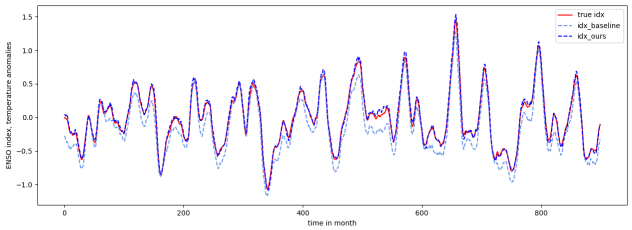


Fig. 3. Evolution of ENSO across time for the true data (red), our reconstruction method (dark blue) from $\sim$ 450 points $\lambda = 10^{-2}, \mu = 100 \sim 450$ points and the Linear300 (light blue) with 2000 points

In order to ensure have generated images that are physically sound, we compute 3 major indexes that represent oscillations of the Earth's surface temperature anomalies: 1) IPO (Interdecadal Pacific Oscillation) through the Tripole Index (TPI) [8], 2) AMO (Atlantic Multidecenal Oscillation) which explains strongly the temperature in the North Atlantic [9], 3) ENSO (El Nino Southern Oscillation) which plays a key role for both climate and ecosystems. As we can see in figure 3, the indexes induced by generated data follow the same patterns as the true ones with a correlation exceeding 0.99 for all indexes except for AMO. Figure (3) also show our performance in reconstructing ENSO index compared to the best baseline. We can conclude that our reconstruction algorithm enforces physics more efficiently than the baseline with fewer points.

C. Physical meaning of the masks

Finally, we investigate the physical significance of our mask.

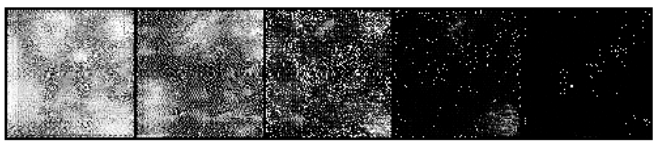


Fig. 4. Masks obtained for various λ, from left to right 10^{-5}, 2.10^{-5}, 4.10^{-5}, 10^{-4}, 10^{-3} with $\mu = 100$ inducing respectively 0.6%, 17%, 33%, 74%, 93% percent of zeros in the mask

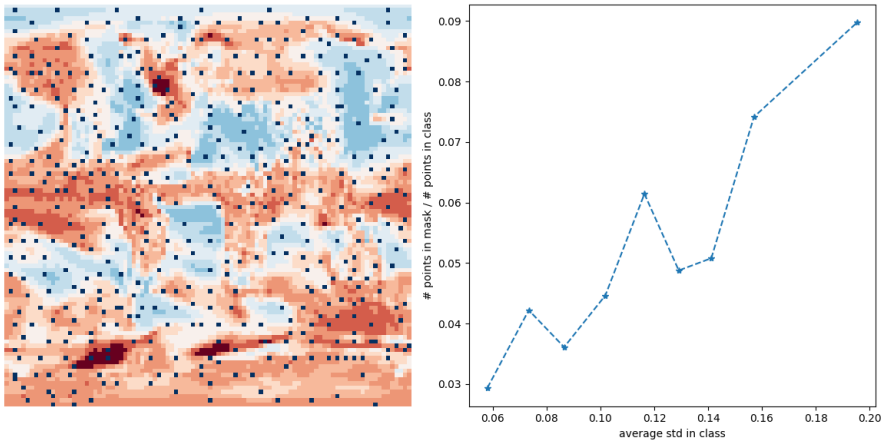


Fig. 5. Left panel: quantized variance of the temperatures anomalies and position of the sampled points (darkest dots) Right panel: evolution of the proportion of sampled points among each class

In the two firsts images of fig (4), we notice that the North-Atlantic and East-Pacific zones are very represented in the masks, which may explain the very encouraging results on physical indexes. As stated, physical indexes represent major variation modes in the temperatures signal. With increasing sparsity, we want to verify that our sampling algorithm selects pixels with the highest variance. In order to corroborate our hypothesis, we compute the standard deviation for each pixel and apply quantization on the obtained standard deviation for simplicity. We can observed on fig (5) that the pixel with the highest variance are more likely to be sampled with nearly 10% of them being sampled when less than 3% of low variance pixels are selected.

V. CONCLUSION

In this paper we presented a way to compute jointly a sampling and a reconstruction operator thanks to a differential formulation and a hard concrete trick. Thanks to the adversarial learning, the generated images have a strong physical sense. Inserting a constraint on where to place the sensors and formulate this problem without supervision are two leads of improvement of this work.

ACKNOWLEDGMENTS

This work was partially funded by ANR project LOCUST - ANR-15-CE23-0027 and by CLEAR - Center for LEArning & data Retrieval - joint lab. With Thales (www.thalesgroup.com)

REFERENCES

[1] IPSL, "Ipsl cma5.2 simulation." http://forge.ipsl.jussieu.fr/igcmg_doc/wiki/DocHconfigAipslcm5a2, 2018.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *arXiv:1406.2661 [cs, stat]*, June 2014. arXiv: 1406.2661.

[3] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," *arXiv:1606.03498 [cs]*, June 2016. arXiv: 1606.03498.

[4] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *arXiv:1611.07004 [cs]*, Nov. 2016. arXiv: 1611.07004.

[5] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral Normalization for Generative Adversarial Networks," *arXiv:1802.05957 [cs, stat]*, Feb. 2018. arXiv: 1802.05957.

[6] C. Louizos, M. Welling, and D. P. Kingma, "Learning Sparse Neural Networks through L_0 Regularization," *arXiv:1712.01312 [cs, stat]*, Dec. 2017. arXiv: 1712.01312.

[7] K. Manohar, B. W. Brunton, J. N. Kutz, and S. L. Brunton, "Data-Driven Sparse Sensor Placement for Reconstruction: Demonstrating the Benefits of Exploiting Known Patterns," *IEEE Control Systems Magazine*, vol. 38, pp. 63–86, June 2018.

[8] B. J. Henley, J. Gergis, D. J. Karoly, S. Power, J. Kennedy, and C. K. Folland, "A Tripole Index for the Interdecadal Pacific Oscillation," *Climate Dynamics*, vol. 45, pp. 3077–3090, Dec. 2015.

[9] M. E. Schlesinger and N. Ramankutty, "An oscillation in the global climate system of period 65–70 years," *Nature*, vol. 367, p. 723, Feb. 1994.

COMPUTING FLOOD PROBABILITIES USING TWITTER: APPLICATION TO THE HOUSTON URBAN AREA DURING HARVEY

Etienne Brangbour[1,2], Pierrick Bruneau[1], Stéphane Marchand-Maillet[2], Renaud Hostache[1], Marco Chini[1], Patrick Matgen[1], Thomas Tamisier[1]

***Abstract*—In this paper, we investigate the conversion of a Twitter corpus into geo-referenced raster cells holding the probability of the associated geographical areas of being flooded. We describe a baseline approach that combines a density ratio function, aggregation using a spatio-temporal Gaussian kernel function, and *TFIDF* textual features. The features are transformed to probabilities using a logistic regression model. The described method is evaluated on a corpus collected after the floods that followed Hurricane Harvey in the Houston urban area in August-September 2017. The baseline reaches a F1 score of 68%. We highlight research directions likely to improve these initial results.**

I. INTRODUCTION

The seminal way of predicting whether a point in space and time will be flooded is to simulate the water flow and runoff resulting from the expected rainfall. For instance, water flow is simulated using a Digital Elevation Model in Lisflood-FP [1]. More recently, an assimilation technique that allows to inject exogeneous observations, and adaptively update results of such simulations, was disclosed [2]. A 2D raster model of the region of interest is considered, with exogeneous probabilities of being flooded assigned to cells. The pixels in *Synthetic Aperture Radar* (SAR) satellite images, classified using a hierarchical split-based approach [3], have been used as such proxy observations in [2].

The usage of Twitter in the context of environmental hazards prevention and mitigation has been explored by several authors in the literature. For example, Twitter is used by Sakaki et al. [4] to help damage detection and reporting in the context of Earthquake events. The TAGGS platform aims at using Twitter for flood impact assessment at a worldwide scale [5]. Twitter has also been used to monitor the spread of the seasonal flu disease [6].

The objective of the present contribution is to perform and evaluate the conversion of a Twitter corpus into a map product analogous to that obtained from SAR images, for example. Data assimilation as described in [2] would then *a priori* benefit from such multiple sources. Our specific contributions are the following: a Gaussian spatio-temporal kernel function that effectively drives feature vector construction, the combination of heterogeneous geographical information (Twitter fields and 2D geo-referenced raster data), and the application to a real-world use case.

II. RELATED WORK

Many papers about using Twitter for improving the response to natural disasters mainly focus on analyzing the spatial dimension of a Twitter corpus. Geographical information is present in tweets in the form of *geotags* (i.e. discrete GPS coordinates) and *bounding boxes* (surface rectangles). Among these, geotags are the most accurate *a priori*, hence the most interesting. However, several studies report only about 1% of all tweets holding a geotag [7]. This proportion has also been observed in a corpus we collected in relation to the Hurricane Harvey use case [8]. In response to this lack of geographical information, several authors have focused on means to localize Twitter content. In [5], the authors focus on toponym detection and disambiguation. This is of prime importance in their context, as their system collects content from any place in the world. Named Entity Recognition (NER) adapted to Twitter is used to extract geographical entities, then combined in a resolution index table. Schulz et al. [9] combine geotags, NER results, bounding boxes, user information and emission time zones in a polygon stacking approach. In our work, we focus on a smaller area of interest, that limits the need for toponym resolution. Also, we analyzed

Corresponding author: P. Bruneau, pierrick.bruneau@list.lu
[1]Luxembourg Institute of Science and Technology
[2]University of Geneva

the geographical surfaces represented by the bounding boxes in a Twitter corpus collected for the Harvey use case, and concluded that approximately 17% of the tweets in the corpus have geographical information relevant for a flood event [8]. In the remainder of this paper, we focus on using this subset of the corpus as means to attach flood probabilities to cells in a 2D geographical raster model. The most immediate way of detecting flood relatedness is to check for the presence of pre-defined keywords in tweet text [7]. For example, a corpus related to Hurricane Harvey has been collected by detecting the keywords *Hurricane Harvey*, *#HurricaneHarvey*, *#Harvey* and *#Hurricane* during the event [10]. In the context of a flu spread analysis, [11] isolated a corpus matching a set of pre-defined keywords, and manually annotated 6500 tweets in order to train a SVM classifier that further filters out false positives. Individual tweets are then aggregated w.r.t. space and time using a density rate function. [12] saves the effort of individual tweet annotation by averaging tweets at city and day scale w.r.t. their *Term-Frequency-Inverse-Document-Frequency* (*TFIDF*) representation [13]. This is a continuous version of the binary bag-of-word model used in [11] for training a SVM classifier. A regression function linking observed rainfall to these feature vectors is then learned. In [6], the authors explicitly account for the time dynamics by fitting a state-space model to the likelihood Twitter users have to catch flu.

III. Proposed Model

Let us define the target variable y as a binary flooding indicator. We consider a constant spatio-temporal resolution, with spatial and temporal coordinates s and t, so that $y_{s,t} = 1$ if (s,t) is flooded, 0 otherwise. A given spatio-temporal cell is represented by its target variable $y_{s,t}$ and its feature vector $x_{s,t}$. These variables are linked according to the following logistic regression model:

$$p(y_{s,t} = 1|\phi_{s,t}) = \sigma(w^T \phi_{s,t}) \tag{1}$$

with $\sigma(a) = (1 + \exp(-a))^{-1}$ and $\phi_{s,t} = (1, x_{s,t}^T)^T$, therefore allowing for an intercept term in the adjustable weights vector w. For a given set of target values $\{y_{s,t}\}$ and their associated feature vectors, there is a single optimal value w^* to Eqn. (1), obtained by minimizing the following convex loss function [14]:

$$\mathcal{L}(w) = c\|w\|_1 - \sum_{s,t} y_{s,t} \ln \sigma(w^T \phi_{s,t}) + (1 - y_{s,t}) \ln(1 - \sigma(w^T \phi_{s,t})) \tag{2}$$

with c a positive regularization constant. The L1 penalty term ensures that the obtained solution is sparse, i.e. with as few non-zero coefficients in w as possible.

IV. Feature Vectors

In order to compute Eqn. (1) and minimize Eqn. (2), feature vectors $x_{s,t}$ have to be defined. Let us consider a collection of tweets $\mathcal{C}_N$:

$$\mathcal{C}_N = \left\{ \{s_n, d_n, t_n, \Omega_n\}_{n \in 1 \dots N} \right\} \tag{3}$$

where s_n is the spatial index of the n^{th} tweet, d_n its spatial dispersion (homogeneous to a standard deviation), t_n the temporal index of the tweet, and Ω_n the array of phrases in the tweet. The geographical information in tweets is composed of discrete geotags, as well as bounding boxes. The latter are abstracted by the dispersion variable in this section. The computation of this dispersion out of actual data is presented in Section V. In Gao et al., the authors define the scalar *Social Media Event Rate (SMER)* of a spatio-temporal cell $x_{s,t}^{SMER}$ as [11]:

$$x_{s,t}^{SMER} = \frac{\sum_{n=1}^{N} K(s, s_n, d_n) I(t, t_n) z_n}{\sum_{n=1}^{N} K(s, s_n, d_n) I(t, t_n)} \tag{4}$$

with K and I the spatial and temporal kernel functions, respectively. The random variable z_n equals 1 if Ω_n overlaps a query defined by the user, 0 else. In [11], the authors use the Epanechnikov kernel function to model the spatial coordinates. For an explicit account of the dispersion information, we rather use a Gaussian spatial kernel function:

$$K(s, s_n, d_n) = (2\pi d_n^2)^{-1} \exp(-\frac{1}{2}\|s - s_n\|_2^2). \tag{5}$$

The spatial L2 norm is computed by combining differences in latitude and longitude as independent dimensions. An indicator function is used as the temporal kernel: $I(t, t_n) = 1$ if $t = t_n, 0$ else. Alternatively, let us consider the union of all V phrases present in $\mathcal{C}_N$. Assuming an arbitrary order of these V tokens defines a V-dimensional feature space. The *term frequency* $\text{tf}_v(n)$ is simply the number of times token v appears in Ω_n. The *term frequency* and *inverse document frequency* of phrase v are, respectively:

$$\text{tf}_v(n) = \left|\{o = v | o \in \Omega_n\}\right|$$
$$\text{idf}_v = \ln(1 + N) - \ln(1 + N_v) + 1$$

with $N_v = \left|\{n \in 1 \dots N | \text{tf}_v(n) > 0\}\right|$ the document frequency of phrase v in $\mathcal{C}_N$. The *TFIDF* feature vector for tweet n is then defined as $\rho_n = (\text{tf}_v(n) \times \text{idf}_v)_{v \in V}$

[13]. As means to build a feature vector that characterizes a spatio-temporal raster cell, we adapt Eqn. (4) as follows:

$$x_{s,t}^{TFIDF} = \frac{\sum_{n=1}^{N} K(s, s_n, d_n) I(t, t_n) \rho_n}{\sum_{n=1}^{N} K(s, s_n, d_n) I(t, t_n)} \quad (6)$$

In practice, we replace the scalar query overlap variable z_n in Eqn. (4) with V-dimensional ρ_n. Performing this spatio-temporal aggregation is also a way to mitigate the fact that *TFIDF* vectors are very sparse in the case of tweets. $x_{s,t}^{TFIDF}$ instead yields denser representations, that are more easily handled by classification models in general - and Eqn. (1) in particular. Both $x_{s,t}^{SMER}$ and $x_{s,t}^{TFIDF}$ are tested in Section V.

V. EXPERIMENTS

A. Methodology

Hurricane Harvey has affected the Houston urban region between mid-August and mid-September 2017, with a peak around the 30th of August. A corpus of tweets collected for the Harvey event has been made available shortly after [10]. It features tweets matching any of the phrases *Hurricane Harvey*, *#HurricaneHarvey*, *#Harvey* and *#Hurricane*. Alternatively, using the Twitter APIs, we collected our own corpus of 7.5M tweets. In order to match the scope and objectives disclosed in the introduction, especially regarding content localization, we did not use textual query filters, and collected all content matching the spatial bounds of the Houston urban surroundings, and the temporal bounds mentioned above. The spatial area of interest has been determined according to prior analyses of Hurricane Harvey impacts [15]. Details about the corpus collection, pre-processing, and descriptive analysis are described at length in [8]. Figure 1 shows a flooding map of the Houston urban area for the 30th of August. It has been obtained by running the Lisflood-FP model [1]. The raster matrix displayed has been scaled to dimensions 1225×1450 pixels (i.e. approximately 1.78M pixels in total), with resolution 2.10^{-3° (approximately 240m) per pixel.

In this matrix, permanent water is indicated by a large value (i.e. 999). Hence we exclude permanent water pixels from our analysis by considering only pixels with water height less than 10m. This leaves a database of 1.47M pixels. Then, following [1], only pixels associated with water height greater than 0.2m are considered as flooded. This defines 80% non-flooded and 20% flooded pixels. The spatio-temporal binary target variable used in Eqn. (1) takes its ground truth values from these pixels. For our experiments,

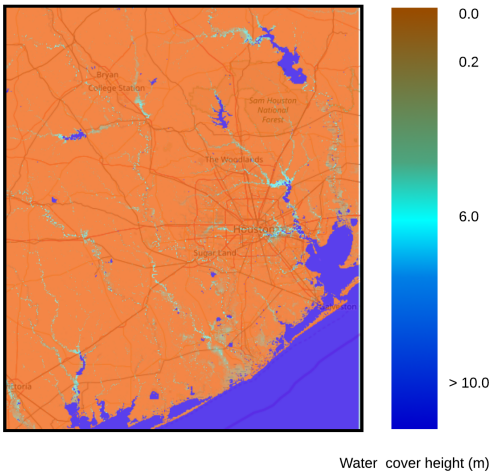


Fig. 1. Flooding map for the Houston urban area on 2017/08/30.

we consider daily temporal resolution, i.e. $I(t, t') = 1$ iif $\text{day}(t) = \text{day}(t')$ in Eqns. (4) and (6).

Our corpus features 319280 tweets matching the 30th of August. Among those, we further select 55214 tweets which hold geographical information sufficiently accurate to be retained in the analysis (see [8] for details about the selection method). We consider this subset as $\mathcal{C}_N$ (see Eqn. (3)). To build $x_{s,t}^{SMER}$ (Eqn. (4)) using this collection, we test the two following query filters: the keywords used in [10], and the keywords *flood* and *harvey*, that intuitively made sense to us. Rectangular bounding boxes are attached to tweets. The dispersion d_n in Eqn. (3) is computed as the square root of the product of the bounding box half-width and half-height. We smooth these values by setting $d_s = \max(d_s, 10^{-3})$ (in degrees, i.e. approximately 100m). Preprocessing is recommended prior to building *TFIDF* vectors [12]. First, URLs and user names are removed from the tweets. Then camel case words, commonly encountered on Twitter, are split as single words (e.g. *HurricaneHarvey* becomes *Hurricane Harvey*). Then punctuation and stop words are removed. Emoticons are not removed, and considered as words. Finally, we apply Porter stemming [16] to all words, that effectively normalizes the text (e.g. *flood*, *flooding* and *flooded* all become *flood*). Obtained feature vectors are normalized so that $\|\rho_n\|_2 = 1, \forall n$ (see Section IV for the definition of ρ_n).

Following the recommendations in [12], we consider the set of 1 and 2-grams as the V-dimensional *TFIDF* space. Simply put, dimensions in $x_{s,t}^{TFIDF}$ are associated to single words in the 1-gram model, whereas 2-word phrases are also included in the 2-gram model. This

Feature Vector	F1 score
SMER	
Hurricane Harvey, #HurricaneHarvey, ...[10]	0.58 ± 0.01
flood, harvey	0.59 ± 0.01
TFIDF	0.68 ± 0.01

TABLE I
TEST F1 ERRORS OBTAINED USING x^{SMER} AND x^{TFIDF}.

yields a very high-dimensional space (29K with 1-grams alone, 203K with 1 and 2-grams). Actually many of these tokens are present only a few times in $\mathcal{C}_N$: as suggested in [12], we retain features v with $N_v > 10$. This yields 3327 1-grams, and 1454 2-grams. The regularization hyper-parameter c in Eqn. (2) is optimized using 5-fold cross-validation. The loss is optimized using the SAGA solver [17]. For each feature vector type, the average F1 score shown in Table I has been computed according to 20 independent runs. For a given run, we have drawn balanced samples of 20K pixels as the training set in Eqn. (1), as well as balanced samples of 2K pixels as the test sets. Then feature vectors are built according to Eqn. (4) and (6). In particular, $x^{TFIDF}_{s,t}$ is renormalized to unity after aggregation.

B. Results

For our experimental setup (ie. 2 balanced classes for training and testing), the F1 score expected by chance is 0.5. In Table I, we see that *SMER* features yield moderate improvement over chance. We get better results using our intuitive keyword set *flood, harvey*, but the difference is not statistically significant. On the other hand, the *TFIDF* feature vector brings approximately 10% performance boost, well beyond significance levels. As we mentioned in Section V-A, the *TFIDF* vector has very high dimensionality, so we focus on identifying and inspecting its most relevant features for the classification problem at hand. First, as we used L1 penalty in Eqn. (2), we obtain a drastic reduction of the number of features used in the model. In the end, 896 features have non-zero weight in average, reduced to 247 if we take the median of our experiments. This is approximately 5% of the total number of features fed to the model.

The magnitude of weights in w can act as simple relevance scores of the respective features [18]. As means to aggregate our independent experimental runs (20), we first rank the features in decreasing relevance order. Then we normalize the scores in $[0, 1]$ using $(\kappa - \text{rank})/\kappa$, with κ the number of features selected in the run. We aggregate runs by averaging these normalized scores. If we consider only features present in all selected feature sets, we get the following features, ordered by decreasing score: *sad, fake, via, eye, today, oop, drive, dalla*. If we extend to the union of all features, the best 20 are: *sad, fake, basic, via, eye, today, told, true, flow, garbag, old, ah, flood, realli, peopl, hockley, learn, guadalup rv, final, work*. Manually inspecting the database, an example of tweet containing the word *sad* is *"mann, pray for my city houston, it's so sad seeing houston like dis!" sad* is then possibly a positive (though unexpected *a priori*) marker of flood relatedness. On the other hand, the token *fake* seems mostly related to fake news discussions, that could possibly be negatively correlated with the flood concept, i.e. people affected by the flood would have other concerns than talking about fake news. Some highly relevant tokens such as *flood* or *flow* make sense intuitively, where the role of others is hard to figure out (e.g. *eye* or *old*). Finally, it is worth noting that only one 2-gram appears in our highlights. This is quite natural, as 2-grams are much more rare than 1-grams on average, and have thus much less leverage on model fitting *a priori*. More striking is the absence of the token *harvey*. One possible explanation is that it used as much in general posts about the hurricane, as by people affected by floods. Its discriminative power would then be lessened.

VI. CONCLUSION

In this paper, we established a baseline for the performance in using Twitter posts to estimate whether a cell in a 2D raster grid is flooded. This baseline obtained a F1 score of 0.68, which is already very significantly better than chance, but also leaves much room for improvement. For this baseline, we use *TFIDF* features, which have recently been superseded by dense representations obtained from deep neural networks. For example, the latent vectors of a BiLSTM model trained for hashtag prediction have been used in the context of Twitter text [19]. However, averaging is not necessarily meaningful in such spaces, so simple stacking strategies such as presented in this paper may not apply directly. For future work, we will also consider the benefit from fields other than text and geographical information, such as the presence of an attached image [20], its content [21], or extracting geographical cues from text [7], [5].

VII. ACKNOWLEDGEMENTS

This work was performed in the context of the Publimape project, funded by the CORE programme of the Luxembourgish National Research Fund (FNR).

REFERENCES

[1] P. Bates and A. De Roo, "A simple raster-based model for flood inundation simulation," *Journal of hydrology*, vol. 236, no. 1-2, pp. 54–77, 2000.

[2] R. Hostache, G. Corato, M. Chini, M. Wood, L. Giustarini, and P. Matgen, "A new approach for improving flood model predictions based on the sequential assimilation of SAR-derived flood extent maps," in *EGU General Assembly Conference Abstracts*, vol. 17, 2015.

[3] M. Chini, R. Hostache, L. Giustarini, and P. Matgen, "A Hierarchical Split-Based Approach for Parametric Thresholding of SAR Images: Flood Inundation as a Test Case," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 6975–6988, 2017.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, "Tweet Analysis for Real-Time Event Detection and Earthquake Reporting System Development," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 4, pp. 919–931, 2013.

[5] J. de Bruijn, H. de Moel, B. Jongman, J. Wagemaker, and J. Aerts, "TAGGS: Grouping Tweets to Improve Global Geoparsing for Disaster Response," *Journal of Geovisualization and Spatial Analysis*, vol. 2, no. 1, 2017.

[6] L. Chen, T. H., P. Butler, N. Ramakrishnan, and B. Prakash, "Syndromic surveillance of Flu on Twitter using weakly supervised temporal topic models," *Data Mining and Knowledge Discovery*, vol. 30, no. 3, pp. 681–710, 2016.

[7] S. Middleton, L. Middleton, and S. Modafferi, "Real-Time Crisis Mapping of Natural Disasters Using Social Media," *IEEE Intelligent Systems*, vol. 29, no. 2, pp. 9–17, 2014.

[8] E. Brangbour, P. Bruneau, S. Marchand-Maillet, R. Hostache, P. Matgen, M. Chini, and T. Tamisier, "Extracting localized information from a Twitter corpus for flood prevention," *arXiv:1903.04748 [cs]*, 2019.

[9] A. Schulz, A. Hadjakos, H. Paulheim, J. Nachtwey, and M. Mühlhäuser, "A Multi-Indicator Approach for Geolocalization of Tweets," in *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, pp. 1–10, 2013.

[10] J. Littman, "Hurricanes Harvey and Irma Tweet ids," 2017. https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/QRKIBW.

[11] Y. Gao, S. Wang, A. Padmanabhan, J. Yin, and G. Cao, "Mapping spatiotemporal patterns of events using social media: a case study of influenza trends," *International Journal of Geographical Information Science*, vol. 32, no. 3, pp. 425–449, 2018.

[12] V. Lampos and N. Cristianini, "Nowcasting Events from the Social Web with Statistical Learning," *ACM Trans. Intell. Syst. Technol.*, vol. 3, no. 4, pp. 1–22, 2012.

[13] T. Joachims, "Text categorization with Support Vector Machines: Learning with many relevant features," in *ECML-98*, pp. 137–142, 1998.

[14] S. Lee, H. Lee, P. Abbeel, and A. Ng, "Efficient L1 Regularized Logistic Regression," in *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, pp. 401–408, AAAI Press, 2006.

[15] P. Matgen, R. Pelich, E. Brangbour, P. Bruneau, M. Chini, R. Hostache, G. Schumann, and T. Tamisier, "Integrating Data Streams from in-situ Measurements, Social Networks and Satellite Earth Observation to Augment Operational Flood Monitoring and Forecasting: the 2017 Hurricane Season in the Americas as a Large-scale Test Case," *AGU Fall Meeting Abstracts*, vol. 31, 2017.

[16] M. Porter, "An algorithm for suffix stripping," *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[17] A. Defazio, F. Bach, and S. Lacoste-Julien, "SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives," *arXiv:1407.0202 [cs, math, stat]*, 2014. arXiv: 1407.0202.

[18] Y. LeCun, J. Denker, and S. Solla, "Optimal Brain Damage," in *Advances in Neural Information Processing Systems 2*, pp. 598–605, 1990.

[19] B. Dhingra, Z. Zhou, D. Fitzpatrick, M. Muehl, and W. Cohen, "Tweet2vec: Character-Based Distributed Representations for Social Media," *arXiv:1605.03481 [cs]*, 2016.

[20] R. Peters and J. de Albuquerque, "Investigating images as indicators for relevant social media messages in disaster management," in *Proceedings of the ISCRAM 2015 Conference*, 2015.

[21] B. Bischke, P. Helber, C. Schulze, V. Srinivasan, A. Dengel, and D. Borth, "The Multimedia Satellite Task at MediaEval 2017," in *MediaEval 2017*, 2017.

MARINE COLD AIR OUTBREAKS: PREDICTION SKILL AND PRECONDITIONS

Iuliia Polkova[1], Hilla Afargan-Gerstman[2], Daniela Domeisen[2], Paolo Ruggieri[3], Panos Athanasiadis[3], Martin King[4], Johanna Baehr[1]

***Abstract*—Marine cold air outbreaks (MCAOs) create conditions for hazardous maritime cyclones known as polar lows, which pose risks to marine infrastructure. For marine management, MCAO predictions would be highly beneficial. Previous studies explain the genesis of MCAOs, while predictability and large-scale causal drivers of MCAOs remain largely unstudied.**

We investigate (i) the ability of the Earth System Model from the Max-Planck Institute for Meteorology (MPI-ESM) to predict MCAOs and (ii) options to improve predictability of MCAOs through their large-scale causal drivers. To identify MCAO preconditions, we utilize the atmospheric reanalysis ERA-Interim in the lagged cross-correlation analysis, composite analysis, and causal effect network (CEN).

The results show that the MPI-ESM has high prediction skill for MCAOs over the Barents Sea (BS), Greenland-Iceland-Norwegian Seas and the Labrador Sea for about 2.5 weeks ahead starting from the November initial conditions. Whereas the lagged cross-correlation analysis indicates relationship between the early-fall atmospheric and sea-ice conditions and the late-fall BS-MCAO index, CEN identifies the causal link only from the Arctic sea ice concentration (SIC).

I. MOTIVATION

Polar lows (PLs) have intensively been studied over the past sixty years [1], [2], [3]. They can be predicted by the nowcasting systems (e.g., BarentsWatch portal www.barentswatch.no/en/polar-low/). However, for planning of marine activities, there is a demand for information about conditions favoring the development of PLs beyond the time horizon that nowcasting can offer [4]. Important conditions for a development of PLs, which can serve as a proxy for PLs, are pronounced and sustained transports of extremely cold air over an ice-free ocean referred to as marine cold air outbreaks (MCAOs) [2], [5].

In contrast to PLs, MCAOs are relatively long-term and large-scale phenomena (PLs: about 1-2 days, 200-1000 km in diameter; MCAOs: about 1-2 weeks, $>$ 1000 km). Due to their large time-space resolution, MCAOs are thought to be predictable for time horizons beyond few days as compared to PLs [6]. Moreover, several recent studies showed that combining statistical and dynamical prediction techniques, it was possible to improve the prediction skill for important climate indices as compared to the skill of dynamical systems alone. In particular in [7], a statistical model based on the autumn North Atlantic Oscillation (NAO) predictors was developed to improve winter NAO predictions from a seasonal prediction system. In [8], cold temperature extremes over land were predicted one month ahead by means of a statistical model, which was based on the key indices for the stratospheric circulation as predictors that were obtained from the seasonal prediction system. Thus, if links between MCAOs and their preconditions are established, this potentially could be used in a hybrid statistical-dynamical prediction system to improve predictability of MCAOs.

Previously, neither prediction skill from subseasonal-to-seasonal (S2S) prediction systems, nor statistical models based on MCAO preconditions were reported. Assessing predictability of MCAOs and establishing causal links between MCAOs and their drivers are our research goals within the EU Horizon 2020 Project Blue Action WP1 (www.blue-action.eu). Some of our results in this direction are described in the current study. In particular, we present results on (i) a prediction skill for MCAOs in regions of highest frequency occurrence of MCAOs, namely in the North Atlantic and the Arctic sectors [9], and (ii) identification of causal relationships between MCAOs and large-scale variability of the sea-ice and the atmosphere, which latter can serve as predictors in a hybrid statistical-

Corresponding author: I Polkova, iuliia.polkova@uni-hamburg.de, [1]Institute of Oceanography, Universität Hamburg, CEN, Hamburg, Germany [2]Institute for Atmospheric and Climate Science, ETH Zürich, Zurich, Switzerland [3]Euro-Mediterranean Center on Climate Change - CMCC, Bologna, Italy [4]NORCE Climate, and Bjerknes Centre for Climate Research, Bergen, Norway

dynamical prediction for MCAOs.

II. METHOD

A. Data and Model

To assess prediction skill for MCAOs, we use a 30-member ensemble of seasonal retrospective predictions (re-forecasts) [10]. An atmospheric reanalysis ERA-Interim [11] is used as a verification data set for the prediction skill assessment and the analysis of MCAO preconditions. The ensembles of re-forecasts are started yearly on November 1 over the period 1980-2016 and are generated by perturbing initial conditions with bred-vectors. The re-forecasts are carried out with the seasonal prediction system that is based on the Max Planck Institute for Meteorology Earth System Model with mixed resolution (MPI-ESM-MR). The model consists of the atmospheric component ECHAM6 with a resolution of T63L95 and the oceanic component MPIOM with 0.4° horizontal resolution and 40 vertical levels. The re-forecasts start from an assimilation run, in which the full fields of the ERA-Interim atmospheric state and the ORAS4 ocean state [12] are nudged into the MPI-ESM-MR [10].

B. MCAO index

The air temperature over the sea ice can fall as low as -40°C, while the temperature over the sea is close to freezing. This creates a strong temperature gradient across the sea-ice edge. Transports of cold air masses from the sea ice toward the ocean create a large vertical temperature gradient, which leads to surface sensible and latent heat-flux release from the ocean. In such conditions, the convective boundary layer emerges. The MCAO index measures atmospheric instability condition that causes convection within the boundary layer over the open ocean [6]. The index is thus obtained from the daily air-sea potential temperature difference

$$\text{MCAO index} = \theta_{SKT} - \theta_{850}, \qquad (1)$$

where θ_{SKT} is potential ocean skin temperature or potential sea surface temperature and θ_{850} is potential air temperature at 850 hPa. For the calculation of potential temperatures, a 1000 hPa level is used as a reference.

The MCAO index only accounts for values over the ocean, thus land grid-points are masked. The positive (> 0 K) potential temperature difference suggests atmospheric instability condition. Further thresholds can be set to discriminate between weak (< 4 K), moderate (4 to 8 K) and strong (> 8 K) conditions as in [13]. We focus on the MCAO index in the late fall to early spring months (November to March) in the North Atlantic and the Arctic sectors.

C. Prediction skill assessment

Prediction skill is assessed by systematically comparing counterpart re-forecasts and observations at each forecast time (lead time) and identifying the margin, at which the re-forecasts start to disagree with observations. The skill can be assessed using (i) deterministic metrics such as correlation coefficients and root-mean square errors and (ii) probabilistic metrics such as relative operating characteristics (ROC; see [14], [15]). Here, we only provide the ROC skill assessed against ERA-Interim that allows to answer the question how fast re-forecasts forget about their initial state.

The ROC skill as function of lead time suggests how well and how long the model can discriminate between MCAO events and non-events. This metric is not sensitive to the amplitude of events. To generate the ROC curve, a set of hit rates (forecast probability of correctly predicted events) is plotted against false-alarm rates (forecast probability of incorrect warnings). The sets of rates are obtained by binning forecast probabilities, which are calculated for each time step over the verification period, by ensemble confidence in the event occurrence (see more details on the ROC metric in [15]). For a good prediction skill, it is expected that ensemble predictions lay over the no-skill diagonal line, thereby showing high hit rates as compared to false alarms. The area under the ROC curve is used as a skill score and for skillful re-forecasts should be > 0.5.

D. Methods to identify/confirm MCAO predictors

Previous studies on relationship between MCAOs and atmospheric circulation indices employed composite and cross-correlation analyses [16], [9], [17]. We also use them, however we are aware that these metrics do not provide information about causality [18]. The recent study [19] on causal links between the mid-latitude winter circulation and its potential drivers gives a promising perspective to amend the analysis on MCAO predictors.

We follow the analysis in [19] that is based on the Causal Effect Network (CEN) algorithm provided by J.Runge (https://github.com/jakobrunge/tigramite) [18]. The CEN algorithm encompasses the following steps: (1) the evaluation of time lagged cross-correlation between potential predictors and the MCAO index (predictand) and between predictors themselves, as well

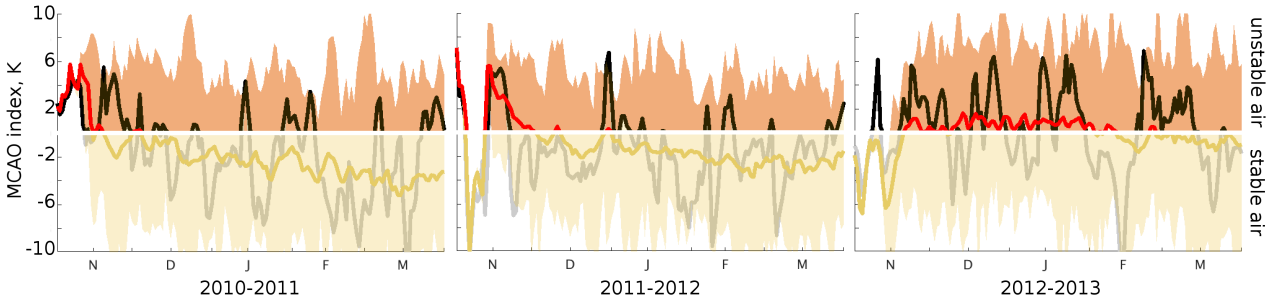


Fig. 1. An example of time series of the daily MCAO index (K) averaged over the BS (68°N–80°N,20°E–50°E) from November 2010 to March 2013 from ERA-Interim in grey and black and the ensemble mean re-forecasts in yellow and red. Positive (red and black) and negative (yellow and grey) values of the index correspond to unstable and stable air conditions, respectively (see section II B). The ensemble range is shown in shading.

as the evaluation of auto-correlation in predictors and predictands; (2) the iterative evaluation of conditional independence between predictors and predictands at given lag accounting for an effect from other predictors. This is done by analyzing partial correlations in the iterative algorithm that is a version of the Markov discovery algorithm. Here, each variable is tested in the role of a predictand with other variables acting as its predictors. If the partial correlation is significant, then there is a direct link between the predictand and the predictor at a given lag. At this step, a refined set of MCAO predictors is obtained. (3) Testing significance of causal links using multi-linear regression models [18].

III. EVALUATION

A. Prediction skill for MCAOs from seasonal re-forecasts

The re-forecasts for the Barents Sea (BS) MCAO index at the beginning have little spread (Fig. 1), however, resemble ERA-Interim rather well. The ROC skill confirms that within about 2.5 weeks after initialization in November; the ensemble re-forecasts show a high ratio of hit rates versus false alarms for the BS-MCAO events, which are defined by the positive values of the MCAO index that stay longer than 2 days (Fig. 2 A). The re-forecasts outperform the baseline damped persistence (from the AR(1) model) by higher value of hit rates. The damped persistence shows a good ROC skill for about a week, within the second week the signal is damped. The ROC skill scores as indicated by the areas under the ROC curves in Fig. 2 B suggest similar length of prediction skill for other regions of high MCAO frequency occurrences, namely for the Greenland-Iceland-Norwegian (GIN) Seas and the Labrador Sea. The November-initialized seasonal re-forecasts also show good skill for winter atmospheric circulation [7] and monthly sea-ice concentration [20].

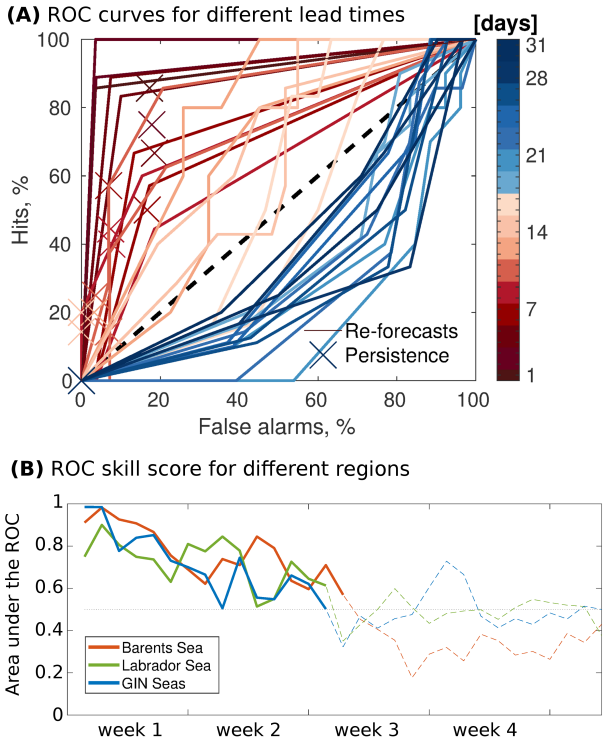


Fig. 2. The MCAO prediction skill: (A) The ROC for the BS-MCAO events (MCAO index $>$ 0 K and $>$ 2 days). Color-coded ROC curves (markers) represent different lead days of the MCAO re-forecasts (damped persistence forecasts). (B) The ROC skill score as function of lead time for the BS (red solid and dashed), the GIN Seas (blue solid and dashed) and the Labrador Sea (green solid and dashed). Solid curves indicate "useful" ROC skill that stays above no-skill line (0.5).

B. Identifying MCAO preconditions

To date, there is little known about large-scale causal drivers of MCAOs. Several studies suggest instantaneous links to local winds [6], the NAO and blocking patterns, which are depending on the phase of the circulation index and the MCAO region suggest changes in the MCAO activity [9], [17], and the upstream sea-ice conditions that might control MCAO properties [21], [22]. The other mechanisms, which were not directly linked to MCAOs but are relevant for high-latitude climate and thus might affect the MCAO-variability, are variability of the meridional ocean temperature gradient and the Arctic sea-ice that affects atmospheric dynamics [23] and the stratosphere that has a downward influence on surface temperatures in polar regions [24], [25]. Thus, in the current setting, we test relationships between the BS-MCAO index and the NAO index, the polar jet stream (JS) position, the Arctic sea-ice concentration (SIC), stratospheric temperature at 100 hPa (T100), BS sea surface temperature (SST) and surface air temperature (SAT).

For all geophysical fields but the well-established indices such as the NAO and the JS [26], we need to specify regions, over which the indices describing the stratospheric and the sea-ice states should be calculated. The outcome of the preconditions analysis can be sensitive to this choice. Following [7], who defined regions of significant lagged cross-correlation between a predictand and predictors, in our example, such regions are obtained from the cross-correlation between September-October T100, SAT, SST, SIC and November-December BS-MCAO index (all but SST are shown in Fig. 3). The indices for T100, SAT, SST and SIC are then obtained by averaging variables over the corresponding regions with significant positive cross-correlation. [7] considered indices obtained in this way as a "first guess" for the NAO and used to select ensemble members that captured the right phase of the NAO as suggested by the "first guess". The "first guess" for the November BS-MCAO index in terms of the September Arctic SIC index is shown in Fig. 3.

As lagged cross-correlation does not imply causal relationship, we further test our potential predictions based on the NAO, JS, SAT, SST, SIC and T100 using CEN. The time series are based on the 10-day averages – the intervals within a "skillful period" for MCAOs and contain anomalies of detrended fields calculated with respect to a 1980-2016 climatology. With respect to BS-MCAO drivers, the CEN analysis confirms only one positive causal link between the Arctic SIC at lag 6 (September 1-10) and the BS-MCAO index (November 1-10). At significance level $\alpha = 0.01$, the strength of coupling is 0.199 and partial correlation is 0.146.

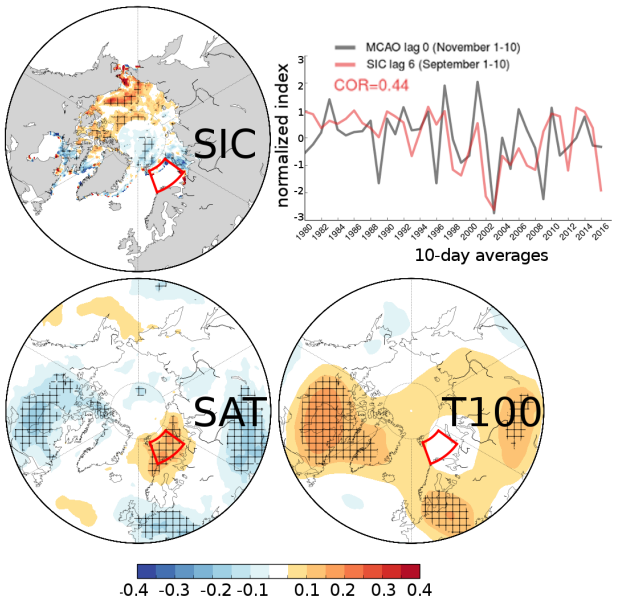


Fig. 3. Lagged cross-correlation between the November-December MCAO index averaged over the BS (red box) and September-October SIC (upper left), SAT (lower left) and T100 (lower right). Hatching indicates significant correlation values estimated at 10 % level using the bootstrap method. The time series show: (red) the SIC index at lag 6 (September 1-10) averaged over the region of significant positive cross-correlation from the upper-left panel and (black) the BS-MCAO index at November 1-10 over 1980-2016.

IV. Conclusions

For a statistical-dynamical approach in predicting MCAOs, we need to establish links between the environment conditions (atmosphere, ocean and sea ice) and MCAOs, and evaluate an ability of the prediction system to predict MCAOs and its preconditions. Along these lines, the prediction system shows high skill for MCAOs on the sub-seasonal timescale, which supports the hypothesis by [6]. In terms of MCAO drivers, the stratospheric state and the NAO, which were suggested in [6] as potential predictors, and the other considered indices though have lagged links, but they do not seem to be conditionally independent. The only confirmed causal link so far appears to be between the September Arctic SIC and the November BS MCAO index which suggests that a decrease in SIC could cause a decrease in the MCAO index, which could further mean a stable atmospheric condition and a decreased frequency of MCAOs (and vice versa for the increase in SIC). The identified SIC predictor is not sufficient for the statistical-dynamical approach. In a next step, we plan to append the analysis with indices describing local

wind components, blocking indices, air-sea fluxes, and the upper ocean state. Further, we would like to confirm existence of these links in the prediction system. The CEN represents a useful complementary tool to test the hypothesized predictors. Finally, we would like to assess predictions and predictors throughout the whole fall-winter-spring season to be able supporting the maritime services. Presented are the first results in this direction.

ACKNOWLEDGMENTS

Funding was provided by the Blue-Action project from the European Union's Horizon 2020 research and innovation programme under grant agreement No 727852, and by the Swiss National Science Foundation to D.D. through project PP00P2 170523. We thank Lara Hellmich for providing the JS time series, Jacob Runge for the tigramite python package and David Nielsen for polarplots python package.

REFERENCES

[1] D. Harley, "Frontal contour analysis of a polar low," *Meteor. Mag*, vol. 89, pp. 146–147, 1960.

[2] E. Rasmussen and J. Turner, *Polar lows: mesoscale weather systems in the polar regions*. Cambridge University Press, 2003.

[3] T. Spengler, C. Claud, and G. Heinemann, "Polar low workshop summary," *Bulletin of the American Meteorological Society*, vol. 98, no. 6, pp. ES139–ES142, 2017.

[4] A. P. Orimolade, O. T. Gudmestad, and L. E. Wold, "Vessel stability in polar low situations," *Ships and Offshore Structures*, vol. 12, no. sup1, pp. S82–S87, 2017.

[5] O. A. Landgren, I. A. Seierstad, and T. Iversen, "Projected future changes in marine cold-air outbreaks associated with polar lows in the northern North-Atlantic Ocean," *Climate Dynamics*, pp. 1–13, 2019.

[6] E. W. Kolstad, "Higher ocean wind speeds during marine cold air outbreaks," *Quarterly Journal of the Royal Meteorological Society*, vol. 143, no. 706, pp. 2084–2092, 2017.

[7] M. Dobrynin, D. I. Domeisen, W. A. Müller, L. Bell, S. Brune, F. Bunzel, A. Düsterhus, K. Fröhlich, H. Pohlmann, and J. Baehr, "Improved teleconnection-based dynamical seasonal predictions of boreal winter," *Geophysical Research Letters*, vol. 45, no. 8, pp. 3605–3614, 2018.

[8] M. Cai, Y. Yu, Y. Deng, H. M. van den Dool, R. Ren, S. Saha, X. Wu, and J. Huang, "Feeling the pulse of the stratosphere: An emerging opportunity for predicting continental-scale cold-air outbreaks 1 month in advance," *Bulletin of the American Meteorological Society*, vol. 97, no. 8, pp. 1475–1489, 2016.

[9] E. W. Kolstad, T. J. Bracegirdle, and I. A. Seierstad, "Marine cold-air outbreaks in the North Atlantic: temporal distribution and associations with large-scale atmospheric circulation," *Climate dynamics*, vol. 33, no. 2-3, pp. 187–197, 2009.

[10] J. Baehr, K. Fröhlich, M. Botzet, D. I. Domeisen, L. Kornblueh, D. Notz, R. Piontek, H. Pohlmann, S. Tietsche, and W. A. Müller, "The prediction of surface temperature in the new seasonal prediction system based on the MPI-ESM coupled climate model," *Climate Dynamics*, vol. 44, no. 9-10, pp. 2723–2735, 2015.

[11] D. P. Dee, S. Uppala, A. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. Balmaseda, G. Balsamo, P. Bauer, *et al.*, "The ERA-Interim reanalysis: Configuration and performance of the data assimilation system," *Quarterly Journal of the royal meteorological society*, vol. 137, no. 656, pp. 553–597, 2011.

[12] M. A. Balmaseda, K. Mogensen, and A. T. Weaver, "Evaluation of the ECMWF ocean reanalysis system ORAS4," *Quarterly Journal of the Royal Meteorological Society*, vol. 139, no. 674, pp. 1132–1161, 2013.

[13] L. Papritz and T. Spengler, "A lagrangian climatology of wintertime cold air outbreaks in the Irminger and Nordic Seas and their role in shaping air–sea heat fluxes," *Journal of Climate*, vol. 30, no. 8, pp. 2717–2737, 2017.

[14] I. T. Jolliffe and D. B. Stephenson, *Forecast verification: a practitioner's guide in atmospheric science*. John Wiley & Sons, 2012.

[15] S. J. Mason and N. E. Graham, "Conditional probabilities, relative operating characteristics, and relative operating levels," *Weather and Forecasting*, vol. 14, no. 5, pp. 713–725, 1999.

[16] E. W. Kolstad and T. J. Bracegirdle, "Marine cold-air outbreaks in the future: an assessment of IPCC AR4 model results for the Northern Hemisphere," *Climate Dynamics*, vol. 30, no. 7-8, pp. 871–885, 2008.

[17] L. Papritz and C. M. Grams, "Linking low-frequency large-scale circulation patterns to cold air outbreak formation in the northeastern North Atlantic," *Geophysical Research Letters*, vol. 45, no. 5, pp. 2542–2553, 2018.

[18] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting causal associations in large nonlinear time series datasets," *arXiv preprint arXiv:1702.07007*, 2017.

[19] M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge, "Using causal effect networks to analyze different Arctic drivers of midlatitude winter circulation," *Journal of Climate*, vol. 29, no. 11, pp. 4069–4081, 2016.

[20] F. Bunzel, D. Notz, J. Baehr, W. A. Müller, and K. Fröhlich, "Seasonal climate forecasts significantly affected by observational uncertainty of Arctic sea ice concentration," *Geophysical Research Letters*, vol. 43, no. 2, pp. 852–859, 2016.

[21] B. Brümmer, "Boundary-layer modification in wintertime cold-air outbreaks from the Arctic sea ice," *Boundary-Layer Meteorology*, vol. 80, no. 1-2, pp. 109–125, 1996.

[22] M. Adakudlu and I. Barstad, "Impacts of the ice-cover and sea-surface temperature on a polar low over the Nordic seas: a numerical case study," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 660, pp. 1716–1730, 2011.

[23] J. Cohen, J. A. Screen, J. C. Furtado, M. Barlow, D. Whittleston, D. Coumou, J. Francis, K. Dethloff, D. Entekhabi, J. Overland, *et al.*, "Recent Arctic amplification and extreme mid-latitude weather," *Nature geoscience*, vol. 7, no. 9, p. 627, 2014.

[24] J. Kidston, A. A. Scaife, S. C. Hardiman, D. M. Mitchell, N. Butchart, M. P. Baldwin, and L. J. Gray, "Stratospheric influence on tropospheric jet streams, storm tracks and surface weather," *Nature Geoscience*, vol. 8, no. 6, p. 433, 2015.

[25] M. Cai and R. Ren, "Meridional and downward propagation of atmospheric circulation anomalies. Part I: Northern Hemisphere cold season variability," *Journal of the atmospheric sciences*, vol. 64, no. 6, pp. 1880–1901, 2007.

[26] T. Woollings and M. Blackburn, "The North Atlantic jet stream under climate change and its relation to the NAO and EA patterns," *Journal of Climate*, vol. 25, no. 3, pp. 886–902, 2012.

ATTRIBUTION OF MULTIVARIATE EXTREME EVENTS

Yanira Guanche García[1,3], Maha Shadaydeh[2], Miguel Mahecha[3,4], Joachim Denzler[2,3]

***Abstract*—The detection of multivariate extreme events is crucial to monitor the Earth system and to analyze their impacts on ecosystems and society. Once an abnormal event is detected, the following natural question is: what is causing this anomaly? Answering this question we try to understand these anomalies, to explain why they happened. In a previous work, the authors presented a multivariate anomaly detection approach based on the combination of a vector autoregressive model and the Mahalanobis distance metric. In this paper, we present an approach for the attribution of the detected anomalous events based on the decomposition of the Mahalanobis distance. The decomposed form of this metric provides an answer to the question: how much does each variable contribute to this distance metric? The method is applied to the extreme events detected in the land-atmosphere exchange fluxes: Gross Primary Productivity, Latent Energy, Net Ecosystem Exchange, Sensible Heat and Terrestrial Ecosystem Respiration. The attribution results of the proposed method for different known historic events are presented and compared with the univariate Z-score attribution method.**

I. INTRODUCTION

The detection of multivariate extreme events is crucial to monitor the Earth system and to analyze their impacts on ecosystems and society. We expect that climate extremes such as droughts and heatwaves will increase as a consequence of climate change [1]. Hence, it is of a paramount importance to understand the drivers of such multivariate extreme events as well as the complex land-atmosphere-biosphere interactions, including those constellations that are not extreme for a single variable but are extreme for a combination of variables, also called compound event [1], [2]. In the last years several studies have gone in this direction, proposing different approaches: [2], [3], or [4] are just some examples. Unlike the different attribution methods proposed in the literature so far, the proposed method is suitable for data with low sampling rate where a pointwise detection and attribution can be applied.

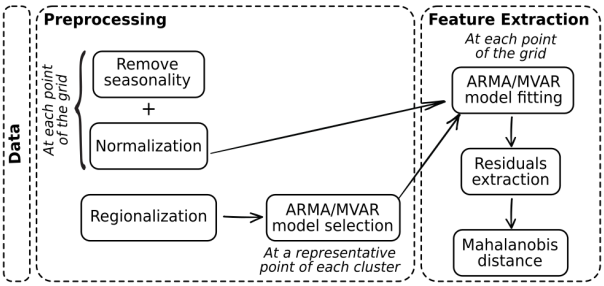


Fig. 1. Flowchart of the anomaly detection method using ARMA/VAR model(s) and Mahalanobis distance.

An abnormal event can be defined as those points within a time series that are not well represented by a previously fitted statistical model [5]. Following this intuitive concept, we have recently proposed a methodology based on linear regression models to detect extreme events in the biosphere [6],[7] (cf. Figure 1). More precisely, in [6] after preprocessing the data, we combine Autoregressive Moving Average Models (ARMA) with the Mahalanobis distance of the residuals between the models and the data to detect those points where the models and the data significantly differ and therefore can be considered as abnormal events. The method was further improved in [7] based on using a Vector Autoregressive (VAR) Model instead of multiple univariate ARMA models. The VAR model allows for presenting the variables with a model that takes into account their inter-dependencies and hence enables better whitening of the residuals and consequently better spatial and temporal detection accuracy of the anomalous events.

In this paper, we present an approach for the attribu-

Corresponding author: Y Guanche, yanira.guanchegarcia@dlr.de
[1] Institute of Data Science, German Aerospace Center, DLR, Jena, Germany [2]Computer Vision Group, Friedrich Schiller University, Jena, Germany [3] Michael Stifel Center for Data driven and Simulation Science, Jena, Germany [4]Max Planck Institute for Biogeochemistry, Jena, Germany

[1]https://www.ipcc.ch/

tion of multivariate anomalous events, where attribution here means to define the contribution of each of the variables involved to making the event an extreme one. The presented approach is based on the decomposition of the Mahalanobis distance of the residuals of the VAR model into components whereby each component presents the contribution of one of the used variables to the Mahalanobis distance. The decomposed form of this metric provides an answer to the question: how much does each variable contribute to this distance metric? The method is applied to the extreme events detected when using five land-atmosphere exchange fluxes (Gross Primary Productivity, Latent Energy, Net Ecosystem Exchange, Sensible Heat and Terrestrial Ecosystem Respiration). The attribution results of the proposed method for different known historical events are presented and compared with the univariate Z-score attribution method.

II. Anomaly Detection with vector autoregressive model and Mahalanobis distance

Let $x_i, i = 1, \cdots, N$ denotes the time series of N Earth observation variables. Each time series $x_i(n), n = 1, \cdots, m$ is a realization of length m real valued discrete stationary stochastic process $X_i, i = 1, \cdots, N$. These N time series can be represented by a pth order VAR model of the form

$$\begin{bmatrix} x_1(n) \\ \vdots \\ x_N(n) \end{bmatrix} = \sum_{r=1}^{p} A_r \begin{bmatrix} x_1(n-r) \\ \vdots \\ x_N(n-r) \end{bmatrix} + \begin{bmatrix} \epsilon_1(n) \\ \vdots \\ \epsilon_N(n) \end{bmatrix}. \quad (1)$$

The residuals $\epsilon_i, i = 1, \cdots, N$ constitute a white noise stationary process with an $N \times N$ residual covariance matrix Σ. The model parameters at time lags $r = 1, \cdots, p$ are defined by

$$A_r = \begin{bmatrix} a_{11}(r) & \cdots & a_{1N}(r) \\ \vdots & \ddots & \vdots \\ a_{N1}(r) & \cdots & a_{NN}(r) \end{bmatrix}. \quad (2)$$

The steps of the anomaly detection method using the VAR(p) model in (1) are summarised in Figure I. After removing seasonality and normalizing the variables as two pre-processing steps, the data are clustered into climate regions according to the Koppen climate classification map [8]. Then for each climate region, a representative point that is geographically centered in the region has been selected. The VAR model order p was defined for every climate region, at each representative point, by means of a Bayesian Criterion [9]. Once the model order p is defined for each region, we proceed with the entire grid, fitting a VAR(p) model for each point in the grid. The residual vector $\mathbf{E}$ is calculated as the difference between the estimated VAR model output and the real value of the used variables. The Mahalanobis distance [10], [11] of the residual vector is then used as a measure of the deviation of the multivariate residuals at a certain time step from their joint distribution. The Mahalanobis distance is defined in the square unit as:

$$d_m(\mathbf{E}) = (\mathbf{E} - \bar{\mathbf{E}})^{\mathbf{T}} \boldsymbol{\Sigma}^{-\mathbf{1}} (\mathbf{E} - \bar{\mathbf{E}}) \quad (3)$$

where $\bar{\mathbf{E}}$ and $\boldsymbol{\Sigma}$ are the mean and covariance matrix of the multivariate residuals vector $\mathbf{E}$ respectively. The mean and the covariance were estimated considering the entire time series. This was the best way to do so in our case due to the short length of the time series used together with its coarse temporal resolution.

When the value of the Mahalanobis distance of the residuals is large, it is assumed that something abnormal occurs in the system and the model is not able to correctly capture it. The easiest way to detect abnormal events is to set a fixed percentile threshold and look for the points with Mahalanobis distance surpassing this threshold. The reader is referred to [6], [7] for further details on different multivariate anomalous event detection methods.

III. Attribution Scheme based on Mahalanobis Distance Decomposition

Once an anomalous event is detected, the next natural question is: *which variables are causing this anomaly?* An intuitive approach to answer this question is to decompose the value of the Mahalanobis distance into components, whereby each component quantifies the contribution of one of the variables to the distance. Garthwaite and Koch [12] recently proposed the corr-max transformation for the decomposition of the Mahalanobis distance which can be easily implemented and provides helpful results from an attribution point of view. The decomposition has the form:

$$d_m(\mathbf{E}) = \mathbf{W}^T \mathbf{W}, \quad (4)$$

where $\mathbf{W} = (W_1, \cdots, W_N)^T$ is a vector with N elements, corresponding to the N variables contributing to the Mahalanobis distance $d_m(\mathbf{E})$, and is calculated by:

$$\mathbf{W} = (\mathbf{S}\Sigma\mathbf{S})^{-1/2}\mathbf{S}(\mathbf{E} - \bar{\mathbf{E}}), \quad (5)$$

where $\mathbf{S}$ denotes a diagonal matrix of the inverses of the standard deviations of the variables of $\mathbf{E}$. The

components of **W** should be uncorrelated, with the transformation chosen to maximize the sum of correlations between the corresponding elements of **S** and **W**.

IV. Experimental Results and Discussion

Data from the Earth System Data Cube (ESDC)[2] developed within the ESDL project has been used as the primary source of land-atmosphere exchange fluxes data for this study. The ESDC comprises spatiotemporal data consisting of: time, latitude, longitude and multivariate Earth Observations. The version used in this study covers the period from January 2001 to December 2012 with 8-daily observations and a spatial grid with a resolution of 0.25°. More than 30 biosphere and atmosphere parameters are included in this database. Out of these variables, we have used those five that mainly measure the terrestrial biosphere activities: Gross Primary Productivity (GPP), Latent Energy (LE), Net Ecosystem Exchange (NEE), Sensible Heat (SH) and Terrestrial Ecosystem Respiration (TER), which were kindly provided by the FLUXCOM[3] initiative [13], [14]. The study area comprises Africa and Europe (see Figure 2). This area was defined as the main study area within the European project BACI: Towards a Biosphere Atmosphere Change Index[4] which is the framework of the current study.

The proposed method is applied for the attribution of two known historic events: the Russian Heatwave of July 2010 and the Drought that affected the Horn of Africa in November 2006. Attribution results of other historic events can be found in [7]. The definition of the temporal and spatial extension of these events was supported by socio-climate experts from the BACI project and is out of the scope of this study.

The results of the proposed method are compared with the univariate Z-score results [15] applied to the same historic extreme events. The Z-score is a measure that compares the distribution of a certain variable within the temporal extent of the detected anomalous event with the distribution of this variable in the entire time series. The Z-score quantifies the discrepancy between these two distributions. This is done separately for each variable. High positive or negative values of the Z-score indicate which variables are most different from their normal behaviour within the time duration of the event.

Figures 2 and 3 show the results obtained for the Mahalanobis distance decomposition and the Z-scores for the two known historic events. Each figure shows the spatial extension of the event (upper left subplot), the Z-scores (upper subplots) and the Mahalanobis distance decomposition (lower subplots). The Z-score subplots show the histograms of the five variables within the time window of the event (red) and the entire time series (grey) together with the value of the Z-score obtained from the comparison of both histograms. The lower subplots show the Mahalanobis distance intensity (map on the left) presenting the spatial extent of the detected anomalous event, in addition to five other maps, each one shows the contribution of one of the used variables to the detected events.

According to the Mahalanobis decomposition using the corr-max transformation, the Russian Heatwave (Figure 2) is most dominantly manifested in LE then in SH. These results of the Mahalanobis decomposition are coherent with the analysis proposed by other authors, [2], [4]. For the case of the Drought in the Horn of Africa, (Figure 3) GPP and NEE, are the most contributing ones. Validating the attribution results of a drought are not trivial since these are very long events where several factors are involved. The Z-score analysis sorts the contribution of the variables differently. These discrepancies between the Z-score approach and the proposed multivariate approach show the relevance of performing a multivariate analysis and the limitations of univariate approaches in such complex systems like biosphere or climate systems. Unfortunately, a detailed quantitative evaluation of the performance of the proposed method is not possible due to the lack of ground truths for the attribution of the extreme events considered in this study.

Acknowledgments

This study has been conducted within the framework of the project BACI: Towards a Biosphere Atmosphere Change Index, funded by the European Union's Horizon 2020 research and innovation program under the grant agreement No 640176.

[2] //http://www.earthsystemdatalab.org/
[3] //http://www.fluxcom.org/
[4] //http://www.baci-h2020.eu

References

[1] M. Reichstein, M. Bahn, P. Ciais, D. Frank, M. D. Mahecha, S. I. Seneviratne, J. Zscheischler, C. Beer, N. Buchmann, D. C.

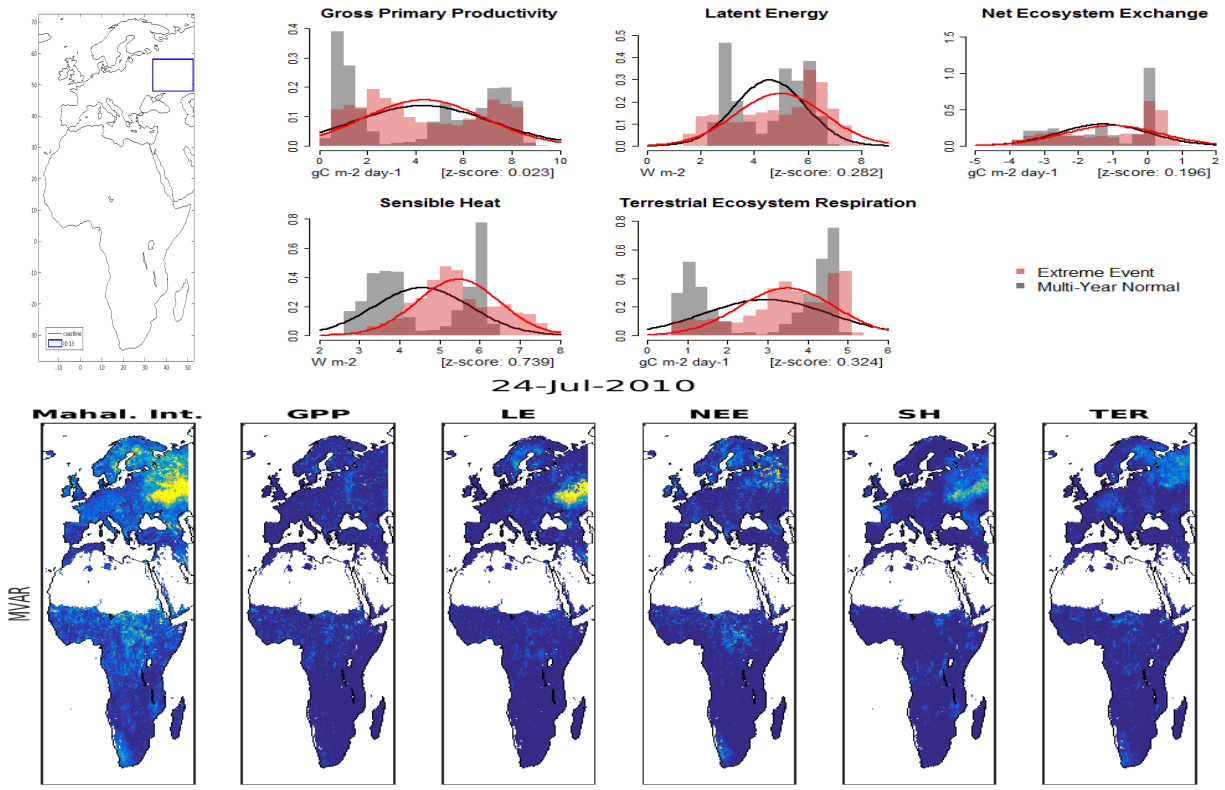


Fig. 2. Attribution scheme for the Russian Heatwave from July 2010. Upper left plot: spatial extent of the Heatwave (blue rectangle), upper right plots: Z-score for the five variables involved, lower plots: Mahalanobis intensity (left) and its decomposition into the five used variables .

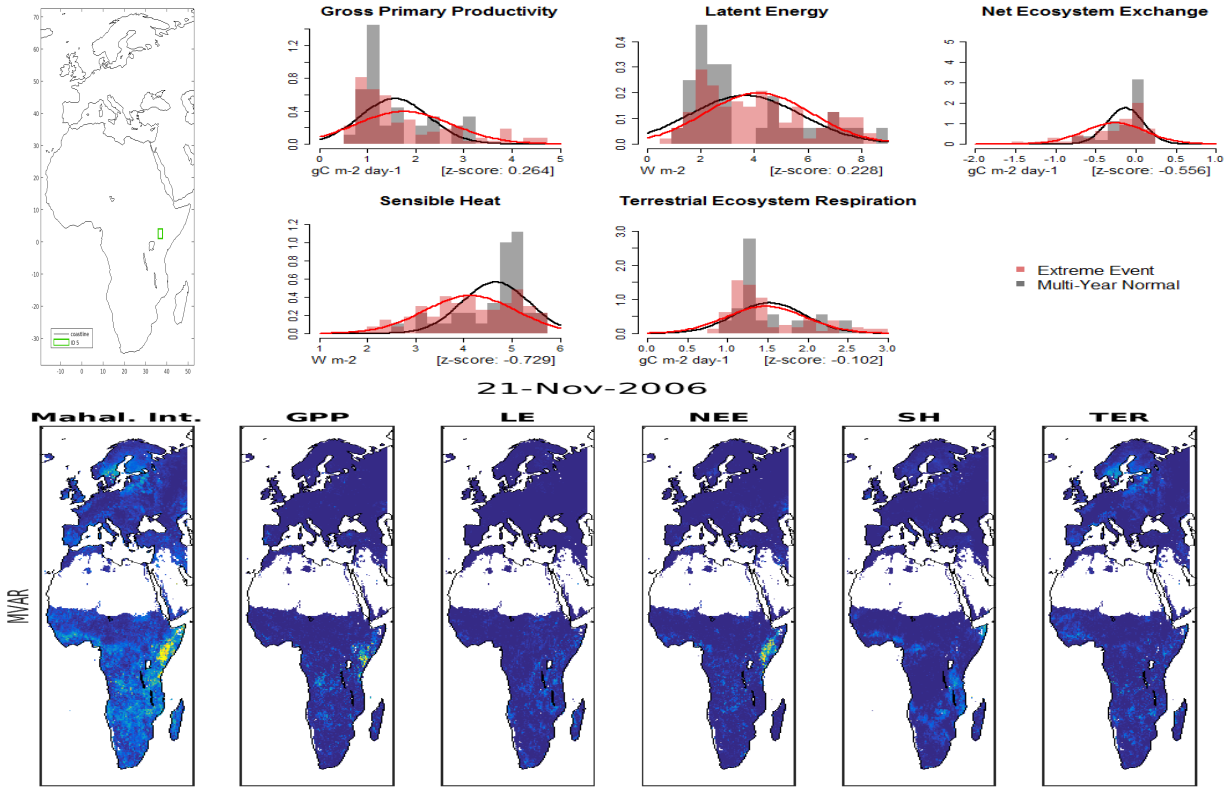


Fig. 3. Attribution scheme for the Drought in the Horn of Africa from November 2006. Upper left plot: spatial extent of the Drought (green rectangle), upper right plots: Z-score for the five variables involved, lower plots: Mahalanobis intensity (left) and its decomposition into the five used variables.

Frank, *et al.*, "Climate extremes and the carbon cycle," *Nature*, vol. 500, no. 7462, p. 287, 2013.

[2] D. G. Miralles, A. J. Teuling, C. C. Van Heerwaarden, and J. V.-G. De Arellano, "Mega-heatwave temperatures due to combined soil desiccation and atmospheric heat accumulation," *Nature geoscience*, vol. 7, no. 5, p. 345, 2014.

[3] J. Zscheischler, M. Reichstein, S. Harmeling, A. Rammig, E. Tomelleri, and M. D. Mahecha, "Extreme events in gross primary production: a characterization across continents," *Biogeosciences*, vol. 11, no. 11, pp. 2909–2924, 2014.

[4] M. Flach, S. Sippel, F. Gans, A. Bastos, A. Brenning, M. Reichstein, and M. Mahecha, "Contrasting biosphere responses to hydrometeorological extremes: revisiting the 2010 western russian heatwave," *Biogeosciences*, vol. 16, pp. 6067–6085, 2018.

[5] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 3, p. 15, 2009.

[6] Y. Guanche, M. Shadaydeh, M. Mahecha, and J. Denzler, "Extreme anomaly event detection in biosphere using linear regression and a spatiotemporal mrf model," *Natural Hazards*, pp. 1–19, 2018.

[7] M. Shadaydeh, Y. Guanche, M. Mahecha, and J. Denzler, "Baci deliverable 5.4: Methods for attribution scheme and near real-time baci." http://www.baci-h2020.eu/index.php/Outreach/Deliverables, 2018.

[8] D. Chen and H. W. Chen, "Using the Köppen classification to quantify climate variation and change: An example for 1901–2010," *Environmental Development*, vol. 6, pp. 69–79, 2013.

[9] G. Schwarz *et al.*, "Estimating the dimension of a model," *The annals of statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[10] P. Mahalanobis, "On the generalised distance in statistics (vol.2, pp.49–55)," *Proceedings National Institute of Science, India.*, 1936.

[11] H. Hotelling, "Multivariate quality control," *Techniques of statistical analysis*, 1947.

[12] P. H. Garthwaite and I. Koch, "Evaluating the contributions of individual variables to a quadratic form," *Australian & New Zealand journal of statistics*, vol. 58, no. 1, pp. 99–119, 2016.

[13] G. Tramontana, M. Jung, C. R. Schwalm, K. Ichii, G. Camps-Valls, B. Ráduly, M. Reichstein, M. A. Arain, A. Cescatti, G. Kiely, *et al.*, "Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms," *Biogeosciences*, vol. 13, pp. 4291–4313, 2016.

[14] M. Jung, S. Koirala, U. Weber, K. Ichii, F. Gans, G. Camps-Valls, D. Papale, C. R. Schwalm, G. Tramontana, and M. Reichstein, "The fluxcom ensemble of global land-atmosphere energy fluxes," in *Scientific Data*, 2018.

[15] J. Reiche, J. Balling, M. Herold, M. Niedertscheider, K. Erb, M. Urban, and C. Schmullius, "Baci deliverable 6.2: Product comparison and validation report." http://www.baci-h2020.eu/index.php/Outreach/Deliverables, 2018.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

GENERATIVE ADVERSARIAL NETWORK FOR CLIMATE DATA FIELD GENERATION

Jussi Leinonen[1], Tianle Yuan[2,3] and Alexis Berne[1]

***Abstract*—Machine learning based on deep neural networks has been previously applied for predictive modeling of geophysical fields. However, machine-learning solutions often fail to reproduce realistic spatial structures, chiefly due to the use of loss functions whose minima correspond to blurry solutions. Generative adversarial networks (GANs) have recently attracted much attention in the machine learning community, as they have proved to be able to generate highly realistic images resembling their training datasets. They have also been recently introduced to climate data analysis. In this paper, we explore the use of GANs to generate precipitation fields and introduce an interpretable, style-based generator architecture for climate data applications.**

I. MOTIVATION

Geophysical fields often exhibit complex spatial structure, yet most analysis methods can only capture this structure in an incomplete statistical sense, unable to generate instances of realistic fields. For instance, various interpolation methods (e.g. Kriging [1]) are commonly used in climate sciences to generate continuous fields from point measurements, but the interpolated field cannot reproduce details that are smaller in scale than the distance between the measurement points. Deep neural networks have proved to be able to distinguish spatial patterns with unprecedented skill, but using them to predict spatial fields also tends to produce blurry, unrealistic results because of the pointwise-applied conventional loss functions such as mean square error. More advanced loss functions have been used to alleviate this problem to some extent (e.g. [2]), but this requires subjective selection and weighting of different metrics. In any case, the problem is usually underdetermined — that is, there are many possible solutions — and thus it is ideally solved using a stochastic approach.

A generative adversarial network (GAN; [3]) learns to generate samples that are visually similar to its training dataset. A GAN consists of two neural networks: a *generator* that takes noise as an input and outputs generated samples, and a *discriminator*. The discriminator is trained to distinguish between real and generated samples, while the generator is trained to "fool" the discriminator as much as possible. Thus, the generator learns a mapping between the relatively simple probability distribution of the noise, which is usually sampled from uncorrelated standard normal variables, and the highly complex and spatially dependent joint probability distribution of the training samples. GANs implemented using deep convolutional neural networks are able to produce complex patterns and have proved capable of generating highly realistic synthetic samples (*e.g* [4], [5], [6]).

GANs are appealing for atmospheric data modeling and analysis as they can generate complex spatial structures in a stochastic manner. The application of GANs to climate data problems has been recently introduced by, *e.g.* [7]. In that work, a conditional GAN was used to generate cloud vertical profiles, as measured by the CloudSat satellite radar, from passive optical observations made by the Moderate-Resolution Imaging Spectroradiometer (MODIS) on the Aqua satellite. The GAN architecture used in that study was relatively simple and required inputs of a specific size. In this paper, we present a style-based, fully convolutional GAN architecture that can be used to generate various geophysical fields, and demonstrate it with radar-based precipitation data.

II. GAN ARCHITECTURE

As mentioned above, a GAN consists of a generator and a discriminator. The discriminator is typically an image-classifier network that passes the image through convolution–downsampling operations to obtain higher-level features, while the generator is the opposite: It

Corresponding author: J. Leinonen, jussi.leinonen@epfl.ch
[1]Environmental Remote Sensing Laboratory, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
[2]Joint Center for Earth Systems Technology, University of Maryland, Baltimore County, Catonsville, MD, USA
[3]Earth Sciences Division, NASA Goddard Space Flight Center, MD, USA

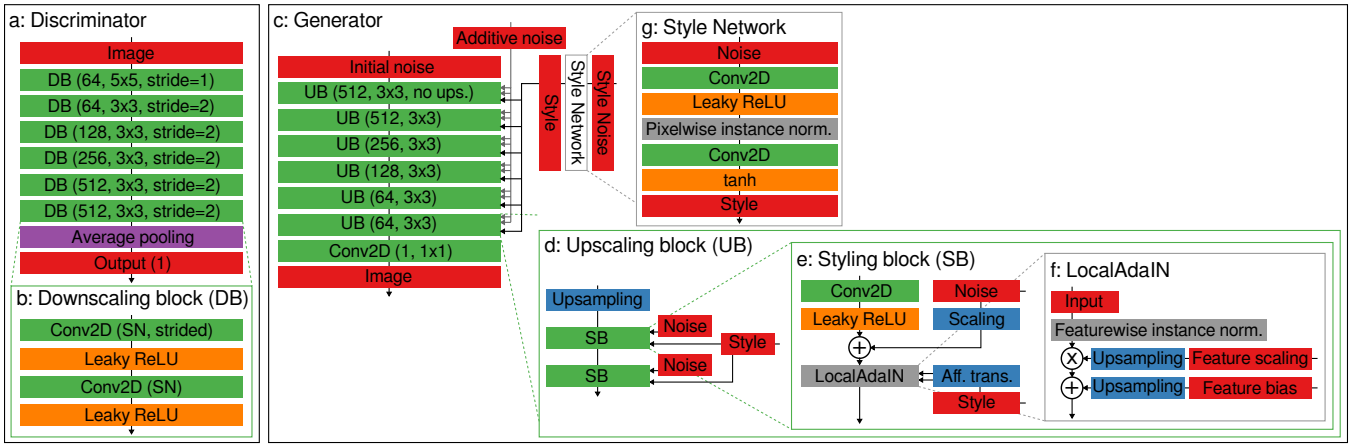


Fig. 1. Diagram of the GAN architecture. The discriminator (a) comprises a series of downscaling blocks (b). The generator (c) is made up of upscaling blocks (d) each containing two styling blocks (e), each of which employs LocalAdaIN (f) modulated by the style network (g). The numbers in convolutional blocks (*e.g.* 512, 3×3) indicate the number of channels and the size of the learned convolution kernels.

starts with high-level features that are then turned into an image through a series of upsampling and convolution operations. A more comprehensive overview of GAN architectures and training is given by [8].

When GANs are used to generate images of objects, each location in the image is semantically different and has different relations to its surroundings. For instance, images of human faces tend to have the face centered in the middle, with key features such as the eyes, nose and mouth located in similar arrangement in each sample. In contrast, geophysical fields follow, to a first approximation, the same rules with respect to their surroundings everywhere in the field. Thus, a natural GAN architecture for generating these fields is fully convolutional, that is, fully connected layers are not used. In such architectures, each pixel shares information only with other nearby pixels at each layer. A further of advantage of fully convolutional GANs is that the output size is not fixed; the same GAN can be used to generate fields of different sizes with the same spatial resolution.

For our discriminator, we use a fully convolutional classifier (Fig. 1b) that processes its input image through six convolutional blocks (Fig. 1a), five of which employ strided convolutions to downsample the image by a factor of 2. The results are then average-pooled to a single value. For example, a 128×128 image is downscaled to 4×4 before pooling, but since the architecture is fully convolutional it can be used for images of any size.

For the generator, we modify the style-based architecture described by [4], where a styling network is used to transform a noise vector $\mathbf{z}$ to a style vector $\mathbf{w}$ (both of dimension 32), which is used to modulate the operation of the main upsampling network at each convolution step using adaptive instance normalization (AdaIN, [9]). In [4], the styling network is fully connected and each style has a global effect on the generated image. This is not necessarily appropriate for geophysical fields, where different processes (and thus different styles) may be dominant in different parts of the image. To enable localized styling, we make the styling network convolutional (Fig. 1f), producing a low-resolution *map* of styles instead of global styles, and introduce a localized AdaIN (henceforth LocalAdaIN; Fig. 1d) layer that applies different styles in different parts of the image. As with [4], white noise is added before each LocalAdaIN layer. The upsampling architecture consists of a total of 12 convolutional layers, each followed by a learned noise addition and LocalAdaIN (Fig. 1g). Five upsampling operations upscale the image to 32 times the size of the initial noise vector.

We trained the GAN as a Wasserstein GAN with gradient penalty (WGAN-GP; [10]), using spectral normalization (SN, [11]) in the convolutional blocks of the discriminator. For us, this combination proved to enable relatively stable training with consistent results across different training runs. On each iteration, we trained the discriminator with five batches of data and the generator with one batch. To encourage sample diversity and particularly to avoid mode collapse (a failure where the generator always outputs the same image) early in the training, we used a batch statistics layer in the discriminator, as described by [4]. The training took approximately 24 hours using an Nvidia Tesla K40 GPU.

III. DATA

We trained the GAN using 128×128 pixel image patches extracted from the rain rate (R) product from the MeteoSwiss radar composite, which combines data from five weather radars to cover Switzerland and parts of the surrounding regions at a resolution of 500 m per pixel [12]. The data were sampled from the year 2018 and balanced such that the maximum $\log(R)$ in images is approximately evenly distributed, and such that the images represent each two-week period of time during the year roughly evenly. Images were also required to contain at least some precipitation, *i.e.* empty regions were not considered. The training set contained a total of 1.44 million images (some of which overlap each other).

The data was normalized by taking $\log_{10}(R)$, then mapping the data to the interval $[0, 1]$. We handle non-raining pixels by mapping them to 0, while the raining pixels are mapped between approximately 0.17 and 1. This procedure allows non-raining data points to be considered as small values by the GAN, which would otherwise require some kind of special processing to distinguish between raining and non-raining points. When postprocessing the generator outputs, we flag all outputs below the 0.17 threshold as non-raining.

Random rotation and mirroring are applied to the images during training to further increase the diversity of the training data.

IV. RESULTS

We show examples of images generated by the trained GAN, along with real images for comparison, in Fig. 2. We see that the generated images are qualitatively similar to the real ones and exhibit different modes of organization from small convective cells to wider regions of stratiform precipitation. The variety produced by the generator is visually similar to the variety of the real images, with no sign of the generator repeating similar fields.

Figure 3 shows the effect of styling on the generator network. In that figure, we use the same style noise for each column of a given row, while the initial and additive noise is different for each sample. We see that the style controls the occurrence and the general mode of organization of the precipitation field, while the random noise represents variability within this organization. As the style is localized, different areas of the image can have different styles. For example, the bottom-right corner in the samples on the top row always contains a relatively uniform precipitating region, the left side on the middle row contains convective precipitation with shorter length scales of organization, and the top-left corner on the bottom row contains more sparsely distributed convective cells. Likewise, the top-left corner on the top row, the right side on the middle row, and the bottom-right corner on the bottom row are mostly empty.

In Fig. 4, we explore the effect of analytically manipulating the style vector $\mathbf{w}$. On both rows i of this image, we interpolate from $\mathbf{w}_i$ to $-\mathbf{w}_i$ to show the effect of style inversion, while using the same initial and additive noise for each image. We see that the precipitation field changes smoothly as a function of $\mathbf{w}$. The inversion mostly controls the occurrence of precipitation in a given area.

V. CONCLUSIONS

As demonstrated by [7], GANs can be used to solve climate data problems where it is desirable to generate realistic synthetic data fields. The generator architecture introduced in this paper makes the generation process more interpretable by introducing localized styling, which controls the presence and the type of organization of the generated field in a given region. This is particularly useful for clouds and precipitation, which exhibit a rich variety of modes of organization, originating from numerous physical processes. Associating styles with corresponding processes alleviates the problem, commonly encountered with deep learning, of having to treat the model as a black box whose inner functionality is poorly understood. The fully convolutional architecture allows the GAN to be used for variable-sized images, although we expect that the spatial extent of the organization of the GAN is limited by the size of the training images.

Since it is relatively straightforward to add conditioning variables to GANs, it should be possible to adapt the proposed architecture to solve more practical problems in climate science. We propose the following as particularly interesting and potentially fruitful topics of further exploration for this architecture and for GANs in general:

1) *Downscaling*, *i.e.* the stochastic generation of high-resolution fields consistent with lower-resolution data, for applications such as super-resolution observations, and climate models parameterizations; already demonstrated for related applications by [13], [14].
2) *Spatiotemporal interpolation* of climate data fields to bridge gaps in observations.

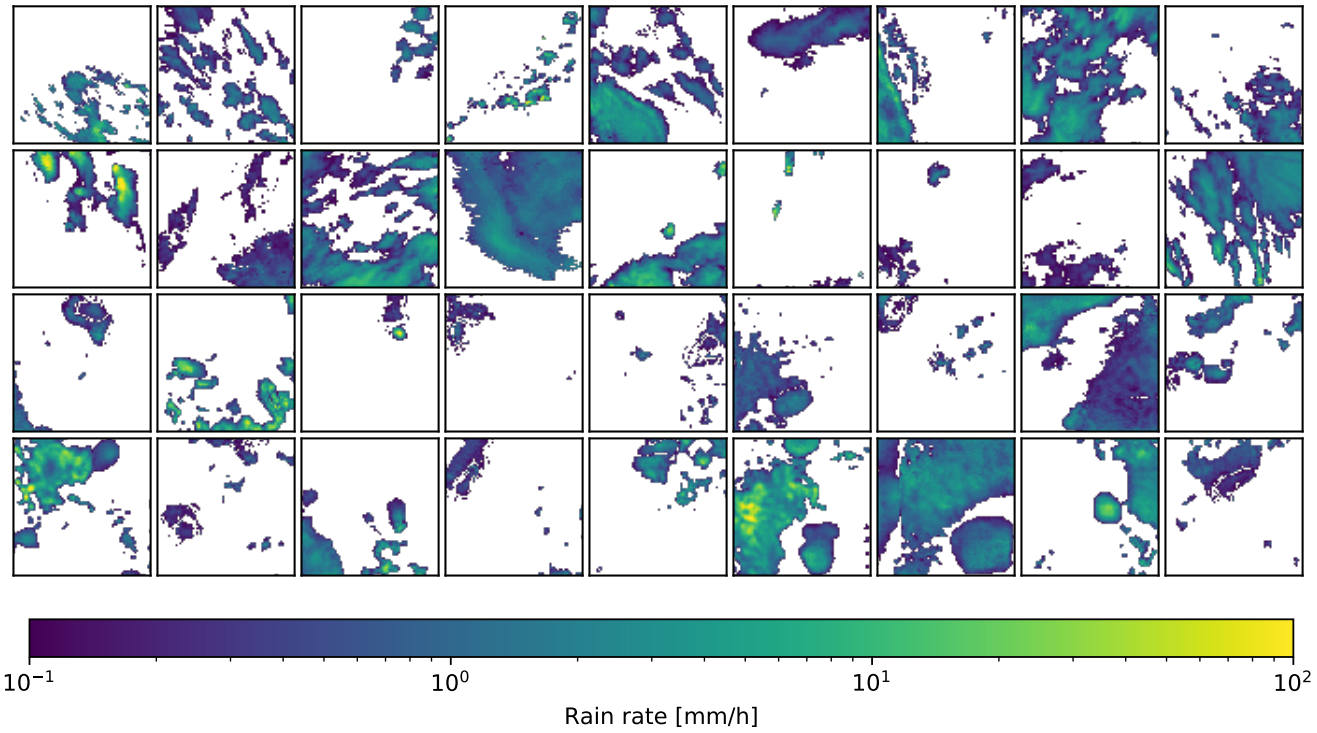


Fig. 2. Real and generated precipitation formations. The first two rows are from the real dataset, while the last two rows were generated with the GAN.

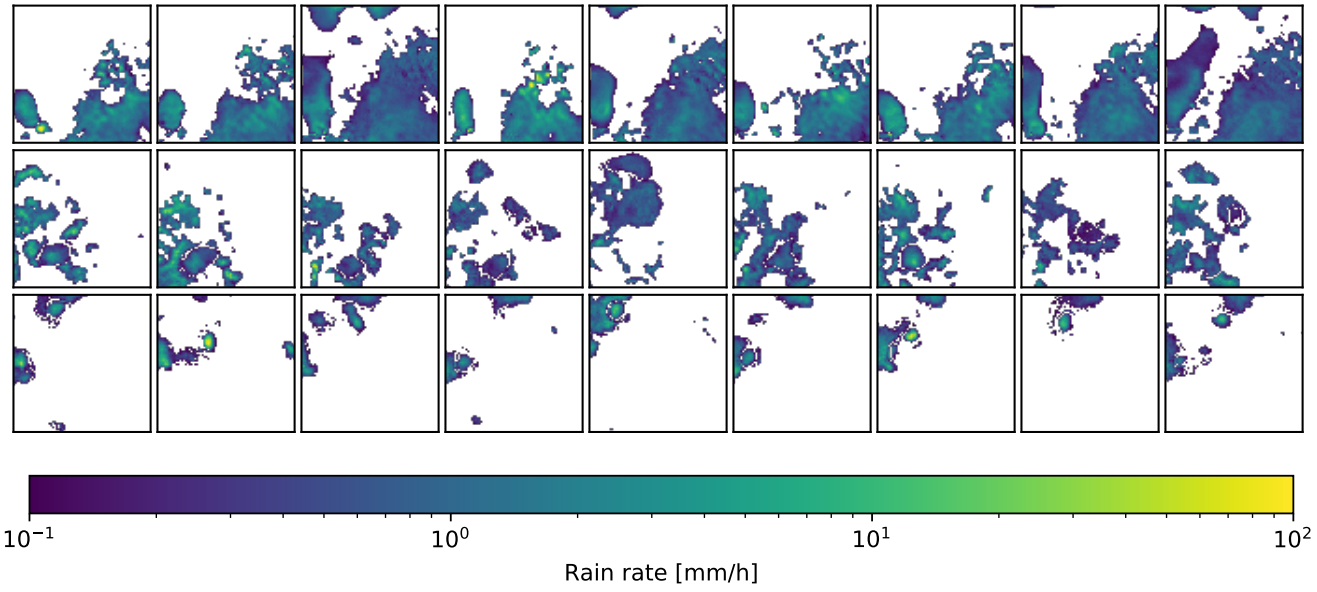


Fig. 3. The effect of styles on the generator. On each row, the style is held constant while the random noise is varied.

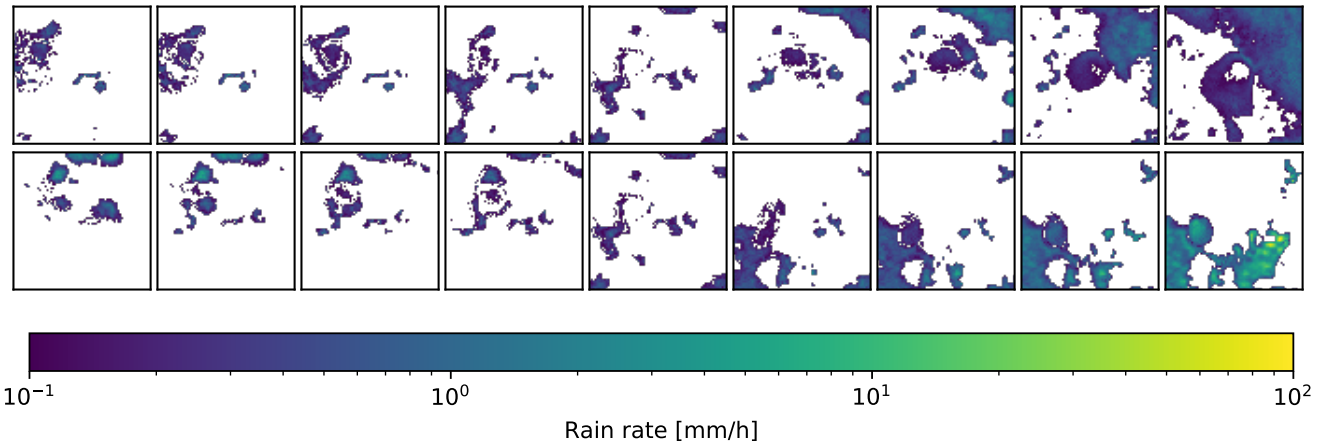


Fig. 4. Style inversion: The noise is held constant for each image while the style is transitioned from $\mathbf{w}_i$ to $-\mathbf{w}_i$ on both rows i.

3) *Nowcasting*, the short-term prediction of the time evolution of data fields where numerical weather predictions are not (or not yet) available.
4) Unsupervised or semi-supervised *pattern classification* using GAN variants that can employ the discriminator as a classifier (*e.g.* [15]).

Moreover, further research is needed into validating the results of GANs and improving the ability to understand their predictions. In the context of climate data, it should be noted that climate change may change the data distributions and thus a distribution learned by a GAN in one climate may not be valid in another. Thus it may be preferable to use GANs that approximate, or are explicitly constrained by, physical models, which will be unchanged in a changing climate.

An implementation of the GAN using Python/Keras [16] and information on obtaining the trained network weights and the training dataset are available at https://github.com/jleinonen/geogan/.

ACKNOWLEDGMENTS

We thank MeteoSwiss for providing the radar data and Daniel Wolfensberger for assisting us with using it, as well as Gionata Ghiggi for feedback.

REFERENCES

[1] M. L. Stein, *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media, 2012.

[2] M. Mathieu, C. Couprie, and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," *arXiv preprint arXiv:1511.05440*, 2015.

[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 2672–2680.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2019.

[6] J. Menick and N. Kalchbrenner, "Generating high fidelity images with subscale pixel networks and multidimensional upscaling," in *International Conference on Learning Representations*, 2019.

[7] J. Leinonen, A. Guillaume, and T. Yuan, "Reconstruction of cloud vertical structure with a generative adversarial network," *Geophys. Res. Lett.*, vol. 46, 2019. doi: 10.1029/2019GL082532

[8] Y. Cao, L. Jia, Y. Chen, N. Lin, C. Yang, B. Zhang, Z. Liu, X. Li, and H. Dai, "Recent advances of generative adversarial networks in computer vision," *IEEE Access*, vol. 7, pp. 14 985–15 006, 2019. doi: 10.1109/ACCESS.2018.2886814

[9] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of Wasserstein GANs," in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5767–5777.

[11] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," in *International Conference on Learning Representations*, 2018. [Online]. Available: https://arxiv.org/abs/1802.05957

[12] U. Germann, M. Boscacci, M. Gabella, and M. Sartori, "Peak performance: Radar design for prediction in the Swiss Alps," *Meteorological Technology International*, no. April 2015, pp.

42–45, 2015.

[13] W. Ma, Z. Pan, J. Guo, and B. Lei, "Super-resolution of remote sensing images based on transferred generative adversarial network," in *2018 IEEE International Geoscience and Remote Sensing Symposium*, 2018. doi: 10.1109/IGARSS.2018.8517442 pp. 1148–1151.

[14] D. Zhu, X. Cheng, F. Zhang, X. Yao, Y. Gao, and Y. Liu, "Spatial interpolation using conditional generative adversarial neural networks," *Int. J. of Geogr. Inf. Sci.*, 2019. doi: 10.1080/13658816.2019.1599122

[15] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2172–2180.

[16] F. Chollet, "Keras," https://github.com/fchollet/keras, 2015.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

A COMPARISON OF TECHNIQUES TO OPTIMISE TROPICAL CYCLONE ENSEMBLE PREDICTIONS

Mohan Smith[1], Ralf Toumi[1]

Abstract — **Several attempts have been made at optimising tropical cyclone ensemble predictions through the formation of a sub-ensemble. These are produced with the goal of excluding any erroneous ensemble members prior to computing the ensemble mean. To date there has been no rigorous comparison between the techniques used to identify erroneous ensemble members, or the forecast variables which benefit most from sub-ensemble forecasting. This paper presents a comparison of techniques trained with outgoing longwave radiative (OLR) images to provide immediate improvements in forecasts of cyclone track, intensity and precipitation. We also compare these results to techniques trained with cyclone position. Working entirely within a model environment we find that cyclone track forecasts improve ≈10% at 24hrs lead time whereas precipitation and intensity forecasts see no improvement for sub-ensembles formed via OLR images or cyclone positions.**

I. MOTIVATION

Tropical cyclones (TC) are the most destructive weather systems on Earth, producing large volumes of precipitation, high velocity winds and storm surges. The huge economic, social and environmental costs associated with them [1] drive the demand for improvements in TC forecasts to mitigate their effects.

Primarily, forecasts are produced through deterministic computational coupled ocean-atmosphere models incorporating data measured from bouys, aircrafts, ships, dropsondes and satellites [2]. Ensemble prediction systems (EPS) are formed from multiple model simulations providing a range of potential weather states. Typically operating at a low spatial resolution (13-18km), their use can be justified as multiple models can accommodate the various model physics schemes and the uncertainties in measured variables [3][4][5]. The mean of the ensemble behaviour, particularly regarding track forecasts, yields higher forecast skill over any one model [6][7]. Work by [8] states that 'forecasters can improve on the ensemble mean by excluding the erroneous ensemble members (models) prior to computing the mean behaviour, provided a consistent tool is available to assist the forecaster in discerning the best models within the ensemble'. As such the tool must be robust, computationally inexpensive and easily applied. Several works have attempted to identify the best sub-ensemble selection tool in an attempt to improve and update the TC forecast in the time between forecast publications.

Work by [9] reduces ensemble size by selecting ensemble members whose forecasted position is within a radius r from the observed position, where r = average position error across the whole ensemble, at small lead times. Following this, [10] present a comparison of sub-ensemble size (N) to the same technique, selecting the closest N members to the target TC to compose the sub-ensemble. These two techniques were validated on 35 and 80 TCs respectively and demonstrate improvements in forecast skill of 30% and 18% at 24 hrs lead time.

In an attempt to balance spread and accuracy, subsequent work by [11] maintains the ensemble size by determining N 'good' and N 'poor' members, based on proximity to observed TC location for 3 TCs. Poor members were replaced with perturbed good members and models re-run for a range of N values. This work targets both track and intensity (P_{min} and V_{max}) with 10% skill improvement over 24 hrs. However, the requirement to rerun models suppresses the ability to update the TC forecast in the information void between forecast publications.

Whilst the aforementioned works deploy a k nearest neighbor (k-NN) algorithm (for details see [12]) on TC positions, [13] demonstrated an improvement (5%) in TC precipitation forecasts through a self-organising map (SOM) based cluster analysis. An SOM produces a synthetic mapping of the dataspace onto a 2d non-periodic grid whilst preserving the topological

Corresponding author: M. Smith, mohan.smith13@imperial.ac.uk [1]Department of Space and Atmospheric physics, Imperial College London

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

relationships between data points, for details on the SOM algorithm see [14] and [15]. Their study involved mapping both the ensemble and measured precipitation fields over Taiwan for short lead times into clusters. The cluster containing the measured data constitutes the sub-ensemble from which large lead time precipitation behaviour is improved.

The above works demonstrate the ability to provide an improvement in TC forecasts with relatively little computational expense, however, none present a comparison of the different techniques to do this. Our work aims to provide a detailed analysis of a range of clustering methods applied to two training variables - OLR images and position. We compare our results to the equivalent algorithm used by [9] and [10], to understand where improvements in cyclone track, P_{min} and total precipitation forecasts can be made in the 12hr information void between EPS forecast publications.

Following the rationale of [16] we use the perfect model approach to validate skill improvements within the ECMWF medium range EPS. Work by [17] shows that the perturbed models composing this EPS demonstrate a spread in behavior equal to the best track error of the ensemble mean. Given this ensemble property, we extract 1 ensemble member as a hypothetical 'target' TC, the remaining 49 ensemble members constitute our baseline EPS forecast, and a sub-ensemble from this baseline constitutes our improved forecast.

We use OLR images to discern erroneous ensemble members by measurement of their euclidean distance (ED) to the target OLR under the following principals:

1) OLR images provide information on the TC cloud structure and heights and are therefore a strong representation of current TC and environmental properties [18].
2) Current TC behaviour is indicative of future TC properties.
3) Similar OLR images determined by ED will identify similar ensemble members.
4) Measured OLR images from satellites are readily available in real time. This enables the techniques validated here to be applied to real TCs.

We apply this framework to ECMWF forecasts (from 2016 – 2018) for 23 TCs across all basins (northern and southern hemisphere, Pacific, Atlantic and Indian Oceans). All TCs achieved a category 4 intensity and improved sub-ensemble forecasts were applied across the entire TC lifespan, both defined by best track data [19]. This provides a complete range of TC characteristics for validation.

II. METHOD

The working dataset is available from the TIGGE Database [20]. The ECMWF EPS from 8th March 2016 is a 50 member ensemble at the highest spatio-temporal resolution available and is in current use.

Within each ensemble member we calculate the track of the TC as the location of the minimum surface pressure at t_n in the vicinity of the TC position at t_{n-1}, initialised at the observed best track position [19] at forecast publication time (t_0). The TC motion from t_{n-1} to t_n time steps defines the along and across track vectors. We calculate the total precipitation as the sum of all precipitation over the prior 6 hrs within 500 km of the storm center.

Model output datasets are produced at 6 hr increments and we apply our sub-ensemble selection techniques at t_1 only (6hrs after the forecast is published) to improve the forecast for larger lead times. Lead times hereafter will refer to the number of hours after t_1.

Within each ensemble forecast publication we extract 1 member as a target TC, where the mean P_{min}, total rain and position, across the remaining 49 members is our baseline forecast. An N member (where N<49) sub-ensemble is our improved forecast. Testing sub-ensemble selection techniques in this way means that with each published forecast we can test each sub-ensemble selection method 50 times. As such, with our whole dataset, we make 18,400 attempts at sub-ensemble selection, far more than any prior works. Additionally, this is many more tests than would be possible with the current generation of models applied to real OLR data over the past 2 years. Finally, working wholly within the model world provides a perfect, complete dataset and enables objective comparison between the target TC, the baseline forecast and the improved forecast.

All model OLR images were centered on their P_{min} at t_1. Sub-ensemble members were selected based upon their similarity to the target OLR image. To replicate, expand and compare to the work by [10] we also produced separate sub ensembles determined by cyclone position only.

We identified several algorithms based on ED to deploy for sub-ensemble extraction. These are either statistical or SOM based. These techniques include:

1) k-NN (used for track position and OLR images) – extracting closest k members to compose a sub-ensemble. Applied for $5 \leq k \leq 35$.
2) k-means [21] clustering (OLR images only) - splitting the ensemble into k clusters, for $2 \leq k \leq 5$.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

3) SOM based (OLR images only) - Locate the target on an i x j SOM and extract the nearest cluster (defined by members mapped to adjacent nodes on the SOM for details see [22]).
4) SOM based (OLR images only) - Extract members residing in a k x k extraction region, on the SOM, centered at the target location. Where the SOM is an n x n node grid and $0.2n \leq k \leq 0.8n$. For details see [23].

Upon identification of the sub-ensemble, we produce the baseline and improved forecasts of P_{min}, position and total precipitation by taking averages (both the arithmetic mean and a weighted mean) of these variables across the constituent members. We define the improvement in forecast skill as:

$$skill = 1 - \frac{\sum_{a=0}^{b} |\overline{SE} - target|}{\sum_{a=0}^{b} |\bar{E} - target|}$$

Where $\overline{SE}$ = mean across the sub-ensemble, $\bar{E}$ = mean across the whole ensemble, b = no. forecasts made (18,400). The sensitivity of forecast improvement to any input parameters required by the algorithms were also tested as were image sizes in the range (1200x1200 – 3500x3500 km).

III. EVALUATION

The techniques used here, trained on both OLR images and track position, have demonstrated an ability for significant improvement in track forecast skill over the baseline EPS. Of the algorithms trained on OLR images, technique 4, used by [23] yields the largest skill improvement. Additionally there is comparable skill in the k-NN algorithm (k = 12). The k-means algorithm (2) and the SOM clustering (3) are the worst performing algorithms with little or no skill. The best k-NN configuration trained on position slightly out performs the algorithms trained on OLR images. Arithmetic means outperformed weighted means across all algorithms as found by [9].

The best Laaksonen [23] configuration utilises 3000x3000 km images mapped on a 10 x 10 SOM. Sub-ensemble members were extracted within an 8x8 node region centered upon the target node. This algorithm configuration yields skill improvements of 8% for 24hrs lead time, and 5 % for 48 hrs lead time. This is slightly lower (2%) than the best k-NN configuration (15 nearest neighbours) trained on cyclone position. Figure 1 demonstrates this improvement in track forecast skill for OLR images vs track positions. Additionally it is shown that skill improvements in P_{min} and precipitation are weak for both training variables.

Analysis of the sensitivity of the improved forecast to

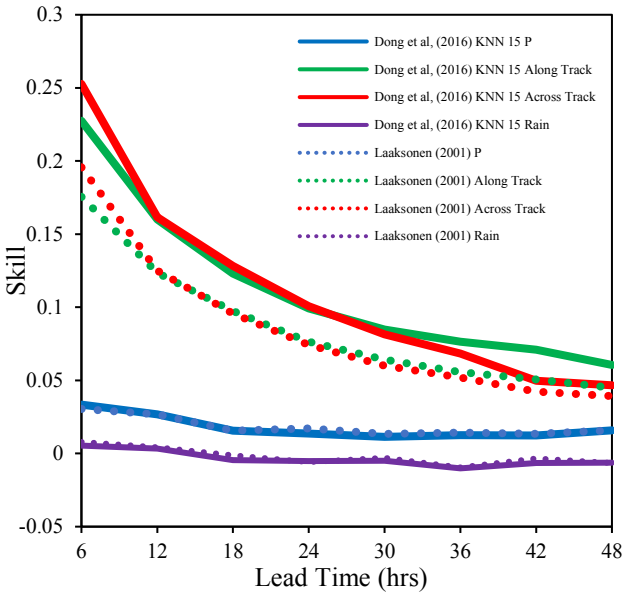


Figure 1. Forecast skills for all variables trained on TC position vs OLR image.

the various configurations of input parameters yields interesting results. Generally larger OLR images, encompassing a radius upto 1800km produce stronger skill across all training algorithms. The input parameters for training algorithms directly or indirectly control the size of the sub ensemble. Tests show that generally a sub-ensemble size between 10-20 members yield the largest skill improvement for both training variables.

The Laaksonen algorithm extracts a variable sub ensemble size controlled by the target data's coherency with the ensemble. The SOM generally maps the images towards the corners, where the k x k extraction region becomes smaller due to SOM non-periodicity. Target images located here have high coherency with significant portions of the ensemble, such that only these (often 5-20) coherent ensemble members are extracted. Conversely, target images mapped towards the SOM center are more incoherent and have a large extraction region. They therefore extract a larger sub-ensemble with a variety of behaviours, replicating the variability in the whole ensemble. This reduces the tendency to weaken the quality of the forecast through sub-ensemble selection for anomalous cyclone behaviours.

Splitting the sub-ensemble according to the OLR images enables analysis of the regions of importance in the image which yield high skills in track forecasts. Comparing the best and worst performing 20% of sub-ensembles demonstrates that the most improved forecasts are found with sub-ensembles that replicate the OLR target image in the front right quadrant at large distances (>500km), see figure 2. Additionally the sub

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

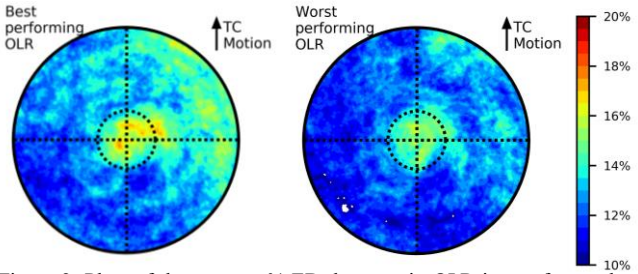


Figure 2. Plots of the average % ED decrease in OLR image from sub-ensemble formation, for the best (left) and worst (right) performing track forecasts. Percentages are calculated as the ED between the target image and the sub-ensemble image normalised by the ED between the target image and the baseline ensemble image. Averages are calculated across the top and bottom 20% performing forecasts. All images are rotated according to target TC motion. A 500 km radius dashed circle is also plotted.

ensemble must replicate the inner (<500km) TC cloud behavior for all quadrants. It is well understood that the environmental flow is the dominant control on cyclone track, whilst the outer tangential wind profile determines the strength and orientation of beta gyres (e.g. see [24][25][26]). More skillful track forecasts are produced when the sub ensemble replicates the outer tangential winds compared to radial winds. This therefore implies that the front right quadrant of the OLR image provides a good indicator of the outer tangential wind profile.

ACKNOWLEDGMENTS

Funding for the authors was provided by NERC National Environmental Research Council.

REFERENCES

[1] E. Strobl, "The Economic Growth Impact of Hurricanes: Evidence from U.S. Coastal Counties", *Review of Economics and Statistics*, vol. 93, no. 2, pp. 575-589, 2011.

[2] L. Magnusson et al., "ECMWF Activities for Improved Hurricane Forecasts", *Bulletin of the American Meteorological Society*, vol. 100, no. 3, pp. 445-458, 2019.

[3] M. Tracton and E. Kalnay, "Operational Ensemble Prediction at the National Meteorological Center: Practical Aspects", *Weather and Forecasting*, vol. 8, no. 3, pp. 379-398, 1993.

[4] F. Molteni, R. Buizza, T. Palmer and T. Petroliagis, "The ECMWF Ensemble Prediction System: Methodology and validation", *Quarterly Journal of the Royal Meteorological Society*, vol. 122, no. 529, pp. 73-119, 1996.

[5] R. Buizza et al., "The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System)", *Quarterly Journal of the Royal Meteorological Society*, vol. 133, no. 624, pp. 681-695, 2007.

[6] J. Goerss, "Tropical Cyclone Track Forecasts Using an Ensemble of Dynamical Models", *Monthly Weather Review*, vol. 128, no. 4, pp. 1187-1193, 2000.

[7] D. Richardson, "Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size", *Quarterly Journal of the Royal Meteorological Society*, vol. 127, no. 577, pp. 2473-2489, 2001.

[8] R. Elsberry and L. Carr, "Consensus of Dynamical Tropical Cyclone Track Forecasts—Errors versus Spread", *Monthly Weather Review*, vol. 128, no. 12, pp. 4131-4138, 2000.

[9] L. Qi, H. Yu and P. Chen, "Selective ensemble-mean technique for tropical cyclone track forecast by using ensemble prediction systems", *Quarterly Journal of the Royal Meteorological Society*, vol. 140, no. 680, pp. 805-813, 2013.

[10] L. Dong and F. Zhang, "OBEST: An Observation-Based Ensemble Subsetting Technique for Tropical Cyclone Track Prediction", *Weather and Forecasting*, vol. 31, no. 1, pp. 57-70, 2016.

[11] J. Li, Y. Gao and Q. Wan, "Sample Optimization of Ensemble Forecast to Simulate a Tropical Cyclone Using the Observed Track", *Atmosphere-Ocean*, vol. 56, no. 3, pp. 162-177, 2018.

[12] E. Fix and J. Hodges, *Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties*. Randolph Field, Tex.: USAF School of Aviation Medicine, 1951.

[13] M. Wu, J. Hong, L. Hsiao, L. Hsu and C. Wang, "Effective Use of Ensemble Numerical Weather Predictions in Taiwan by Means of a SOM-Based Cluster Analysis Technique", *Water*, vol. 9, no. 11, p. 836, 2017.

[14] T. Kohonen, "Self-organized formation of topologically correct feature maps", *Biological Cybernetics*, vol. 43, no. 1, pp. 59-69, 1982.

[15] T. Kohonen, "Essentials of the self-organizing map", *Neural Networks*, vol. 37, pp. 52-65, 2013.

[16] L. Leslie, R. Abbey and G. Holland, "Tropical cyclone track predictability", Meteorology and Atmospheric Physics, vol. 65, no. 3-4, pp. 223-231, 1998.

[17] S. Lang, M. Leutbecher and S. Jones, "Impact of perturbation methods in the ECMWF ensemble prediction system on tropical cyclone forecasts", *Quarterly Journal of the Royal Meteorological Society*, vol. 138, no. 669, pp. 2030-2046, 2012.

[18] V. Dvorak, "Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery", *Monthly Weather Review*, vol. 103, no. 5, pp. 420-430, 1975.

[19] "IBTrACS - International Best Track Archive for Climate Stewardship", *Ncdc.noaa.gov*, 2019. [Online]. Available: https://www.ncdc.noaa.gov/ibtracs/index.php?name=wmo-data.

[20] P. Bougeault et al., "The THORPEX Interactive Grand Global Ensemble", *Bulletin of the American Meteorological Society*, vol. 91, no. 8, pp. 1059-1072, 2010.

[21] J. Hartigan and M. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm", *Applied Statistics*, vol. 28, no. 1, p. 100, 1979.

[22] G. Lin and M. Wu, "A hybrid neural network model for typhoon-rainfall forecasting", *Journal of Hydrology*, vol. 375, no. 3-4, pp. 450-458, 2009.

[23] J. Laaksonen, M. Koskela, S. Laakso and E. Oja, "Self-Organising Maps as a Relevance Feedback Technique in Content-Based Image Retrieval", *Pattern Analysis & Applications*, vol. 4, no. 2-3, pp. 140-152, 2001.

[24] M. Fiorino and R. Elsberry, "Some Aspects of Vortex Structure Related to Tropical Cyclone Motion", *Journal of the Atmospheric Sciences*, vol. 46, no. 7, pp. 975-990, 1989.

[25] L. Carr and R. Elsberry, "Models of Tropical Cyclone Wind Distribution and Beta-Effect Propagation for Application to Tropical Cyclone Track Forecasting", *Monthly Weather Review*, vol. 125, no. 12, pp. 3190-3209, 1997.

[26] J. Chan and J. Kepert, *Global perspectives on tropical cyclones*. Singapore: World Scientific Pub. Co, 2010.

LATENT SPACE REPRESENTATION AND RNN FOR IMAGE-BASED TYPHOON INTENSITY ANALYSIS AND PREDICTION

Clément Playout[1,2], Asanobu Kitamoto[2]

***Abstract*—Currently employed image-based methods for intensity estimation of typhoons rely either on handcrafted features which are not guaranteed to be optimal or on subjective interpretation of images. In this paper, we propose a general pipeline based on unsupervised features extraction from typhoons images with Convolutional Autoencoders and we demonstrate the ability of Recurrent Neural Network to learn temporal model from these extracted features. We evaluate our methodology on different tasks: classification (identification of tropical versus extra-tropical cyclone and estimation of tropical cyclone categories, with a respective accuracy of 94.89% for the first task and 65.90% for the later), regression (central pressure estimation) and forecast. Finally, we also propose a simple methodology to assess the consistency of best tracks data based on cross-years evaluation.**

I. MOTIVATION

The term "tropical cyclones" refers to low-pressure storms that occur on tropical latitude. These potentially highly catastrophic storms are at the core of many investigations from a large array of fields such as climatology, meteoinformatics or economics. Investigating the causes, behaviors or consequences of cyclones, brings two critical challenges: (I) estimating the intensity and (II) analyzing the past and future evolution, specifically in the perspective of the different scenarios of climate warming. The main difficulty in the first case is finding an objective and consistent metric that applies to cyclones in every regional basin. In the latter case, to deal with potentially inconsistent historical measurement, partially due to different ways of evaluating cyclones intensity over time. Thanks to geostationary satellites, we have access to continuous and global observations of cyclones during their lifetimes. These observations allow estimating their intensities, historically, with the Dvorak technique [1]. This technique proposes a procedural interpretation of cloud patterns to estimate tropical cyclone changes. It is still commonly used but suffers from being subjective and time consuming. To alleviate these limitations, automatic pattern recognition methods have been introduced in [2] (*Objective Dvorak Technique*) and extended in [3] with the *Advanced Dvorak Technique*. Other approaches, not relying on the Dvorak scale, have also been proposed. For instance, Velden et al. [2] suggest an approach based on Deviation Angle Variance, improved in [4]. But all these approaches have in common that they rely on handcrafted features. Instead, using Convolutional Neural Network (CNN), Pradhan et al. [5] train an architecture for the classification of cyclones in 8 categories and the estimation of their sustained wind speed. Their work doesn't take into account the temporal consistency intrinsic to the sequence structure. Another limitation of this work is the splitting between train and test set that is done frame-wise instead of typhoons-wise. This induces a strong bias in the evaluation procedure, as adjacent frames (looking very much alike) might be in distinct sets.

Our work mainly focuses on challenge (I) (intensity estimation) by proposing a novel neural network pipeline using only satellite IR images as input variables, based on Convolutional AutoEncoders (CAE) for features extraction and a Recurrent Neural Network (RNN) trained on different tasks. We consider in this paper: (1) tropical cyclone (TC) vs extra-tropical cyclone (ETC) classification, (2) TC classification and (3) central pressure regression. For this last task, we show the ability of the model to provide future forecasts. We also bring a contribution to challenge (II) (analyzing past and future evolution of typhoons) by discussing the notion of cross-years evaluation in order to assess the temporal consistency (and therefore the reliability) of the best-tracks. Indeed, given the potentially subjective nature of manual estimation of typhoon intensity, we want to verify this consistency quantitatively. By training a model on only a small recent subset of the data

Corresponding author: C Playout, clement.playout@polymtl.ca
[1]École Polytechnique de Montréal[2]National Institute of Informatics

and evaluating it on different past time-intervals, we estimate the generalization capacity of the network over time. Temporally inconsistent best-tracks were expected to lead to unstable performances over time. We did not observe such behavior, leading to the conclusion that our model is not sensitive to potential inconsistencies in the data. This suggests that the variability over time of the data is lower than the model's intrinsic error. This is a valuable result for climatology studies that exploit these data.

II. METHOD

A. Data

This work focuses on Pacific Ocean cyclones (*typhoons*). The data we used come from the Digital Typhoon project [6], which consists of an exhaustive gathering of satellite-based IR images of typhoons and their corresponding best tracks. The database is composed of the observations of 971 typhoons, dating from 1978 to 2017. Each typhoon sequence is composed of several frames, acquired every hour (except from data anterior to 1986, acquired every 3 hours), and their corresponding best tracks information compiled by the *Japan Meteorological Agency* (JMA). In total, there are 164627 frames in the database, captured by a total of 9 different satellites. The frequency of the best tracks provided by JMA is every 6 hours. To match the frames frequency, cubic spline interpolation is used for approximating missing values. The center of the typhoons are obtained from the best tracks in order to crop and center each image on their geographical coordinates. Each frame becomes a 256×256 temperature map, with values ranging from $T_{min} = 160$ Kelvin (K) to $T_{max} = 310$ K. Each frame is normalized to the range $[-1; 1]$ using T_{min} and T_{max}.

B. Features extraction

Each map m_t (where $t \in \{1, ..., T\}$ stands for *time*, indexing each frame of a sequence) represents a high-dimensional vector in the space $[-1; 1]^{256\times256}$. In this space, finding statistical significance is an extremely complex task. To alleviate this complexity, we extract features using a CAE to extract a latent space representation z_t of each image m_t. Obtained with an *encoder*, z_t represents a highly compressed version of m_t that ideally contains all the necessary information to reconstruct m_t, which is the task of the *decoder*. Both the encoder and the decoder are convolutional neural networks trained jointly. The architectural details of the structure, called TyNet (for *Typhoon Network*) are described in Figure 1. The training objective function is an addition of three components. First, inspired by [7], we choose to train our model using a perception loss $\mathcal{L}_p$. Given an external pretrained model, the loss consists in the distance between features extracted with this model from both the original image m_t and its reconstruction $\hat{m}_t$. We note the former features F_{org} and the latter F_{rec}. The perception loss is simply obtained by computing the mean-square error:

$$\mathcal{L}_p = ||F_{org} - F_{rec}||^2 \quad (1)$$

We use the MobileNetV2 network [8] as the external pretrained network. This choice is determined by hardware limitation, that imposes to choose a light-weight model. This loss allows capturing dependencies between images in a more abstract way (based on content) than pixel-level based losses, which are known to lead to very blurry image reconstructions. As stated in [7], by itself, this loss may lead to high-frequency artifacts. To counter-balance this effect, they advocate for an adversarial loss, that forces the decoder to build realistic reconstructions following the same pixel distribution as the original images. The adversarial loss is obtained by adding a discriminator network $\mathcal{D}$, which takes as an input both the original images and their reconstructions, and is trained to distinguish between then, by minimizing the loss:

$$\mathcal{L}_\mathcal{D} = \frac{1}{2}(\mathbb{E}[\mathcal{D}(\hat{m}_t)^2] + \mathbb{E}[(\mathcal{D}(m_t) - 1)^2]) \quad (2)$$

In this configuration, $\mathcal{D}$ is pushed to predict label 1 for m_t and label 0 for $\hat{m}_t$. In reverse, the TyNet is trained to fool the discriminator, by minimizing:

$$\mathcal{L}_{adv} = \mathbb{E}[(\mathcal{D}(\hat{m}_t) - 1)^2] \quad (3)$$

The discriminator is trained following the procedure described in [9], that aims at minimizing the Wasserstein distance between the original images distribution and the reconstructed one, while forcing a Lipschitz constraint on the discriminator by using Gradient Penalty. Nonetheless, neither $\mathcal{L}_{adv}$ nor $\mathcal{L}_p$ guarantee $\hat{m}_t$ to be close to m_t pixel-wise; as they both act as metrics of similarity in terms of distribution, helping $\hat{m}_t$ to *look* realistic but not necessarily close to m_t. Therefore, we also added an additional constraint under the form of an image-similarity loss. We opted for the Mean Structural Similarity Index (MSSIM), as introduced in [10], that compares local patterns of images. Overall, the global loss function used to train the TyNet is:

$$\mathcal{L}_{TyNet} = \mathcal{L}_{adv} + \mathcal{L}_p - \lambda_{MSSIM} \times \text{MSSIM}(m_t, \hat{m}_t) \quad (4)$$

The factor λ_{MSSIM} has been set to reinforce the similarity property between the input and its reconstruction. We observe that $\lambda_{MSSIM} = 10$ gives qualitatively slightly better reconstruction result. The latent space

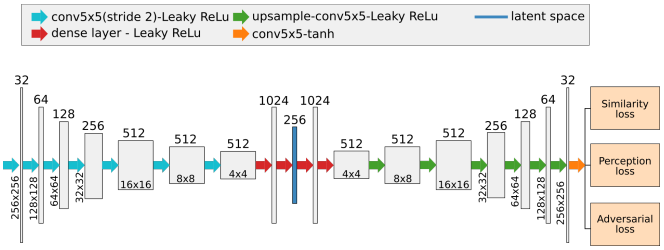


Fig. 1: Autoencoders architecture used for latent space extraction. The encoder compresses the input image in a vector of size 256 that the decoder uses to reconstruct the input image.

representation z_t obtained for each input m_t is a vector of size 256. The TyNet is distributed and trained in parallel on two GPUs (Quafro RTX 5000). Figure 2 illustrates samples reconstructed with TyNet.

As a baseline, we extracted features from the image space using Incremental Principal Component Analysis [11].

For the sake of completeness, we also observed that the features extracted from a pre-trained network (the GoogleNet [12] in our experimentation) were robust features for our temporal model. Nonetheless, because of the lack of interpretability of such features (due to the absence of a reconstruction step), we did not extend our experiment with these features and leave it for future work.

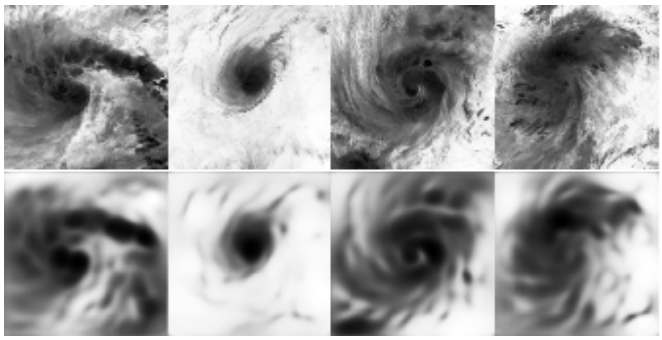


Fig. 2: Encoding-Decoding of typhoons images using TyNet: First row: original frames. Second row: TyNet reconstruction from the latent space.

C. Temporal Series Model

For temporal classification, regression and forecast, we adopted a model based on Recurrent Neural Networks. This network is composed of two layers of *Long Short-Term Memory* (LSTM), introduced by [13] and an output layer being a simple RNN-cell. This output layer was the only part of the network that was modified according to the different tasks, otherwise, the network architecture was preserved. This network is fed by temporal sequences of the feature vectors extracted by the TyNet.

1) Data imputation for missing values: Some sequences contains missing frames. Let's denote $\tilde{z}_t$ a missing value in a sequence. After evaluating different imputation strategies, we observed than a RNN-cell $\mathcal{R}$ trained on predicting the missing value gave the highest results on the validation. We therefore adopt an imputation method based on missing value estimation $\tilde{z}_t = \mathcal{R}(z_{t-1}, z_{t-2}...z_1)$.

2) Uncertainty estimation: Dropout is inserted between the LSTM layers to limit the effect of overfitting during training. But as demonstrated in [14], dropout can also be used as a Bayesian Approximation to estimate the model uncertainty. Uncertainty can be interpreted as the standard deviations of outputs produced by a large number of different models trained on the same data. Instead of actually training different models, at evaluation time, different connections are dropped in a loop for a given pass, leading to slightly different outputs at every iteratveryion. The final estimation is an average of these different outputs and the uncertainty is their standard deviation.

III. EVALUATION

For evaluation purposes, we split the database in a training set and a test set. The test set is composed of sequences posterior to 2013 (103 sequences), the remaining ones being used for training (from 1978 to 2013 included). From the training set, 20% of the sequences were randomly extracted and used as validation set. Because of the high correlation between frames of the same sequence, it is important to split the sets into distinct sequences rather than solely distinct frames. All following metrics are computed per sequence and then averaged over the whole test set.

A. Tropical Cyclone vs Extra-tropical Cyclone

Many Tropical Cyclones (TC) turn into Extra-tropical Cyclones (ETC) at the end of their lifetimes, during a phenomena called extratropical transition. TC and ETC require different forecasting strategies [15]. Therefore being able to discriminate between both is of great importance for meteorologists. Our first network is trained to identify the category of each frame based on

present (z_t) and past ($z_{1<i<t}$) samples, corresponding to a scenario of real-time analysis of incoming typhoon. Note that extending this model to post-event analysis (using not only past and present but also future samples) is easy, by simply turning all the RNN-like cells into bidirectional-ones. We provide results for both scenarios in Table I, but it appears that bidirectional cells, for this task, don't improve the performance. By looking at the F1-score (harmonic average between precision and recall) we can appreciate in particular the improvement provided by using the TyNet over the baseline (PCA). Specifically, a higher recall (ratio between predicted true positive and positive samples) indicates a better capacity to recognize extra-tropical cyclones (considered as the positive class in our task), which are significantly less represented than tropical cyclones in our database (important class imbalance). As the precision is maintained in this case, this increase is not a consequence of a higher number of predicted false positive.

TABLE I: Tropical Cyclones VS Extra-Tropical Cyclones

	Accuracy	Precision	Recall	F1
Real-time analysis				
PCA+RNNs	94.99%	77.21%	75.76%	71.67%
TyNet+RNNs	94.89%	75.47%	92.67%	79.30%
Post-event analysis (bidirectional RNNs)				
Tynet + RNNs	94.77%	78.33%	89.50%	77.98%

B. *Tropical Cyclone Intensity Classification*

In the JMA best tracks, TCs are classified in four categories, corresponding to different Maximum Sustained Wind during 10min (MSW) values, expressed in knots (kt). These categories, and their MSW intervals are: Tropical Depression (TD, ≤ 33 kt), Tropical Storm (TS, $34-47$ kt), Severe Tropical Storm (STS, $48-63$ kt), Typhoon (T, ≥ 64kt). The same network is used as in the previous section, as the exception of the last layer that now outputs four categories. Results of classification are summarized in Table II. Unlike in the previous task, post-event analysis using bidirectional cells is shown to improve classification performance. By looking at the confusion matrix (Figure 3), we observe that the errors are mostly done between adjacent classes, which can be explained by the arbitrary nature of these classes' boundaries.

C. *Central Pressure Estimation*

So far, we have estimated the typhoon intensity assuming discrete categories. Nonetheless, physical properties of typhoons are continuous values. We mentioned

TABLE II: TC intensity-class estimation

	Accuracy	Precision	Recall	F1
PCA+RNNs	59.92%	70.04%	59.92%	60.30%
TyNet+RNNs	65.90%	75.33%	65.90%	66.44%
Post-event analysis (bidirectional RNNs)				
Tynet + RNNs	68.17%	72.25%	68.17%	67.50%

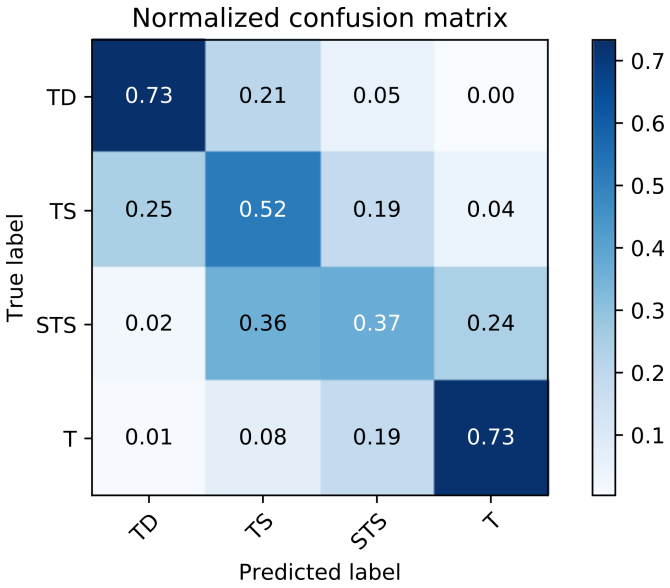


Fig. 3: Classification results for TC identification. Higher values in the diagonal indicates a correct classification.

the MSW as one way to quantify typhoon intensity, which closely related to the central pressure (CP) of the typhoon's eye. However, in the best tracks, MSW is sparser and noisier than CP (in particular, MSW is not available for ETC), which has justified the use of the latter as the intensity metric. For this task, we not only estimate the CP of the current frame, but we also forecast the expected future values (n-hours ahead). To evaluate the quality of the regression, we compute the mean-absolute error (MAE) per sequence, and we compute the average of this score on the given test set. Results are shown in Figure 4. Since JMA also provides their 1, 2 and 3-days ahead annual CP forecast errors[1], we compare these with the performance of our model retrained on data from 1978 to 2005 and tested on years 2006-2017 (years corresponding to the errors provided by JMA). The results are shown in Figure 5: it appears clearly that our model improves the performance compared to JMA. Moreover, our results are also comparable to the ones obtained in [16], using SHIPS (Statistical Hurricane Intensity Prediction Scheme). Nonetheless, our model relies solely on images whereas SHIPS is

[1]Resource in Japanese: http://www.data.jma.go.jp/fcd/yoho/typ_kensho/table_prs.html

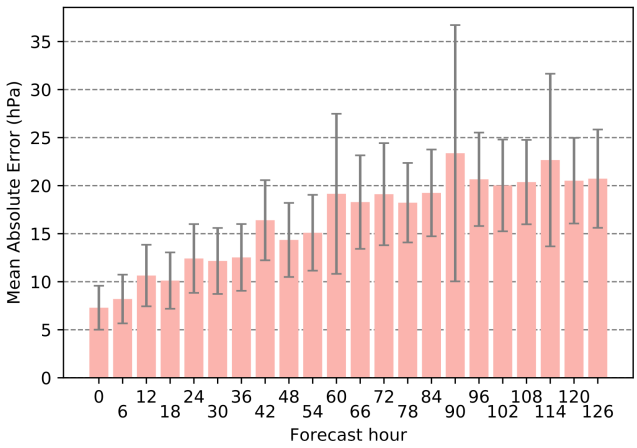


Fig. 4: Predicted central pressure error per forecast hour.

based on linear regression from multiples predictors (from climatology simulation, physical measurement, etc.). Thanks to the flexibility of RNNs, we hypothesize that the fusion of these predictors with image latent-space representation should improve even more the prediction performance, but this is left for future work.

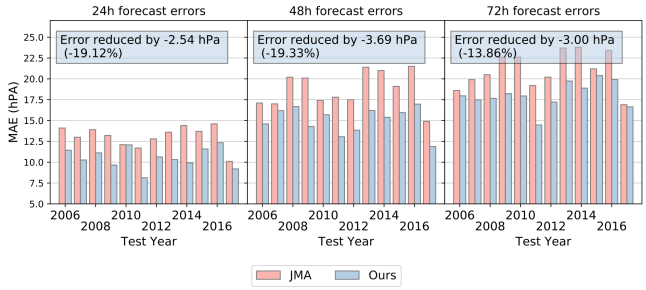


Fig. 5: Comparison of the errors obtained by JMA and our model. In average, our model produces an error that is 19% smaller than JMA's for 24h and 48h forecast, and 13,9% smaller for 72h forecast.

D. Cross-Years Evaluation

We now aim at identifying possible inconsistencies in the best tracks data in order to help creating more reliable groundtruth. The model is trained for central pressure estimation on the period 2010-2017 and then evaluated on different time periods. For each time-interval, we also indicate the mean uncertainty of the model. A higher uncertainty tends to point out unusual typhoons patterns and might therefore indicate a potential outlier sequence. Figure 6 shows the results obtained. The model does considerably worst on the two periods 1978-1982 and 1983-1986 corresponding to the periods were the acquisition is done every three hours. On the other periods, the model performs surprisingly evenly, which goes in the sense of consistent best tracks over the different time periods. Indeed, in case of best tracks inconsistencies on a given period (due, for example, to the adoption of a new methodology), the model should have produced a higher error on this corresponding period. Therefore, if there are inconsistencies, they only lead to errors considerably smaller than the model's intrinsic error.

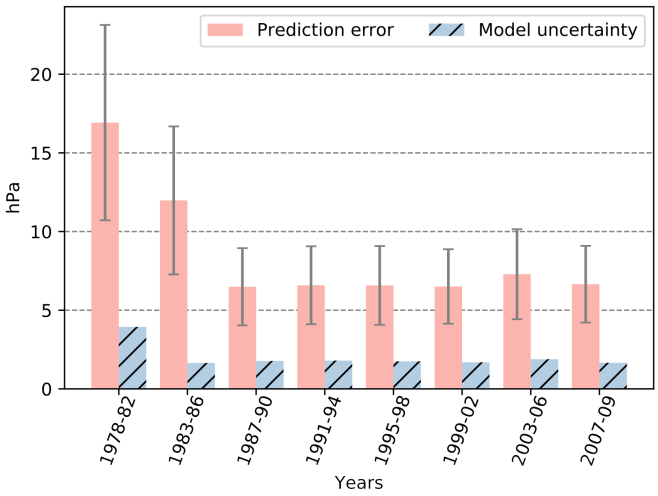


Fig. 6: Generalization with limited training data: given only a limited amount of data (sequences from 2010 to 2017), we evaluate the ability of the network to generalize on the whole database.

ACKNOWLEDGMENTS

Funding for the authors was provided by the National Institute of Informatics (NII) International Internship Program. We also like to acknowledge Dr. Shimada from the Meteorological Research Institute of JMA for the fruitful discussion and Maria Astefanoaei for revising this manuscript.

REFERENCES

[1] V. F. Dvorak, "Tropical Cyclone Intensity Analysis and Forecasting from Satellite Imagery," *Monthly Weather Review*, vol. 103, pp. 420–430, May 1975.

[2] C. S. Velden, T. L. Olander, and R. M. Zehr, "Development of an Objective Scheme to Estimate Tropical Cyclone Intensity from Digital Geostationary Satellite Infrared Imagery," *Weather and Forecasting*, vol. 13, pp. 172–186, Mar. 1998.

[3] T. L. Olander and C. S. Velden, "The Advanced Dvorak Technique: Continued Development of an Objective Scheme to Estimate Tropical Cyclone Intensity Using Geostationary Infrared Satellite Imagery," *Weather and Forecasting*, vol. 22, pp. 287–298, Apr. 2007.

[4] E. A. Ritchie, G. Valliere-Kelley, M. F. Piñeros, and J. S. Tyo, "Tropical Cyclone Intensity Estimation in the North Atlantic Basin Using an Improved Deviation Angle Variance Technique," *Weather and Forecasting*, vol. 27, pp. 1264–1277, June 2012.

[5] R. Pradhan, R. S. Aygun, M. Maskey, R. Ramachandran, and D. J. Cecil, "Tropical Cyclone Intensity Estimation Using a Deep Convolutional Neural Network," *IEEE Transactions on Image Processing*, vol. 27, pp. 692–702, Feb. 2018.

[6] A. Kitamoto, "The Development of Typhoon Image Database with Content-Based Search," in *Proceedings of the 1st International Symposium on Advanced Informatics*, pp. 163–170, 2000.

[7] A. Dosovitskiy and T. Brox, "Generating Images with Perceptual Similarity Metrics based on Deep Networks," in *Advances in Neural Information Processing Systems 29* (D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, eds.), pp. 658–666, Curran Associates, Inc., 2016.

[8] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018.

[9] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved Training of Wasserstein GANs," in *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 5767–5777, Curran Associates, Inc., 2017.

[10] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, Apr. 2004.

[11] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, "Incremental Learning for Robust Visual Tracking," *International Journal of Computer Vision*, vol. 77, pp. 125–141, May 2008.

[12] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Boston, MA, USA), pp. 1–9, IEEE, June 2015.

[13] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, pp. 1735–1780, Nov. 1997.

[14] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," in *Deep Learning Workshop, ICML*, vol. 1, p. 2, 2015.

[15] S. C. Jones, P. A. Harr, J. Abraham, L. F. Bosart, P. J. Bowyer, J. L. Evans, D. E. Hanley, B. N. Hanstrum, R. E. Hart, F. Lalaurette, M. R. Sinclair, R. K. Smith, and C. Thorncroft, "The Extratropical Transition of Tropical Cyclones: Forecast Challenges, Current Understanding, and Future Directions," *Weather and Forecasting*, vol. 18, pp. 1052–1092, Dec. 2003.

[16] M. Yamaguchi, H. Owada, U. Shimada, M. Sawada, T. Iriguchi, K. D. Musgrave, and M. DeMaria, "Tropical Cyclone Intensity Prediction in the Western North Pacific Basin Using SHIPS and JMA/GSM," *Sola*, vol. 14, pp. 138–143, 2018.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

CLIMATE CHANGE-INDUCED WATER SCARCITY AND CROP PLANNING: CASE STUDY OF MKOMAZI IRRIGATION

Oseni Taiwo Amoo[1], Abdultaofeek Abayomi[2], Solomon Bilewu Olakunle[3], Wahab Salami Adebayo[4], Israel Edem Agbehadji[5]

Abstract— **In the context of climate change, the hydrological regime of Mkomazi River Basin (MRB), KwaZulu Natal Province, South Africa is likely to be affected with tendencies to alter the water availability for the downstream populace. This current study uses a Flow Duration Curve (FDC), and a combination of multivariate statistical techniques and ombrothermic diagram to evaluate the hydro-meteorological trends, seasonal fluctuation patterns and available water in determining the watershed area that could be managed in an integrated manner to mitigate water scarcity for optimal crop planning and yield. Optimisation linear programming techniques in Microsoft Excel solver was utilized to maximise total crop net benefit during a planting season while minimising total water utilised under changing land use and climate variabilities. The findings from this research are of importance in determining crop water requirement especially during water scarcity, in allocating the appropriate quantity of water to various crops on farmland for optimum yield as climate change is likely to have both compensating and underpinning effects on agricultural crops yield.**

I. INTRODUCTION

The need for integrated land and water resources management in reducing poverty and ensuring food security cannot be overemphasised. The hydrological regime of MRB, KwaZulu Natal Province, South Africa's downstream irrigated water use for farming has been affected due to the recent varying global climate change. South Africa has been a water-stressed country [1]. Its irrigation water use is about 60% of the country's total water requirements, while urban requirements are up to 25% as the second-largest user. The outstanding 15% is shared by the other sectors [2]. The MRB's natural assets, as well as its traditional values, are of tremendous importance to all the riparian users. They are essential for industrial and domestic use, tourism and irrigation. They are also a means of livelihood for subsistence farming, fish production and commercial afforestation at the downstream sector [3,4,5]. Thus, accurate estimation of water use is one of the challenges facing the irrigation water supply sector. The ability to manage an irrigation system is contingent on an accurate estimate of the percentage of pumped water that becomes available for crop use. This is because not all the water taken from a source reaches the root zone of the plants as some are lost during transport through canals and fields [6,7].

Therefore, irrigated farm land's water allocation to each of the cropping areas through the furrow canal needs to take into consideration the specific constraints of each site to guard against waterlogging and flooding problems. Thus, efficient planning and management strategies are essential for the optimum utilization of water resources for continuous improvement and sustainable development [3,8]. This research study intends to optimize total crop net yield over a planting season while minimizing total water utilization under changing land use and climate variabilities.

II. MATERIALS AND METHODS

The hydro-meteorological data utilized was collected from the South African Weather Information System (SAWs), and the Agricultural Research Council (ARC). These data were subjected to the Mann-Kendall, Sen's Slope, ombro-thermic diagram and Factor analysis (FA)

Corresponding author: A. Abayomi, 21451441@dut4life.ac.za
[1]Department of Civil Engineering and Geomatics; [2,5]ICT and Society Research Group: [1,2,5]Durban University of Technology, Durban, South Africa. [3,4]Department of Water Resources and Environmental Engineering, University of Ilorin, Ilorin, Nigeria.

to relate the trend, pattern, fluctuation into wet or dry season in order to identify hydroclimatic variables responsible for streamflow characteristics as a basis for determining available water. The Factor Defining Variables (FDV) for the rotated component matrix, greater than 0.6 from the rotated matrix formed the major variance in each component identified. The Principal Component Analysis (PCA) as a choice of FA explains the contribution of the unobserved common features in a target event from observed ones in order to reduce variety of data matrix to form few selected component variables derived to form a true representative of its original sets and explain seasonal variation in water availability in the environments. Spatial analysis in Geographical Information System (GIS) was used to measure changes in the present and future land use and land cover pattern to determine the size, slope, shape, and imperviousness of the drainage area to reveal the physical processes acting upon the topographic conditions on the distribution of soils, vegetation and occurrence of water.

Development of the model application:

The optimization and development of a water management model for surface irrigation project is estimated as the minimum flow discharge throughout the year [9, 10]. The crop planning optimization problem in this study was conducted for a planting season at Mkomazi. A farmland with an area of 100 000 m^2 and maximum water quota of 4000 m^3 per ha/annum was selected as a case study. Four different crops which include groundnuts, maize, pecan nuts and lucerne are planted on the piece of farmland. Furthermore, we adopted in this study, an assumption that all the crops are not rain-fed but rely solely on irrigation. The mathematical objective function optimization is as follows:

(1) Decision variables and objectives:

Minimize Irrigation Water Use

The mathematical equation for minimizing total irrigation water use is presented in equation 1:

$$Minimize \ \ WU_{vol} = \sum_{i=1}^{n} (CWR_i \times A_i) \qquad (1)$$

where WU_{vol} is the total irrigation water use in m^3 and CWR_i represents total annual estimated gross crop water requirements under flood irrigation in mm for crop i, selected from Table 1.

Table 1: Crop water requirement (Grove, 2011 [11]).

SN	Crop	Crop water requirement (mm)
1	Maize	620
2	Groundnuts	810
3	Lucerne	1,800
4	Pecan nuts	1,920

(2) Problem constraints

The objective crop planning optimisation problem is subject to the following constraints:

Constraint 1: Available Total Land Area

The sum of areas A_i where the crops are grown must not be greater than the total land area available for farming. This constraint is presented in equation (2):

$$A = \sum_{i=1}^{n} (A_i) \leq 1,000,000 \qquad (2)$$

Constraint 2: Maximum Crop Planting Areas

The minimum and maximum planting areas for each crop constitute the boundary constraints of the problem. Each crop is planted in at least 100, 000 m^2 to avoid crop scarcity which may lead to hike in selling prices of food while the maximum planting areas ensure there will not be excess/surplus so that farmers will not have storage or selling glitches.
The computation of maximum crop planting areas in this study is as follows:

Since the minimum planting area for each crop =100,000 m^2, Then the 4 crops will occupy a minimum of 100,000 x 4 = 400,000 m^2 : This leaves (1,000,000 – 400,000)m^2 = 600,000 m^2 *as the maximum area available for all crops. Therefore,* 600,000 m^2 *is chosen as the maximum planting area for all the crops.*

The boundary constraint for this problem is therefore specified in equation 3 as:

$$1000000 \leq A_i \leq 600000 \quad (3)$$

Constraint 3: Irrigation Canal Capacity

The quantity of water available on the farm annually is limited by the capacity of the irrigation canal. Water is supplied to the farm through feeder canals with a maximum capacity of 150m^3 per hour for 5½ days in a week [7, 12]. To avail consistency in computation; the canal capacity is converted to volumetric units' m^3 as follows:

Amount of water available per day = 120 m^3/hour x 24 hours = 2880 m^{3}daily. Water available for 5½ days per week = 5.5 x 2880 = 15840 m^3 weekly. Therefore w*ater available for a month = 4 x 15840 = 63,360*m^3.

Hence, the canal is able to supply a maximum of (63360 x 12) = 760,320 m^3 of water annually.

It is thus required that total irrigation water use does not exceed the maximum that can be supplied by the feeder canal. This constraint is presented in equation 4:

$$WU_{vol} \leq 760320 \qquad (4)$$

III. RESULTS AND DISCUSSION

Figure 1 shows the mean monthly meteorological data variation for the Basin while the seasonal Sen's slope and Mann-Kendall Trend Test result is depicted in Table 2 for the station's year 2008-2015. The cross correlation coefficients as shown in Table 3 indicate similarities in the mechanisms that causes a phenomenon to be exhibited in the basin. The results of the PCA and FA are useful for reducing and interpreting large multivariate datasets with their underlying linear structures as shown in Tables 4 and 5. The most significant reduced variables that affect the discharge flow as explained by the component defining variables while FA predictors result is shown in Table 5 and relates the unsuspected relationships and the total water demand [13].

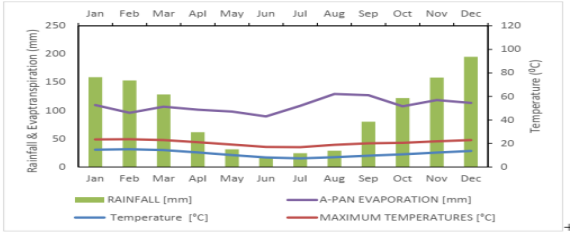


Figure 1: Mean monthly meteorological data variation for the Basin.

Table 2: Seasonal Sen's slope and Mann-Kendall test

	Unit [2]	Kendall's tau [3]	S [4]	Var(S) [5]	p-value (Two-tailed) [6]	Sen's slope [7]	Trend [8]
MaxT	°C	0.009	34	0.824	0.05	-0.026	Increasing
MinT	°C	-0.053	-191	0.2	0.05	-0.037	Decreasing
Solar	MJ/m^2	-0.083	-296	0.046	0.05	-0.026	Decreasing
windsp	[m/s]	-0.032	-115	0.442	0.05	-0.004	Decreasing
MaxRH	%	-0.205	-737	< 0.0001	0.05	-0.546	Decreasing
MinRH	%	-0.168	-604	< 0.0001	0.05	-0.475	Decreasing
R Evap	Mm	0.067	242	0.104	0.05	0.318	Increasing
Rain	Mm	-0.166	-596	< 0.0001	0.05	-0.408	Decreasing
Runoff	m^3/s	-0.039	-136	0.36	0.05	0	Decreasing

The Factor Analysis (FA) was applied to remove the multi-collinearity from identified variables while the morpho-climatic correlation matrix assessed the interrelationship between the meteorological dataset and the streamflow as shown in Table 3. A strong relationship exists between the hydro morpho-climatic data for the basin which implies that the influx of rainfall increases the runoff. The PCA from FA gives insight of the structural relationship between streamflow and rainfall as shown in Table 4.

Table 3: Morpho-climatic matrix

Variables	MaxT	MinT	Solar	windsp	MaxRH	MinRH	Revo	Rain	Runoff
MaxT	**1**	**0.809**	**0.645**	-0.253	**0.427**	**0.376**	**0.454**	**0.194**	**0.463**
MinT	**0.809**	**1**	**0.632**	-0.100	**0.625**	**0.749**	**0.373**	**0.463**	**0.545**
Solar	**0.645**	**0.632**	**1**	0.032	**0.234**	**0.330**	**0.628**	**0.526**	**0.506**
windsp	-0.253	-0.100	0.032	**1**	-0.386	-0.075	**0.223**	0.025	-0.096
MaxRH	**0.427**	**0.625**	**0.234**	-0.386	**1**	**0.800**	-0.085	**0.316**	**0.186**
MinRH	**0.376**	**0.749**	**0.330**	-0.075	**0.800**	**1**	0.084	**0.436**	**0.375**
Revo	**0.454**	**0.373**	**0.628**	**0.223**	-0.085	0.084	**1**	**0.257**	**0.277**
Rain	**0.194**	**0.463**	**0.526**	0.025	**0.316**	**0.436**	**0.257**	**1**	**0.441**
Runoff	**0.463**	**0.545**	**0.506**	-0.096	**0.186**	**0.375**	**0.277**	**0.441**	**1**

Values in bold are different from 0 with a significance level alpha=0.05

The bold squared cosine values depict the most significant uncommon variables that affect the discharge flow. Based on the pre-screening of the data using PCA, the data were classified into two main components namely PC1 and PC2.

Table 4: Factor loading of the variables.

	Factor Loadings								
Variables	**F1**	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**F8**	**F9**
MaxT	**0.629**	0.004	0.235	0.038	0.008	0.046	0.022	0.006	0.013
MinT	**0.869**	0.006	0.000	0.036	0.016	0.011	0.026	0.016	0.020
Solar rad	**0.606**	0.186	0.009	0.003	0.037	0.033	0.122	0.004	0.000
Windsp	0.027	**0.372**	0.369	0.142	0.061	0.019	0.000	0.010	0.000
MaxRH	0.419	**0.441**	0.013	0.036	0.018	0.003	0.011	0.057	0.002
MinRH	**0.552**	0.169	0.150	0.045	0.009	0.035	0.004	0.026	0.010
Revo	0.230	**0.498**	0.028	0.041	0.049	0.147	0.005	0.002	0.000
Rain	**0.377**	0.014	0.236	0.204	0.122	0.011	0.034	0.001	0.001
Runoff	**0.440**	0.027	0.000	0.247	0.257	0.020	0.003	0.006	0.000

Values in bold correspond to each variable, the factor for which the squared cosine is the largest

PC1 is a more significant component than PC2. Using the corresponding factors loading value, the scores on PC1 can be computed as shown in equation 5, while the scores on PC2 can also be estimated as in equation 6.

PC1= **0.629** × TempMax + **0.869** × TempMin +**0.606**× Solar+ **0.027** × Windsp + **0.419** × RHMax+**0.552**x RHMin + **0.23** × ET_o + **0.377**× Rainfall (5)

PC2 = **0.004** × Tempmin + **0.006** × Tempmax + **0.186** × Solar + **0.372** × wind speed + **0.441** × R. Humiditymax + **0.169** × R.Humiditymin +**0.498**× ET_o + **0.014**× Rainfall (6)

Table 5: Factor analysis of the Eigen value.

Variables (%)	F1	F2	F3	F4	F5	F6	F7	F8	F9
Eigenvalue	4.15	1.72	1.04	0.79	0.58	0.33	0.23	0.13	0.05
Variability	46.10	19.07	11.56	8.81	6.42	3.62	2.52	1.39	0.51
Cumulative	46.10	65.18	76.73	85.55	91.99	95.58	98.10	99.49	100.00

It can be observed that factor loading values in bold - F1 and F2 explains most of the variability associated with the variables. Table 5 shows that F1 accounts for the highest variance of 46.10% while F2 account for 19.07% of the total variance. The F1 variables are strongly correlated with streamflow prediction. Factor 1 has the highest loading for temperature and solar radiation followed by relative humidity. This shows that the evaporation process is the dominating factor for water availability in this semi-arid region.

Figure 2 shows the Basin ombrothermic diagram in relation to its sub-basins water use. It is a line graph plot of ratio Mean Annual Precipitation (MAP) and Mean Annual Evopotranspiration (MAE) for wet and dry periods [9].

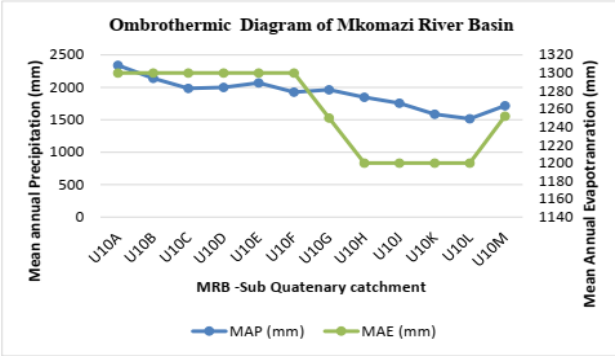


Figure 2: The MRB ombro-thermic diagram.

The result in Figure 2 reveals that the sub-basins U10A-U10F have surplus water while the subsequent sub-basins U10G-U10M have water deficits. The resultant wet and dry periods that prevails in some sub-basin calls for effective crop planning, planting and yield cultivation.

Also, Figure 3 illustrates the FDC and seasonal annual rainfall distribution plot result for the area. The FDC plot of discharge vs. percentage of the time that a particular discharge was equaled or exceeded shows an extreme value of flow magnitude of 0.89mm^3 and a fair seasonality annual fitting for rainfall distribution.

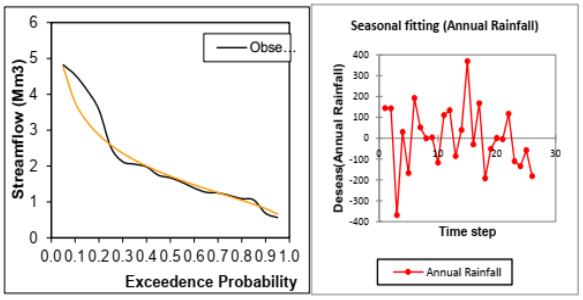


Figure 3: The FDC and seasonal annual rainfall distribution plot.

The crop planning model when maximizing total planting area and minimizing irrigation water using the Excel LP solver is shown in Figure 4.

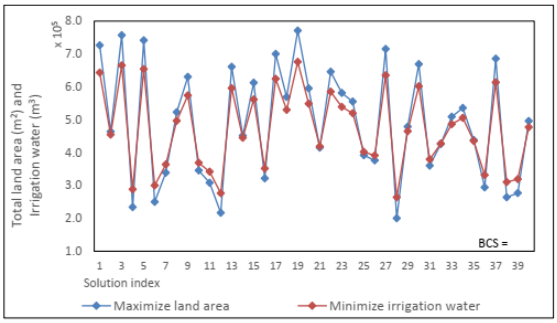


Figure 4: Excel coded solution for maximum total planting area and minimum water usage.

The excel solver optimisation solution indicates the total land area (m^2) and irrigation water (m^3) that was required to grow the prominent crops including maize, groundnuts, lucerne and pecan nuts. The results indicate 7 x 10^{6}m^2 of land planting area will require a minimum amount of 6.5 x10^{6}m^3 of irrigation water to optimise the

four-cropping pattern cultivation. The application of linear optimisation algorithms with the aid of quantitative tools provides an analytical user-friendly model that determines both crop plantation and yield endogenously with efficient water use. The effects of morpho-climatic in recent times reveals changes was witnessed most at the foot slopes of the mountain where large commercial afforestation has been established [14]. This has adversely affected the downstream and ecological systems leading to conflict over water demand, which calls for careful planning to alleviate the growing water challenges especially for agriculture and food production. This study reveals crop water production modelling and that water stress caused by insufficient rainfall plays a significant role in mixed crop production. Thus, optimal water availability, the removal of alien species, adequate landscape control among users; and rehabilitation of the eroded land are crucial.

IV. CONCLUSION

The ever-growing population and the urgency for food and economic advancement calls for efficient use and management of water resources to step up agricultural and industrial production. This study has successfully demonstrated the ability of multivariate statistics and linear programming algorithms to generate a set of the solutions in investigating sustainable water use under the changing climatic conditions and land use in determining optimum crop yield with minimum water requirements. The results obtained are of utmost importance in resolving crop planning problems and for future possibilities of local adaptation in any river basin with similar attributes to the study area in establishing a minimum water requirement baseline in farmlands.

ACKNOWLEDGMENTS

The authors expressed their appreciation to the Department of Water Affairs and Forestry, KZN, South Africa who supported this research with logistic data.

REFERENCES

[1] K. Crowley and A. J. A. van Vuuren, “South Africa Water Expenditure Needs Rise to $71 Billion,” 2013.

[2] M. N. Nkondo, F.C. Van ZYL, H. Keuris and B. Schreiner, “Proposed National Water Resources Strategy 2 (NWRS2): Summary. Cape Town”, Department of Water Affairs, South Africa. 2012.

[3] Department of Water Affairs and Forestry (DWAF), "Volume 3: Ecoclassification and Ewr Assessment On the Mkomazi, Umngeni, and Mvoti Rivers. 7.KwaZulu-Natal Province," Dwaf.Gov.Za 7 Gazette. 2014.

[4] W.A. FLÜGEL and M. MÄRKER, "The response units concept and its application for the assessment of hydrologically related erosion processes in semiarid catchments of Southern Africa," *Spatial Methods for Solution of Environmental and Hydrologic Problems—Science, Policy, and Standardization.* ASTM International, 2003.

[5] W.A. FLÜGEL, M. MÄRKER, S. MORETTI, G. RODOLFI, and A. SIDROCHUK, "Integrating geographical information systems, remote sensing, ground truthing and modeling approaches for regional erosion classification of semi-arid catchments in South Africa," *Hydrological Processes,* 17**,** 929-942, 2003.

[6] FAO, "*Fertilizer use by crop* [Online]," Rome, Italy: Food and Agriculture Organization, 2002, Accessed on 3/8/ 2015.

[7] FAO, *South Africa, "Water and Food Security Country Profiles," Accessed on* 3/8/2015.

[8] K. DEB, "Multi-Objective Optimization Using Evolutionary Algorithms. *An Introduction," Multi-objective Evolutionary Optimisation for Product Design and Manufacturing,* 2011.

[9] FAO, "Irrigation potential in Africa: A basin approach," Land and Water Bulletin 4. Food and Agriculture Organization of the United Nations, Rome. 1997.

[10] J.A.Adeyemo and F.A.O. Otieno, "Multi-objective differential evolution algorithm for solving engineering problems," *Journal of Applied Sciences*, *9*(20), pp.3652-3661, 2009.

[11] B. Grové, "Review of whole-farm economic modeling for irrigation farming," Water SA, 37 (5), pp. 789-796, 2011.

[12] A. SINGH, "Simulation-optimization modeling for conjunctive water use management," *Agricultural Water Management,* 141**,** pp. 23-29, 2014.

[13] H.H. Zhang and D.F. Brown, "Understanding urban residential water use in Beijing and Tianjin, China," *Habitat International*, *29*(3), pp.469-491, 2005.

[14] Department of water and Sanitation (DWS), “*Regional Water Allocation Reform Implementation Plan: KwaZulu-Natal Province,2013 – 2018,”* Department of water and Sanitation; Republic of South Africa, 2016.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

DEEP LEARNING FOR ENVIRONMENTAL SENSING TOWARD SOCIAL WILDLIFE DATABASE

Clement Duhart[1], Spencer Russell[1], Felix Michaud[2], Gershon Dublon[1], Brian Mayton[1], Glorianna Davenport[1], Joseph Paradiso[1]

***Abstract*—Climate change and environmental degradation are causing species extinction worldwide. Automatic wildlife sensing is an urgent requirement to track biodiversity losses on Earth. Recent improvements in machine learning can accelerate the development of large-scale monitoring systems at high resolution that would help track conservation targets and outcomes. This would offer also unique opportunities for studying wildlife sociology at individual scale. In this paper, we present our efforts to develop suitable tools for building machine learning databases for wildlife detection, identification, acoustic source separation and geolocalization. These tools work on data collected at the Tidmarsh Wildlife Sanctuary, the site of the largest freshwater wetland restoration in Massachusetts.**

I. INTRODUCTION

Ubiquitous sensing technologies [1] can be used to capture aspects of ecosystem function and ecological transformation with minimal impact at high resolution over long periods of time. However, in part due to recognition challenges, automatic wildlife sensing remains mostly out of reach. In the ecological research community, wildlife surveys are still conducted by experts estimating a given species population at a specific time. Intensive manual effort is required, even with the help of recordings and modern signal processing tools. Field surveyors need to maintain perceptual awareness and attention to detail; they also need to conduct surveys at different times of day and on many days throughout the year as the animal populations migrate and breed.

Efforts to automate surveys are vital to gaining a real-time understanding of a massive wave of species extinction. This represents a significant opportunity for Artificial Intelligence (AI) systems, which thrive on big data, and might one day be able to analyze and characterize wildlife populations around the globe. Accurate and continuous wildlife detection, identification and geolocalization would transform wildlife surveys into high-resolution activity maps that can update in real time and at geographic scale. Recent advances in Deep Learning enable recognition of rare species that otherwise produce low-occurrence signals in evolving and noisy environments, and a distributed sensing approach to studying Wildlife might also uncover localized interactions and social behaviors that would be difficult to identify and track manually.

Optical and acoustic sensing provide complementary information. For example, the biophony is intrinsically complex in terms of vocalizations (e.g. birds) and diverse regarding species candidates. However, the biophony is mainly produced by creatures that are difficult to spot. A multi-modal sensing approach can help separate noisy geophony and anthrophony from the desired wildlife signal. One crucial requirement is a system's ability to detect new species in an area, especially in a dynamic restoration such as the one presented in Section II.

Some recent contributions have demonstrated the ability of Deep Learning to scale biologists' efforts to identify wildlife. For example, automatic animal identification from camera trap images using a VGG model trained on 1.4 million images over 48 classes was shown to have 96.8% accuracy [2]. AI can be leveraged to save time when used with human volunteers. For acoustic identification of wildlife, several new contributions focus on deep learning technology, including [3] for amphibians, [4] bats, [5] insects, and [6] bird vocalization segmentation. We expect interesting contributions in the future thanks to the Bird Audio Detection challenge [7], [8], [9]. Such lab-based results are essential to accelerating field deployments.

In this paper, we present our own efforts to monitor wildlife activity at the Tidmarsh Wildlife Sanctuary, the site of the one of the largest-ever freshwater wetland restorations in the northeastern United States. In this deployment, our Deep Learning models have been running 24/7 on data streaming from microphones and cameras in real-time over the last 3 years, and our sys-

Corresponding author: C Duhart, duhart@mit.edu [1]Responsive Environment, MediaLab, Massachusetts Institute of Technology, USA [2]Laboratoire d'Acoustique de l'Universite du Mans, FRANCE

tem has been used by biologists, restoration scientists, and other practitioners. These resources also provide an acoustic and visual database for machine learning researchers to build systems that are able to detect, identify and geo-localize wildlife interaction patterns at the individual level. As the capability improves, the system can provide unique scientific data for studying wildlife sociology in natural environments at this critical juncture for shrinking populations and ecosystems.

II. TIDMARSH WILDLIFE SANCTUARY

The Tidmarsh Wildlife Sanctuary is a 485 acre former cranberry farm in south-eastern Massachusetts that was actively restored to a freshwater wetland (2010-2016) . Different types of sensors, illustrated in Figure 1, are permanently deployed on the site to monitor its evolution, including environmental changes to water quality and temperature, wetland surface, stream channels, soils, atmosphere, plants and animal life etc. Such a data collection provides a high resolution environmental map revealing multi-scale dynamic interactions over time. Relying solely on theoretical frameworks to model such a complex environment is challenging: how do we determine chains of cause and effect? For example, how might we link changes to animal behavior to new microbial populations in the soils through intermediate effects on the plant community? One opportunity may be to map an ecosystem across many variables over time, space and multi-scales.

Fig. 1: Sensors are fully autonomous nodes with wireless communication and solar energy harvesting for collecting data from ground and atmospheric probes.

A. Wildlife Sensing Framework

Our 'Tidzam' wildlife detection system monitors wildlife, leveraging 24 custom-designed microphones and 6 cameras deployed across four different areas at Tidmarsh, as illustrated in Figure 2.

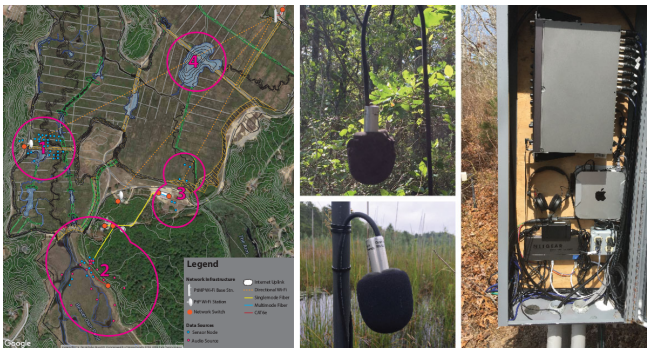


Fig. 2: Sensors, microphones, and cameras are deployed in four regions of interest. They have been specially designed to withstand wetland conditions year-round.

In the Tidzam framework, we implement and deploy Deep Learning techniques from the literature to detect, identify and geolocalize wildlife activities. Over the last four years, we have tested a number of different approaches leveraging bio-acoustics and computer vision.

1) Bio-Acoustic Classifiers: The Tidmarsh bio-acoustic ecosystem has evolved dramatically over years of restoration progress. Dynamic environments require continuous learning to make classifiers robust to both episodic and permanent acoustic changes – especially concerning the identification of as-yet unseen species. To that end, we developed a semi-automatic database augmentation mechanism using a confidence function detailed in [10]. A flow controller limits the recording volume and parameterizes the extraction balance between unidentified and uncertain predictions. Our 'Tidplay' platform, introduced in Section II-B, allows human experts to annotate and discuss these recordings while building a local acoustic database used to iteratively refine the classifiers. At the time of writing, the database is composed of 400,000 500 ms recordings distributed over 66 classes including system failure modes (e.g. microphone crackling due to water ingress), geophonic scenes (e.g. rain, wind, quiet), anthrophonic sounds (e.g. cars, airplanes, human voices), and finally, bio-acoustic events from insects (e.g. crickets, cicadas), amphibians (e.g. spring peepers, green frogs), and bird vocalizations across 42 species.

Several classifier models have been tested, presented in Table I. The classifier is retrained from scratch every 2 months, taking into consideration new recording

annotations. The average accuracy gain increases significantly at each training iteration, with the extent of the improvement depending on the number of new classes, diversity of vocalizations, and quality of the extracted recordings. Our current bio-acoustic classifier is based on a revisited expert architecture [11] running on one Titan X GPU. It continuously analyzes overlapped 500 ms Mel-Spectrogram windows from 24 discrete microphones and 3 on-camera microphones.

Architecture	F_1
64RBM-16RBM + SAE + CE	73%
121C-2P-16C-2P-1024FC-1024FC + CE	85%
121C-2P-16C-2P-3EA(1024FC-1024FC) + CE	88%
121C-2P-16C-2P-1024FC-1024FC + T-Lost	87%

TABLE I: Testing F1 scores on the Tidmarsh dataset using a Restricted Boltzman Machine (RBM) with Stacked Auto-Encoder on Cross-Entropy (CE), Convolution (C) with Pooling (P) and Fully Connected (FC) layers, Expert Architecture (EA) and Triplet-Lost

2) Camera Trap Classifiers: Camera traps use movement detectors to trigger video recording. In an outdoor environment such as Tidmarsh, non-animal movements dominate the trigger. Common causes include rain, wind, and water flow, which together produce a large number of irrelevant video recordings.

Deep Learning can provide high level visual semantic descriptions, saving volunteer time. We have experimented with and deployed different types of computer vision models to pre-filter our motion video databases. These include CNN, Fast R-CNN [12], and Yolo v3 [13]. As illustrated in Figure 3, precise species identification is still a challenging task, given the number of possible classes and the lack of a sufficient training dataset for general-purpose wildlife recognition. It would be a massive challenge to build a database containing every species present on the planet, and it may even be impossible to build a corresponding classifier model. As a result, we use our Tidplay platform to build a locally-dependent visual database to refine the pre-trained classifier model. This platform allows volunteers to create new classes and add new bounding boxes to video frames automatically extracted by a confidence function similar to the one used in our bio-acoustic classifier. Our current system is based on the Yolo v3 model and analyzes video recordings coming from 6 network cameras at Tidmarsh.

B. Tidplay Annotation Platform

Tidplay is an open-source, crowd-sourcing annotation web platform that we have designed to build training

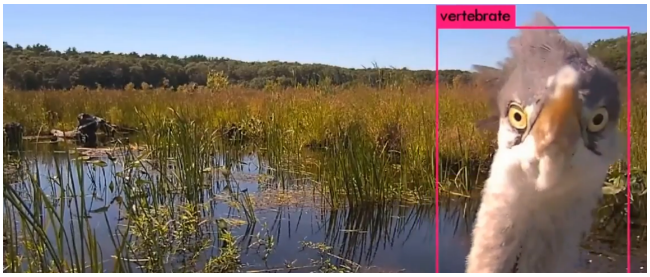


Fig. 3: Wildlife detection on the Herring site by a Yolo v3 model that has not been refined by a local database.

databases from audio and video sources. Users can upload, download and share audio and video files, write down annotations and comments, and create their custom databases while learning about wildlife. Tidplay has two intended user bases. First, wildlife ecologists can use Tidplay to share data for collaborating on the construction of annotated databases. Second, a tutorial mode can be used for public engagement and student training. Users can learn how to distinguish different sounds coming from geophony, anthropophony and biophony, progressively developing their abilities to identify challenging bird calls, for example. The multiple training levels available allow users to extend their bio-acoustic skills by comparing their answers and discussing ambiguous recordings with other users ranging from novices to experts. Recordings extracted automatically by Tidzam classifiers are integrated into Tidplay for cross-validation by multiple wildlife experts before being integrated into training databases. The Tidplay platform can also be used for annotating new audio as shown in Figure 4, for drawing video bounding boxes around objects or wildlife of interest, and for sketching subjects' body pose from video frames, all for use by the machine learning system.

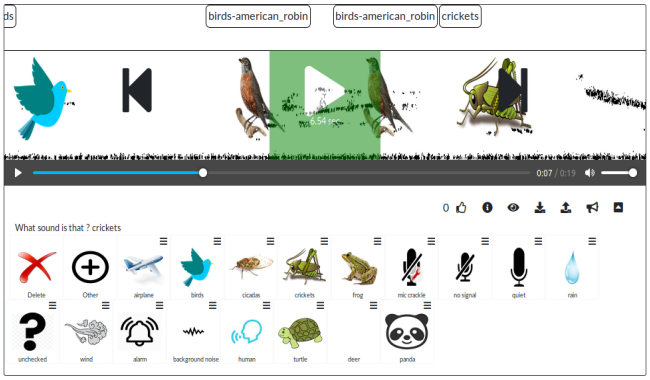


Fig. 4: The audio module of the Tidplay annotation platform shows the recording's spectrogram during listening to facilitate the annotation task.

III. Discussions and Future Work

Our Tidzam framework shows how Deep Learning technology can be used to detect and identify wildlife activities. However, its effectiveness for identifying or tracking individual animals and achieving accurate density estimation is an open question requiring additional data collection and validation. Currently, ecologists at Tidmarsh use correlations between Tidzam's detection density maps and periodic field surveys to estimate the wildlife population dynamics over years of restoration. Our current development effort is focused on how deep learning frameworks can help in tracking individuals through acoustic source separation, and how subjects can be localized more precisely. This work is summarized in the next subsection. Our main goal is to publish a complete database for ecological scientists and machine learning practitioners to study as is and apply in their own settings. The database will contain all the audio recordings from our microphone array with corresponding acoustic event annotations, their source separation masks, and estimated location coordinates.

A. Species Acoustic Source Separation

Recent enhancements have been made in acoustic source separation thanks to Deep Learning technology especially with the U-NET architecture [14]. Our early works based on such approach applied on our recordings is promising as illustrated in Figure 5. We are still investigating model limits regarding acoustic outdoor environment and a procedure for the database constitution regarding of detected species from Tidzam. Live deployment is also a concern regarding computation costs and the various acoustic streams collected from Tidmarsh. On demand analyze should be suitable based on Tidzam species identification.

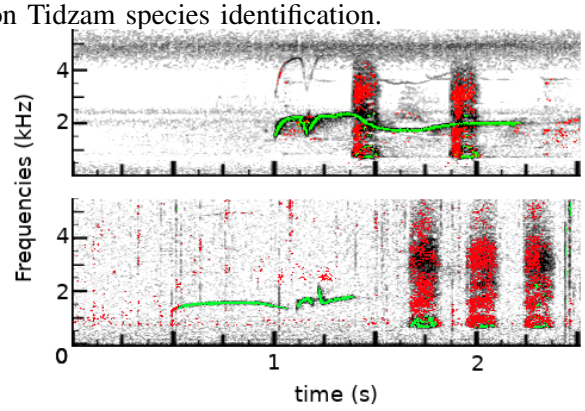


Fig. 5: Two examples of source separation masks when Tidzam has detected two birds : one american crow (red) and one eastern wood pewee (green) with cicadas (top) and rain (bottom) acoustic backgrounds.

B. Individual Acoustic Source Localization

While Tidzam is focused on classifying the the sounds recorded at our microphones, we are also working on methods to use all microphones jointly to localize wildlife within the environment. In addition to giving valuable ecological information, estimating the location of the sound can also inform the source separation process. The vast majority of existing source-separation research operates either in the single-channel context or with a small array with sources in the far-field. For instance, recent work has extended single-channel Deep Clustering using the phase information from an array [15]. However, spaced microphones provide a number of additional challenges because of the longer inter-microphone time delays and varying source-to-microphone propagation paths. Our preliminary work applying the framework of the Spatial Likelihood Function (SLF) [16] has been promising, as you can see in the map in Figure 6. This figure shows a heat map of where a given source (in this case a crow) is likely to be. While so far this work has been focused on classical digital signal processing (DSP) techniques, we are working on several approaches to integrating the DSP and AI frameworks to improve our localization estimates, as well as designing experiments to more rigorously evaluate the performance of our algorithms.

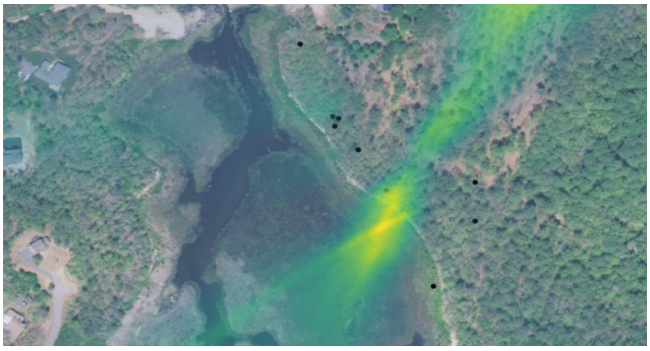


Fig. 6: Spatial Likelihood Function based on audio at a subset of the deployed microphones (black dots). The displayed area is roughly 500m x 350m

IV. Conclusion

We have presented an ongoing effort to deploy Deep Learning tools for automatic wildlife surveying Our work shows how Deep Learning can advance significant opportunities for ecological research efforts, restoration science, and public engagement. Our long term goal is to distribute an annotated database for machine learning practitioners, providing an unprecedented view towards analyzing wildlife in a restoration setting at the individual and species interaction level.

REFERENCES

[1] J. Paradiso, "Our extended sensoria - how humans will connect with the internet of things," *The Next Step: Exponential Life, Open Mind Collection*, vol. 1, no. 1, p. 47–75, 2016.

[2] M. S. Norouzzadeh, A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer, and J. Clune, "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning," *Proceedings of the National Academy of Sciences*, vol. 115, no. 25, pp. E5716–E5725, 2018.

[3] J. Strout, B. Rogan, S. M. M. Seyednezhad, K. Smart, M. Bush, and E. Ribeiro, "Anuran call classification with deep learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2662–2665, March 2017.

[4] O. Mac Aodha, R. Gibb, K. E. Barlow, E. Browning, M. Firman, R. Freeman, B. Harder, L. Kinsey, G. R. Mead, S. E. Newson, I. Pandourski, S. Parsons, J. Russ, A. Szodoray-Paradi, F. Szodoray-Paradi, E. Tilova, M. Girolami, G. Brostow, and K. E. Jones, "Bat detectivedeep learning tools for bat acoustic signal detection," *PLOS Computational Biology*, vol. 14, pp. 1–19, 03 2018.

[5] I. Kiskin, B. P. Orozco, T. Windebank, D. Zilli, M. Sinka, K. J. Willis, and S. J. Roberts, "Mosquito detection with neural networks: The buzz of deep learning," *CoRR*, vol. abs/1705.05180, 2017.

[6] I. Potamitis, "Deep learning for detection of bird vocalisations," *CoRR*, vol. abs/1609.08408, 2016.

[7] D. Stowell, M. D. Wood, H. Pamua, Y. Stylianou, and H. Glotin, "Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge," *Methods in Ecology and Evolution*, vol. 10, no. 3, pp. 368–380, 2019.

[8] E. Cakir, S. Adavanne, G. Parascandolo, K. Drossos, and T. Virtanen, "Convolutional recurrent neural networks for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1744–1748, Aug 2017.

[9] S. Adavanne, K. Drossos, E. akir, and T. Virtanen, "Stacked convolutional and recurrent neural networks for bird audio detection," in *2017 25th European Signal Processing Conference (EUSIPCO)*, pp. 1729–1733, Aug 2017.

[10] C. Duhart, G. Dublon, B. Mayton, and J. Paradiso, "Deep learning locally trained wildlife sensing in real acoustic wetland environment," in *Advances in Signal Processing and Intelligent Recognition Systems* (S. M. Thampi, O. Marques, S. Krishnan, K.-C. Li, D. Ciuonzo, and M. H. Kolekar, eds.), (Singapore), pp. 3–14, Springer Singapore, 2019.

[11] M. I. Jordan and R. A. Jacobs, "Hierarchies of adaptive experts," in *Advances in Neural Information Processing Systems*, pp. 985–992, Morgan-Kaufmann, 1992.

[12] R. Girshick, "Fast r-cnn," in *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[13] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018.

[14] A. Jansson, E. Humphrey, N. Montecchio, R. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks," *In proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017.

[15] Z.-Q. Wang, X. Zhang, and D. Wang, "Robust TDOA estimation based on time-frequency masking and deep neural networks," in *Interspeech 2018*, p. 322326, ISCA, Sep 2018.

[16] P. Aarabi, "The fusion of distributed microphone arrays for sound localization," *EURASIP Journal on Applied Signal Processing*, vol. 2003, p. 338347, 2003.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Living Observatory and the Mass Audubon Tidmarsh Wildlife Sanctuary for the opportunity to realize the audio deployment at this location. The NVIDIA GPU Grant Program has provided the two TITAN X which are used by Tidzam. Clement DUHART has been supported by the PRESTIGE Fellowship of Campus France and the Pôle Léonard de Vinci. We also thank the Elements Collaborative and the sponsors of the MIT Media Lab for their support of this work.

CONNECTIONS BETWEEN DATA ASSIMILATION AND MACHINE LEARNING TO EMULATE A NUMERICAL MODEL.

Julien Brajard[1,2], Marc Bocquet[3], Alberto Carrassi[1,4], Laurent Bertino[1]

***Abstract*—Is it possible to emulate a numerical model from noisy and sparse observations? How realistic and skilful can it be? Recent progress in machine learning has shown how to forecast a model from observations. We will show that by leveraging on data assimilation techniques, it is possible to produce realistic and skilful surrogate models of the underlying dynamics given sparse and noisy observations. It is also shown that data assimilation is equivalent to a machine learning problem and that it can efficiently be combined with classical deep learning algorithms. The approach is illustrated with a multi-dimensional chaotic system. The surrogate model shows both forecast skills and abilities to reproduce the climate (i.e. spectral properties and statistical moments) of the underlying dynamical model on long-term simulations.**

I. INTRODUCTION

In geophysics, numerical models based on physical laws are widely used both to simulate and forecast the Earth system. Deficiencies in these models can be corrected using parametrisations or sub-models driven by observations. This work is thus addressing the general problem of emulating a numerical model from observations. This objective was addressed in two seemingly different fields: data assimilation (DA) for error models [1], [2], [3], [4], [5], [6], [7] and machine learning (ML) trained on observations [8], [9], [10], [11], [12], [13], [14], [15], [16]. This work connects these two fields by linking the results presented separately in two previous papers (see [17], [18]). In both papers, DA is used to build a surrogate model from sparse and noisy observations. In the first [17], it is shown how to infer an ordinary differential equation (ODE) using DA by specifying only the very general form of the ODE. This is shown to be equivalent to a ML regression algorithm. In the second [18], DA is used in combination with a convolutional neural network to emulate the underlying numerical model from the observations. In the present paper, some results obtained in emulating numerical models will highlight the strong connections between machine learning and data assimilation.

Corresponding author: J. Brajard julien.brajard@locean-ipsl.upmc.fr [1]Nansen Environmental and Remote Sensing Center , Bergen, Norway [2]Sorbonne University, Paris, France [3]CEREA, joint laboratory École des Ponts ParisTech and EDF R&D, Université Paris-Est, Champs-sur-Marne, France [4]Geophysical Institute, University of Bergen, Norway

II. STATEMENT OF THE PROBLEM

Let us consider a time series of multi-dimensional observations $\mathbf{y}_k^{\text{obs}} \in \mathbb{R}^{N_y}$ of an unknown process $\mathbf{x}_k \in \mathbb{R}^{N_x}$:

$$\mathbf{y}_k^{\text{obs}} = \mathcal{H}_k(\mathbf{x}_k) + \boldsymbol{\epsilon}_k^{\text{obs}}, \tag{1}$$

where $0 \leq k \leq K$ is the index corresponding to the observation time t_k, and $\mathcal{H}_k : \mathbb{R}^m \to \mathbb{R}^p$ is the observation operator (supposed to be known). The observation error $\boldsymbol{\epsilon}_k^{\text{obs}}$ is assumed to follow a normal distribution of zero mean and covariance matrix $\mathbf{R}_k$, supposed diagonal of the form $\mathbf{R} = \sigma_y^2\mathbf{I}$. For the sake of simplicity, we will also consider a regular time discretisation such that: $t_{k+1} - t_k = \Delta t$ for all k.

We suppose that $\mathbf{x}_k$ is a time-discretisation of a continuous process $\mathbf{x}$, which obeys an unknown ordinary differential equation (ODE) of the form

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \boldsymbol{\phi}(\mathbf{x}), \tag{2}$$

where $\boldsymbol{\phi}(\mathbf{x})$, called the flow-rate, is unknown.

Forecasting $\mathbf{x}_{k+1}$ from $\mathbf{x}_k$ can be achieved by integrating the flow-rate between t_k and t_{k+1} yielding the following "resolvent" of the model:

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \int_{t_k}^{t_{k+1}} \boldsymbol{\phi}(\mathbf{x})\,\mathrm{d}t. \tag{3}$$

The algorithms described in the following sections will be illustrated using the 40-variable Lorenz model [19] hereafter denoted L96, which is used to generate the synthetic observations. In this idealized

case, the surrogate model can be compared with the real underlying dynamics, called the "true" model. The model L96 is defined on a periodic one-dimensional domain by the following set of ODEs:

$$\frac{\mathrm{d}x_n}{\mathrm{d}t} = (x_{n+1} - x_{n-2})x_{n-1} - x_n + F, \qquad (4)$$

where x_n ($0 \leq n < N_x$) is the scalar state variable, $x_{N_x} = x_0$, $x_{-1} = x_{N_x-1}$, $x_{-2} = x_{N_x-2}$, $N_x = 40$ and $F = 8$. The model is integrated using a fourth order Runge-Kutta scheme (RK4) with an integration time step h equal to the observation time step $h = \Delta t = 0.05$. The L96 model with the current choices for m and F is chaotic, with the largest Lyapunov exponent being $\Lambda_1 \approx 1.67$. In the following, we use the Lyapunov time unit $t_\Lambda = \Lambda_1 t$ where t is the time in model unit: one Lyapunov time unit corresponds to the characteristic time for the error to grow by a factor e. An independent integration of the model given other initial conditions is use in the following as a test dataset to compute the score presented in the following.

The paper is organised in two parts: The ODE will first be inferred relying on a parametric expression of the flow rate. This first part is based on a data assimilation framework (a weak constraint 4D-VAR). We will show that the approach is equivalent to a standard ML regression problem. In the second part, we will combine a data assimilation algorithm with a neural network to emulate directly the resolvent of the model without inferring the flow-rate itself.

III. Inferring the ODE

A. Representing the ODE

We define a surrogate model based on an ODE as in Eq. (2) assuming that the flow rate can be written in the form

$$\boldsymbol{\phi}_{\mathbf{A}}(\mathbf{x}) = \mathbf{A}\mathbf{r}(\mathbf{x}), \qquad (5)$$

where $\mathbf{A} \in \mathbb{R}^{N_x \times N_p}$ is a matrix of real coefficients to be estimated and $\mathbf{r} : \mathbb{R}^{N_x} \mapsto \mathbb{R}^{N_p}$ is a map that defines regressor functions of $\mathbf{x}$. $\mathbb{R}^{N_p}$ is the latent space in which the flow rate is linear. We choose to build a map of monomials up to second-order functions of $\mathbf{x}$:

$$\mathbf{r}(\mathbf{x}) = \left[1, \{x_n\}_{0 \leq n < N_x}, \{x_n x_m\}_{0 \leq n \leq m < N_x}\right], \qquad (6)$$

where x_n is the n-th component of the vector $\mathbf{x}$. In absence of additional assumptions, the number of regressors is $N_p = \binom{N_x+1}{2} = \frac{1}{2}(N_x+1)(N_x+2)$. In the case the state $\mathbf{x}$ is the discretisation of a spatial field, we add two assumptions to reduce the number of regressors:

- Physical locality of the dynamics: all multivariate monomials in the ODEs have variables x_n that belong to a stencil, i.e. a local arrangement of grid points around a given node. In 1D and with a stencil of size $2L+1$, the size of the dense $\mathbf{A}$ is $N_x \times N_\mathrm{a}$ where $N_\mathrm{a} = \sum_{l=L+1}^{2L+2} l = \frac{3}{2}(L+1)(L+2)$.
- $\boldsymbol{\phi}_{\mathbf{A}}$ is invariant by translation. $\mathbf{A}$ thus becomes a vector of size N_a.

Given the ODE, the surrogate model is integrated to simulate the field for the whole observational period $\mathbf{x_{0:K}}$. To compute the resolvent of the model between t_k and t_{k+1} in Eq. (3) we assume that Δt is a multiple of the integration time step h such as $\Delta t = N_c h$. We will consider two integration schemes: RK4 (i.e. the same as the true model) or a second order Runge-Kutta (RK2). The resolvent between two observation times is denoted $\mathbf{F_A}$ such as

$$\mathbf{x}_{k+1} = \mathbf{F}_{\mathbf{A}}(\mathbf{x}_k) + \boldsymbol{\epsilon}_k^\mathrm{m}, \qquad (7)$$

where $\boldsymbol{\epsilon}_k^\mathrm{m}$ are unbiased Gaussian errors of covariance matrices $\mathbf{Q}_k$.

B. Optimization

The two quantities to be determined are the parameters of the ODE $\mathbf{A}$ and the state of the system $\mathbf{x_{0:K}}$. In a Bayesian framework, we are seeking the maximum of probability density function $p(\mathbf{A}, \mathbf{x}_{0:K} | \mathbf{y}_{1:K})$. If we neglect the prior probability on $\mathbf{A}$ and $\mathbf{x_{0:K}}$, under the Markovian assumption of Eq. (7) and the normal distribution of model and observational errors, the problem is equivalent to the minimisation of the cost function

$$\mathcal{J}(\mathbf{A}, \mathbf{x}_{0:K}) = \sum_{k=0}^{K} \left\| \mathbf{y}_k^\mathrm{obs} - \mathcal{H}_k(\mathbf{x}_k) \right\|^2_{\mathbf{R}_k^{-1}} + \sum_{k=1}^{K} \left\| \mathbf{x}_k - \mathbf{F}_{\mathbf{A}}(\mathbf{x}_{k-1}) \right\|^2_{\mathbf{Q}_k^{-1}}. \qquad (8)$$

The cost function is minimised using a quasi-Newton BFGS optimiser relying on the computation of the gradient of $\mathcal{J}$. Using adjoint modelling, the gradient can be explicitly derived from the Runge-Kutta integration scheme and the parametric form of the ODE given in Eq. (5). More details on this formal computation can be found in [17].

C. Results

The first numerical experiment has been conducted using an identical integration scheme (RK4) and time step ($h = \Delta t$, i.e. $N_c = 1$) for the surrogate model and the true model. In this favourable case, the numerical

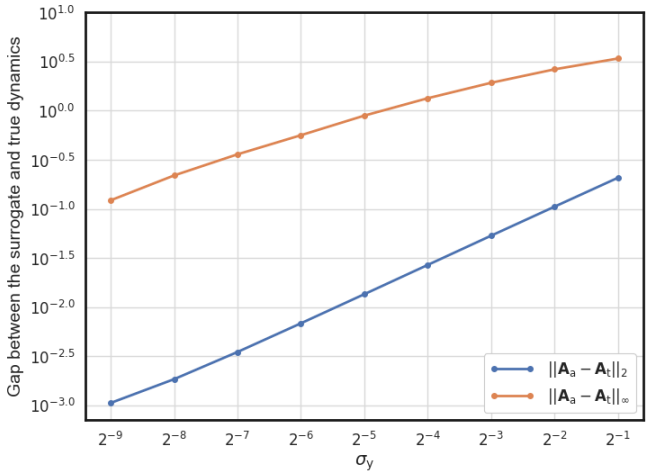


Fig. 1: Gap (matrix norm of the difference of the true parameter $\mathbf{A}_t$ with the parameter of the surrogate model $\mathbf{A}_a$) between the surrogate and the (identifiable) true dynamics as a function of the observation error standard deviation σ_y.

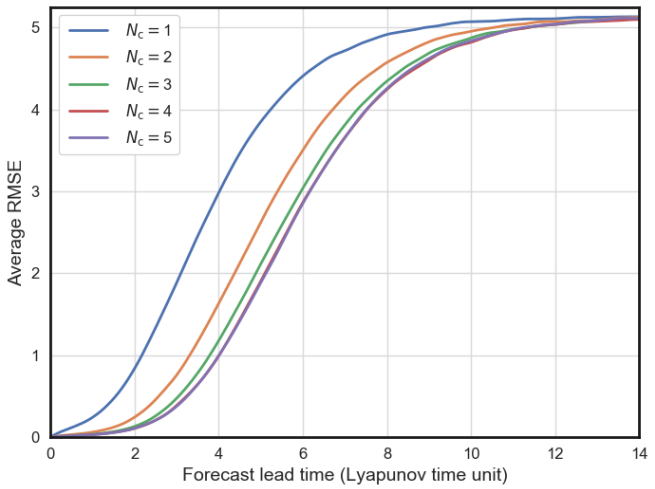


Fig. 2: Average root mean square error (RMSE) of the surrogate model compared to the true model as a function of the forecast lead time (in Lyapunov time unit) for an increasing number of compositions.

model is said to be "identifiable". The field is fully observed ($\mathcal{H}_k$ is the identity) and $\epsilon_k^m = 0$ (no observation noise). We thus retrieve the ODE parameters with an error of 10^{-13} (infinity norm), which is a perfect reconstruction up to the numerical machine precision. We have also introduced observational noise ranging from 2^{-9} to 2^{-1}. The retrieval accuracy of the ODE parameters is shown in Fig. 1. The parameters retrieval degrades with increasing noise but the errors remain low. Another experiment was conducted with $\epsilon_k^m = 0$ and an RK2 integration scheme with N_c ranging from 1 to 5. Note that Δ_t being fixed, defined by the observation sampling rate, higher values of N_c correspond to smaller integration time steps h. In this case, the model is no longer identifiable but it is still possible to assess the forecast skill of the surrogate model. The results in Fig. 2 show that the surrogate model forecast skills improve with a smaller integration time step but the performances saturate with $h \leq \Delta t/4$. More results with partial and noisy observations and with other dynamical systems can be found in [17].

IV. CONNECTION BETWEEN DATA ASSIMILATION AND DEEP-LEARNING

We describe here how the DA approach described previously is equivalent to a ML problem. DA, when implemented properly, is faithful to Bayes' rule. Therefore, ML, when equivalent to DA, should benefit from a similar Bayesian interpretation with all its benefits (e.g., it should be able to provide uncertainty estimates, although this capability is not exploited systematically).

It was shown in [20], [21], that a dynamical system can be represented by residual blocks in a neural network. Similarly, each time step integration of the model can be seen as a layer in a deep learning architecture. The observations $\mathbf{y_{1:K}}$ used in the minimisation constitute the training dataset of ML. By considering the case $\mathbf{y}_k = \mathbf{x}_k$ ($\mathcal{H}_k$ is the identity and observation noise is zero), the cost function in Eq. (8) becomes

$$\mathcal{J}(\mathbf{A}) = \sum_{k=1}^{K} \|\mathbf{x}_k - \mathbf{F}_\mathbf{A}(\mathbf{x}_{k-1})\|^2_{\mathbf{Q}_k^{-1}}\,, \qquad (9)$$

which is the standard regression cost function in machine learning. The adjoint modelling used for the optimization is equivalent to a gradient backpropagation. The same equivalence was already highlighted from the point of view of ML in [22], [23]

The assimilation of observations is equivalent to a training phase, and the use of the model for forecasting is equivalent to the inference of a neural network. Finally, the locality assumption plays a similar role to that of convolutional layers.

V. EMULATION OF THE NUMERICAL MODEL

A. Algorithm description

The other algorithm presented in [18] is aiming at emulating directly the resolvent of the model described

in Eq. (3) through a neural network expressed as the parametric function

$$\mathbf{x}_{k+1} = \mathcal{G}_{\mathbf{W}}(\mathbf{x}_k) + \boldsymbol{\epsilon}^{\mathrm{m}}_k, \tag{10}$$

which plays the role of the dynamical system described in Eq. (7). The neural architecture used in this work is described in Fig. 3. Note that following the remark made in section IV, the neural network is a residual architecture, whose details can be found in [18]. As the numerical experiments are conducted on a periodic spatial domain, (see section II), the input of the neural network is 1-D periodic. The resulting number of weights to estimate is 9391. Because the weights have a priori no obvious spatial structure, no local or homogeneity assumptions on the weights can be made here contrary to section III. Still, convolutive neural networks are acting locally and homogeneously on the input field. As a consequence, the locality and homogeneity assumptions made in the previous section are still implicitly enabled through the choice of the architecture. The hyperparameters of the neural network (e.g. number/type of layers, activation functions, optimizer and others) has been determined by cross-validation experiments

To estimate both $\mathbf{x_{0:K}}$ and $\mathbf{W}$, the proposed algorithm iterates over a two-step cycle as described in Fig. 4. In the first step, the neural network is fixed as a forecasting model and a finite-size ensemble Kalman filter (EnKF-N) [24] is used as a DA technique to estimate $\mathbf{x}^{\mathrm{a}}_{\mathbf{0:K}}$. However, our approach is not tied to any particular DA algorithm and is straightforwardly applicable to any adequate DA method for the problem at hand. For example, a smoother would lead to more accurate results but would be more costly and is thus less common in the operational DA community.

In the second step, the dataset $\mathbf{x}^{\mathrm{a}}_{\mathbf{0:K}}$ estimated by DA in the previous step is used as a training set to estimate the weights $\mathbf{W}$ of the neural net.

One drawback of the method is the computation cost. The algorithm needs to apply one complete data assimilation procedure (equivalent to 30 forward model runs to propagate the ensemble members in our case) and one neural network training (equivalent to 20 forward model runs in our case) for each cycle until convergence. In this work, we have not optimized the computational cost of the algorithm, but there are several avenues for improvement of that aspect. We could, for example, set an optimal stopping criterion, or leverage from concurrent computing (the neural network could start training before the end of the data assimilation, using the portion of the assimilation

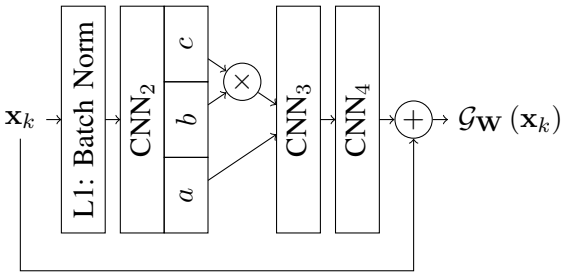


Fig. 3: Architecture of the residual neural network used as a surrogate model (9391 weights).

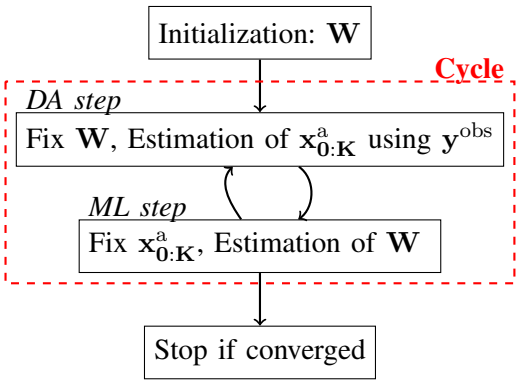


Fig. 4: Scheme of the algorithm. The two-step procedure DA (data assimilation) followed by ML (machine learning) constitutes one cycle of the algorithm.

run which is already analysed). We could also more systematically benefit from accelerations provided by the deep-learning libraries (e.g. GPU).

B. Results

This two-step algorithm is tested with the L96 model. The observation operator $\mathcal{H}_k$ is defined as a sub-sampling operator that draws randomly $N_y = 20$ values at each time step (corresponding to 50% of the field) from a uniform distribution changing the observation locations at each time step. The observation interval is the same as the integration time step of the model. The standard deviation of the observational error in Eq. (1) is $\sigma_y = 1$ (about 5% of the model's variability). The code for running this experiment is publicly available[1].

Figure 5 shows a simulation of the true model (top panel) compared with the one of the surrogate model (middle panel) given by the neural network after optimisation. The differences (bottom panel) between the true and the surrogate simulations increase with

[1]https://doi.org/10.5281/zenodo.2925547

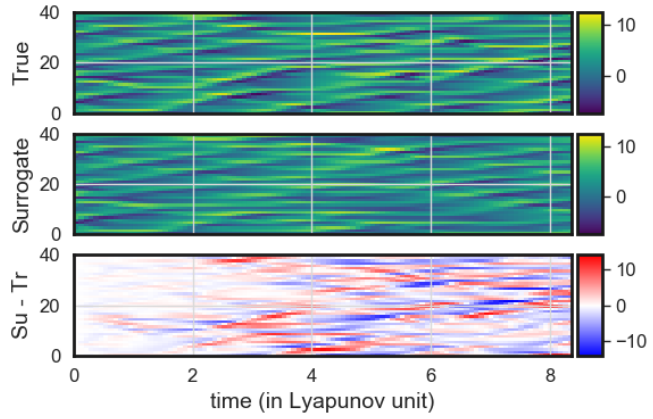


Fig. 5: Hovmøller plot of one trajectory of the true model, the surrogate model and their difference given the same initial condition as a function of lead time.

respect to time, as expected for a chaotic model. The surrogate model shows forecast skills up to 2 Lyapunov time units. Then the trajectory of the surrogate model diverges significantly from the true model and the errors saturate after 4-5 Lyapunov time units.

Further results and metrics can be found in [18]. In particular, it is shown that the surrogate model is also able to reproduce the long-term properties of the dynamical system over long runs, such as the positive Lyapunov exponents.

VI. CONCLUSION

In this work, we have presented links between data assimilation and machine learning supported by two algorithms producing a surrogate model of a system given only partial and noisy observations.

It is therefore natural to use data assimilation in a machine learning framework: both approaches rely on the same theoretical background and are naturally supported by a Bayesian framework. In term of practical implementation, it has been shown that different combinations of machine learning algorithms (e.g. convolutional neural networks) and data assimilation methods (both the variational 4D-Var and the sequential EnKF-N) can produce skilful and reliable surrogate models.

REFERENCES

[1] R. Aster, B. Borchers, and C. Thurber, *Parameter Estimation and Inverse Problems (International Geophysics)*. Elsevier, 2005.

[2] Y. Trémolet, "Accounting for an imperfect model in 4D-Var," *Quarterly Journal of the Royal Meteorological Society*, vol. 132, pp. 2483–2504, 10 2006.

[3] A. Carrassi and S. Vannitsem, "State and parameter estimation with the extended Kalman filter: An alternative formulation of the model error dynamics," *Quarterly Journal of the Royal Meteorological Society*, vol. 137, no. 655, pp. 435–451, 2011.

[4] M. Bocquet, "Parameter-field estimation for atmospheric dispersion: Application to the Chernobyl accident using 4D-Var," *Quarterly Journal of the Royal Meteorological Society*, vol. 138, no. 664, pp. 664–681, 2012.

[5] J. J. Ruiz, M. Pulido, and T. Miyoshi, "Estimating Model Parameters with Ensemble-Based Data Assimilation: A Review," *Journal of the Meteorological Society of Japan. Ser. II*, vol. 91, no. 2, pp. 79–99, 2013.

[6] P. N. Raanes, A. Carrassi, and L. Bertino, "Extending the square root method to account for additive forecast noise in ensemble methods," *Monthly Weather Review*, vol. 143, no. 10, pp. 3857–3873, 2015.

[7] P. Sakov, J.-M. M. Haussaire, and M. Bocquet, "An iterative ensemble Kalman filter in the presence of additive model error," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, pp. 1297–1309, 4 2018.

[8] D. Park and Yan Zhu, "Bilinear recurrent neural network," in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 3, pp. 1459–1464, IEEE, 2002.

[9] D. C. Park, "A time series data prediction scheme using bilinear recurrent neural network," in *2010 International Conference on Information Science and Applications, ICISA 2010*, pp. 1–7, IEEE, 2010.

[10] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems 28* (C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, eds.), pp. 802–810, Curran Associates, Inc., 2015.

[11] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, "Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data," *Chaos*, vol. 27, p. 121102, 12 2017.

[12] X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Deep learning for precipitation nowcasting: A benchmark and a new model," in *Advances in Neural Information Processing Systems*, pp. 5617–5627, 2017.

[13] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz, "Chaos as an intermittently forced linear system," *Nature Communications*, vol. 8, p. 19, 12 2017.

[14] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, "Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach," *Physical Review Letters*, vol. 120, p. 024102, 1 2018.

[15] E. de Bezenac, A. Pajot, and P. Gallinari, "Deep learning for physical processes: Incorporating prior scientific knowledge," *arXiv preprint arXiv:1711.07970*, 2017.

[16] R. Fablet, S. Ouala, and C. Herzet, "Bilinear residual neural network for the identification and forecasting of dynamical systems," in *EUSIPCO 2018, European Signal Processing Conference*, (Rome, Italy), pp. 1–5, 2018.

[17] M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino, "Data assimilation as a learning tool to infer ordinary differential equation representations of dynamical models," *Nonlinear Processes in Geophysics*, vol. 26, no. 3, pp. 143–162, 2019.

[18] J. Brajard, A. Carrassi, M. Bocquet, and L. Bertino, "Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: a case study with the lorenz 96 model," *Geoscientific Model Development Discussions*, vol. 2019, pp. 1–21, 2019.

[19] E. N. Lorenz and K. A. Emanuel, "Optimal sites for supplementary weather observations: Simulation with a small model," *Journal of the Atmospheric Sciences*, vol. 55, no. 3, pp. 399–414, 1998.

[20] B. Chang, L. Meng, E. Haber, F. Tung, and D. Begert, "Multi-level Residual Networks from Dynamical Systems View," *arXiv preprint arXiv:1710.10348*, 2017.

[21] W. E, "A Proposal on Machine Learning via Dynamical Systems," *Communications in Mathematics and Statistics*, vol. 5, pp. 1–11, 3 2017.

[22] W. W. Hsieh and B. Tang, "Applying neural network models to prediction and data analysis in meteorology and oceanography," *Bulletin of the American Meteorological Society*, vol. 79, no. 9, pp. 1855–1870, 1998.

[23] H. D. Abarbanel, P. J. Rozdeba, and S. Shirman, "Machine Learning: Deepest learning as statistical data assimilation problems," *Neural Computation*, vol. 30, no. 8, pp. 2025–2055, 2018.

[24] M. Bocquet, P. N. Raanes, and A. Hannart, "Expanding the validity of the ensemble Kalman filter without the intrinsic need for inflation," *Nonlinear Processes in Geophysics*, vol. 22, pp. 645–662, 11 2015.

PREDICTING ANALOG FORECASTING ERRORS USING DYNAMICAL SYSTEMS

Paul Platzer[1,2,3], Pascal Yiou[1], Pierre Tandeo[2], Philippe Naveau[1], Jean-François Filipot[3]

***Abstract*—Analog forecasting has been used to produce short-range to long-range forecasts in many atmospheric and oceanic applications. Analog forecasting is often treated as a purely empirical method, independent from physical equations. In this paper, we investigate analog forecasting error from a dynamical systems point of view, linking data-driven and model-driven predictions. Assuming that analogs follow the same dynamics as the system of concern, we evaluate statistical properties of analog forecasting errors. We further design dynamics-based systematic error correction methods for standard analog forecasting techniques. These procedures are tested on the 3-dimensional Lorenz-63 system.**

I. MOTIVATION

Two states of a system are called "analogs" when they meet a similarity criterion such as a low Euclidean distance. Analog forecasting (AF) is based on the assumption that similar states will have similar evolution, and produces forecasts based on the "successors" in time of the analogs of the current state. AF has been used in a wide variety of applications in atmospheric prediction, see for instance [1], [2] or [3]. Although AF is less precise than forecasting methods based on the numerical resolution of physical equations, AF can provide statistical forecasts at a low computational cost, outperforming persistence and climatological forecasts. Also, increasing observational data and computer memory make analogs a promising forecast tool.

Although the concept of analogs was originally introduced by [4] to gain information on the dynamics of the atmosphere, today AF is mostly used as a purely empirical method and the link between AF and the underlying dynamics of the system is rarely mentioned. However [5] used a dynamical systems framework to study analogs, but focusing on recurrence time statistics rather than AF performance. Using general properties of dynamical systems, [6] showed that the skills of AF depend on the weights' type and normalization. They used adapted weights based on dynamical properties. Here we make the link between the dynamical equations and AF error.

Here we propose to improve our understanding of forecasting errors associated with AF by expressing them as a function of the dynamics of the system of concern. We assume, as a first step, that the analogs and the system follow exactly the same dynamics. With calculations similar to [7], we give a simple approximation for the error of standard AF techniques for small lead times. This enables us to evaluate AF performance. We further propose two systematic error correction methods for standard AF techniques. We test our methods on the famous Lorenz-63 system [8] in numerical experiments.

This study aims at bridging the gap between purely data-driven AF and purely model-driven methods. We show that understanding the role of the system's dynamics can not only provide statistical information about AF error, but also help improving AF performances.

II. METHOD

Let $\mathbf{x}_0$ represent the system state at time $t = 0$. We are interested in estimating $\mathbf{x}_t$, for time $t > 0$. AF starts with finding a finite number of analogs of $\mathbf{x}_0$ inside a large database called the catalog. We note $\mathbf{a}_0^k$ the k-th analog of $\mathbf{x}_0$. Then AF considers their successors at time t, where the k-th is noted $\mathbf{a}_t^k$. A common AF technique is the Locally Constant (LC) forecast [9], which takes a weighted average of all successors (with weights ω_k) as a forecast. LC forecast is written:

$$\mathbf{LC}_t = \sum_k \omega_k \mathbf{a}_t^k \,. \tag{1}$$

The Locally Incremental (LI) forecast [9] makes use of the differences between successors and analogs called "increments", i.e. $(\mathbf{a}_t^k - \mathbf{a}_0^k)$, rather than successors $\mathbf{a}_t^k$. LI takes the sum of the initial state $\mathbf{x}_0$, and the weighted average of increments $\sum_k \omega_k(\mathbf{a}_t^k - \mathbf{a}_0^k)$. It can also be written:

Corresponding author: P Platzer, paul.platzer@imt-atlantique.fr
[1]Laboratoire des Sciences du Climat et de l'Environnement, Saclay, France [2]IMT Atlantique, Lab-STICC, UMR CNRS 6285, F-29238, Plouzané, France [3]France Énergies Marines, Plouzané, France

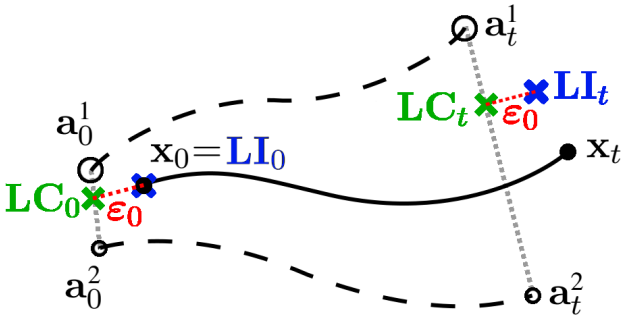


Fig. 1. Illustration of two basic AF techniques. Full circles: $\mathbf{x}_0$ and $\mathbf{x}_t$. Circles: analogs (with larger circles for larger weights). Green crosses: $\mathbf{LC}_0$ and $\mathbf{LC}_t$. Blue crosses: $\mathbf{LI}_0$ and $\mathbf{LI}_t$ (AF techniques). Full line: dynamical trajectory from $\mathbf{x}_0$ to $\mathbf{x}_t$. Dashed lines: analog trajectories. Dotted red lines: $\boldsymbol{\varepsilon}_0$. Dotted grey lines: straight lines between analogs.

$$\mathbf{LI}_t = \mathbf{LC}_t - \boldsymbol{\varepsilon}_0, \tag{2}$$

where $\boldsymbol{\varepsilon}_0 = \mathbf{LC}_0 - \mathbf{x}_0$. Fig. 1 illustrates LC and LI in the simple case of two analogs.

We suppose that $\mathbf{x}$ follows the dynamical equation $\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{f}(\mathbf{x})$, as well as all $\mathbf{a}^k$. We write $\mathbf{J}_t = \nabla\mathbf{f}(\mathbf{x}_t)^{\mathrm{T}}$ (T-superscript is the transpose and ∇ is the gradient operator) the Jacobian matrix of $\mathbf{f}$ along the trajectory $\mathbf{x}$ at time t. If we further assume that $\omega_k\|\mathbf{a}_0^k - \mathbf{x}_0\|$ is small enough (up to a given norm) for all k, the error for the LC forecast is:

$$\mathbf{LC}_t - \mathbf{x}_t \approx \left[\mathbf{I} + t\mathbf{J}_0 + \frac{t^2}{2}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0)\right]\boldsymbol{\varepsilon}_0, \tag{3}$$

and the error for the LI forecast is:

$$\mathbf{LI}_t - \mathbf{x}_t \approx \left[t\mathbf{J}_0 + \frac{t^2}{2}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0)\right]\boldsymbol{\varepsilon}_0. \tag{4}$$

where $\mathbf{I}$ is the identity matrix and $\dot{\mathbf{J}}_0 = \frac{\mathrm{d}\mathbf{J}_0}{\mathrm{d}t}$. The derivation of these formulae is similar to the calculations of [7]. From those formulae we can:

1) infer statistical properties of AF errors associated with LC-type and LI-type AF techniques,
2) apply a systematic error correction to improve the skills of those techniques.

From Eqs. (3–4), it follows that LF and LI are unbiased as long as $\boldsymbol{\varepsilon}_0$ is of zero mean. Squared error of LC can be evaluated up to the order t^2 by taking the square of Eq. (3):

$$\begin{aligned}\|\mathbf{LC}_t - \mathbf{x}_t\|^2 \approx \boldsymbol{\varepsilon}_0^{\mathrm{T}}\{\,&\mathbf{I} + 2t\mathbf{J}_0 + \\ &t^2(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0 + \mathbf{J}_0^{\mathrm{T}}\mathbf{J}_0)\,\}\,\boldsymbol{\varepsilon}_0\,,\end{aligned} \tag{5}$$

and taking the square of Eq. (4) gives the square error of LI up to the order t^3:

$$\|\mathbf{LI}_t - \mathbf{x}_t\|^2 \approx \boldsymbol{\varepsilon}_0^{\mathrm{T}}\left\{t^2\mathbf{J}_0^{\mathrm{T}}\mathbf{J}_0 + t^3\mathbf{J}_0^{\mathrm{T}}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0)\right\}\boldsymbol{\varepsilon}_0. \tag{6}$$

Eqs. (5–6) involve products of $\mathbf{J}_0$-dependent terms and $\boldsymbol{\varepsilon}_0$-dependent terms. We assume that $\mathbf{J}_0$ and $\boldsymbol{\varepsilon}_0$ are statistically independent. Thus, when we estimate the RMSE of LC and LI, we can calculate separately averages of $\mathbf{J}_0$-dependent terms, and averages of $\boldsymbol{\varepsilon}_0$-dependent terms, and multiply them. The first averages are taken over the attractor's invariant distribution, which is equivalent to taking an average over a very long trajectory. The second averages are taken over the attractor's invariant distribution (because $\boldsymbol{\varepsilon}_0$ depends on $\mathbf{x}_0$) and over possible realizations of the catalog (because $\boldsymbol{\varepsilon}_0$ depends on $\mathbf{LC}_0$). Both could be estimated using the analog database called the "catalog".

Finally, we define the following error-corrected AF techniques:

$$\mathbf{GDC}_t = \mathbf{LI}_t - \left[t < \mathbf{J}_0 > + \frac{t^2}{2} < \mathbf{J}_0^2 >\right]\boldsymbol{\varepsilon}_0\,, \tag{7}$$

where GDC stands for "global dynamics correction", and

$$\mathbf{LDC}_t = \mathbf{LI}_t - \left[t\mathbf{J}_0 + \frac{t^2}{2}(\mathbf{J}_0^2 + \dot{\mathbf{J}}_0)\right]\boldsymbol{\varepsilon}_0\,, \tag{8}$$

where LDC stands for "local dynamics correction". The means (symbol $<>$) are taken over the invariant distribution of the attractor. GDC only needs average information from the Jacobian matrix, which can be inferred offline. LDC needs local information on the dynamics, which needs to be inferred online. Both global and local information about the Jacobian could be estimated using the catalog.

The quality of the estimation of $\mathbf{J}_0$-dependent quantities from the catalog depends on the dimension of the attractor and the size and quality of the catalog. However, this problematic is not studied further here and in our numerical experiments we use directly analytical expressions of the Jacobian at $\mathbf{x}_0$.

III. EVALUATION

We use the 3-dimensional Lorenz Model [8] in its standard non-dimensional form, with the standard set of parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$. Time integration is done through a 4th-order Runge-Kutta finite difference numerical scheme with a time step of 0.01 (non-dimensional units). For this model the Jacobian at any point $\mathbf{x}_0$ is:

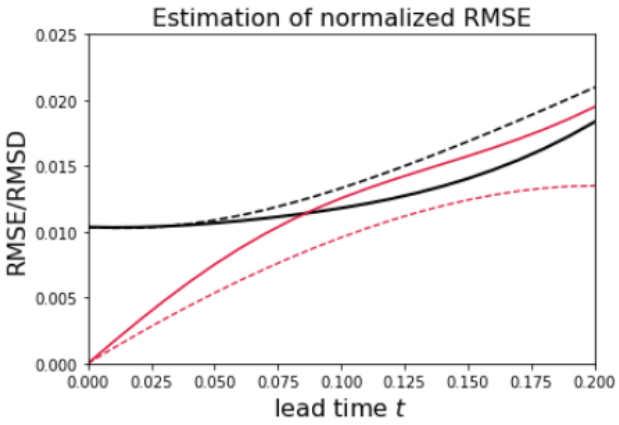


Fig. 2. RMSE associated with locally constant and locally incremental analog forecasting, normalized by the root mean squared distance between two points randomly selected on the attractor. Black, full: LC, empirical. Black, dashed: LC, predicted using Eq. (5). Red, full: LI, empirical. Red, dashed: LI, predicted using Eq. (6).

$$\mathbf{J}_0 = \begin{pmatrix} -\sigma & \sigma & 0 \\ \rho - z_0 & -1 & -x_0 \\ y_0 & x_0 & -\beta \end{pmatrix} , \qquad (9)$$

where $\mathbf{x}_0 = (x_0, y_0, z_0)$. We build 100 independent catalogs of 100000 analogs each. For each catalog we draw 1000 initial vectors $\mathbf{x}_0$ from the invariant distribution of the attractor. For each of these points we apply AF with our four different methods (LC, LI, GDC and LDC) at 50 different lead times from $t = 0.01$ to $t = 0.5$.

Each analog forecast goes through the following initial steps:

1) select the 40 analogs of $\mathbf{x}_0$ with the lowest Euclidian distance to $\mathbf{x}_0$,
2) if two or more selected analog follow each other in time by one integration time step, make a group of all these analogs and keep only the one with the lowest Euclidian distance to $\mathbf{x}_0$,
3) use Gaussian kernels for the weights $\omega_k \propto \exp\left(-0.5\|\mathbf{x}_0 - \mathbf{a}_0^k\|^2/\lambda^2\right)$ with λ set to the median of $\|\mathbf{x}_0 - \mathbf{a}_0^k\|$ inside the small set of analogs used for the forecast.

For GDC, the averages of $\mathbf{J}_0$-dependent terms are estimated offline, using Eq. (9) and within a random trajectory of 300000 points on the attractor. For LDC, $\mathbf{J}_0$-dependent terms are computed online, using Eq. (9). As mentioned earlier, all those terms could be estimated without knowledge of the model equations, provided that the catalog is large and precise enough.

Empirical RMSE associated with LC and LI are shown in the full lines (black and red) of figure 2, with estimations from Eqs. (5–6) in dashed lines. The averages of $\mathbf{J}_0$-dependent terms are estimated on a trajectory of 300000 points on the attractor. The averages of ε_0-dependent terms are estimated over the 100*100000 different analog forecasts, which span 100 different catalogs and 10 million different points $\mathbf{x}_0$ inside the attractor. Estimations based on Eqs. (5-6) are perfect for $t \to 0$, and as t grows several neglected terms influence the validity of our estimations. In particular, we neglected higher-order terms in the t-expansion of Eqs. (5-6) and non-linear terms in ε_0. Our interpretation for the overestimation of the RMSE of LC, and the underestimation of the RMSE of LI, is that the terms neglected in Eqs. (5-6) are high-order non-linear terms that are, in average, better captured by LC than LI. Indeed, those neglected terms are connected with the structure of the attractor, and since LI adds an offset to the trajectories (the last term of Eq. 2) it can generate forecasts outside the attractor, which is why those neglected terms could have a negative influence on LI forecasts and a positive one on LC forecasts. It should be reminded that these results hold for the Lorenz-63 system considered here, but the approximations may behave slightly differently with other dynamical systems. However, our estimations capture correctly the behavior and order of magnitude of RMSE for all lead times considered here.

RMSE and mean absolute error (MAE) associated with our four different forecasting methods are shown in figure 3 and compared with persistence and climatological forecast. The climatological forecast value is the climatological mean, and the persistence forecast value is the last observed state $\mathbf{x}_0$. The MAE of LC (full, green) and LI (dashed, blue) are smaller than their RMSE, this is due to rare but large values of AF errors. At small lead times, LDC (dash-dotted, brown) outperforms GDC (dotted, red), which outperforms LI (dashed, blue), which outperforms LC (full, green). All forecasts outperform climatological (full, grey) and persistence (dashed, grey) forecasts. However, our different types of error corrections have a negative influence on the performances for lead times larger than 0.1 in Lorenz-63 time. The low-order approximations that we used for small lead-time correction create inconsistent forecasts at large lead-times, predicting nonphysical trajectories far away the attractor. All different AF methods have larger normalized RMSE than MAE, which is a signature of rare but large forecasting errors. All forecasting methods proposed here seem to have a

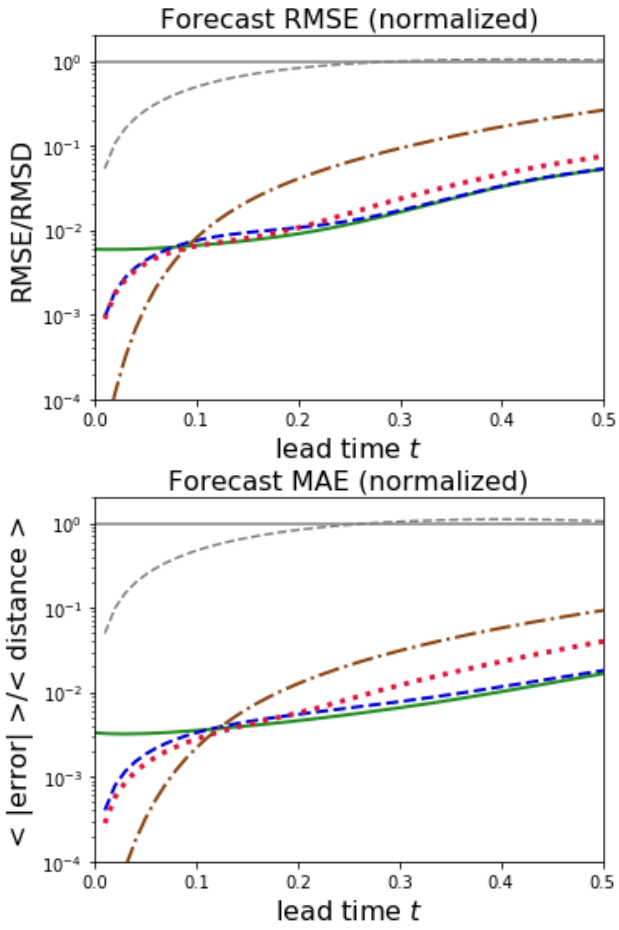


Fig. 3. Empirical RMSE and MAE (mean absolute error) associated with different analog forecasting techniques. Full, green: locally constant (LC). Dashed, blue: locally incremental (LI). Dotted, red: global dynamics correction (GDC). Dash-dotted, brown: local dynamics correction (LDC). Full, grey: climatology. Dashed, grey: persistence.

similar tendency to produce such outliers.

IV. Conclusion

AF is an empirical method, but its performances can be interpreted in the framework of dynamical systems, allowing for systematic error correction. We have expressed the leading terms of the errors of common AF techniques for short lead times. We then used those expressions on the Lorenz-63 system as a means of estimating the RMSE and for systematic error correction, yielding positive results for short lead times.

Our work is limited to AF error due to imprecision of the analogs, i.e. non-vanishing ε_0, but we wish to extend this formalism to the case the analogs and the system state follow different dynamics, creating additional error with potential bias. We should then compare the way AF variance is estimated empirically in AF applications (as in [10] and [9]) with our estimations based on the dynamics. Also, we could use more general formulae than Eq. (4) to find bounds for the AF error, possibly independent from local dynamics, to use it as error bound for AF. Finally, this formalism should be extended to the case of partial observations of a high-dimensional dynamical system, which is a more realistic situation. Indeed, in most atmospheric and oceanic applications only a few physical variables (such as temperature, pressure, humidity...) are observed, and observations are limited in time and space due to operational constraints. In such situations, AF is usually combined with Takens' time-lagged embeddings, using analogs on partial observations at current and past times [11].

Acknowledgments

Funding for the authors was provided by ANR No. 10-IEED-0006-26 and ERC grant No. 338965-A2C2.

References

[1] L. Delle Monache, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, "Probabilistic Weather Prediction with an Analog Ensemble," *Monthly Weather Review*, vol. 141, no. 1974, pp. 3498–3516, 2013.

[2] A. Ayet and P. Tandeo, "Nowcasting solar irradiance using an analog method and geostationary satellite images," *Solar Energy*, vol. 164, no. July 2017, pp. 301–315, 2018.

[3] P. Yiou and C. Déandréis, "Stochastic ensemble climate forecast with an analogue model," *Geoscientific Model Development*, vol. 12, no. 2, pp. 723–734, 2019.

[4] E. N. Lorenz, "Atmospheric Predictability as Revealed by Naturally Occurring Analogues," *Journal of the Atmospheric Sciences*, vol. 26, no. 4, pp. 636–646, 1969.

[5] C. Nicolis, "Atmospheric Analogs and Recurrence Time Statistics: Toward a Dynamical Formulation," *Journal of the Atmospheric Sciences*, vol. 55, pp. 465–475, 1998.

[6] Z. Zhao and D. Giannakis, "Analog forecasting with dynamics-adapted kernels," *Nonlinearity*, vol. 29, no. 9, pp. 2888–2939, 2016.

[7] C. Nicolis, P. Perdigao, and S. Vannitsem, "Dynamics of Prediction Errors under the Combined Effect of Initial Condition and Model Errors," *Journal of the Atmospheric Sciences*, vol. 66, pp. 766–778, 2009.

[8] E. N. Lorenz, "Deterministic nonperiodic flow," *Journal of the atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.

[9] R. Lguensat, P. Tandeo, P. Ailliot, M. Pulido, and R. Fablet, "The analog data assimilation," *Monthly Weather Review*, vol. 145, no. 10, pp. 4093–4107, 2017.

[10] P. Yiou, "AnaWEGE: a weather generator based on analogues of atmospheric circulation," *Geoscientific Model Development*, vol. 7, pp. 531–543, apr 2014.

[11] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980*, pp. 366—-381, Springer, 1981.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

TESTING RANDOM FOREST IMPUTATION FOR LAND HYDROLOGY DATA

Verena Bessenbacher[1], Lukas Gudmundsson[1], Sonia I. Seneviratne[1]

***Abstract*—Despite the importance of terrestrial water storage and fluxes in the Earth system, the inhomogeneity and fragmentation of relevant remote sensing observations hinders advancing our understanding of land-atmosphere dynamics. Imputation of missing values in remote sensing data can alleviate fragmentation issues but so far has been confined to one predictor variable. We consider algorithms that exploit the covariance structure of different variables to mutually gap-fill satellite data. In the first step described in this paper, we benchmark a Random Forest Imputation method using ERA5 reanalysis in a "perfect dataset approach". The Random Forest imputation outperforms simpler imputation methods for all fractions of missingness (measured by RMSE of "true" vs imputed value) and shows more realistic estimated soil moisture patterns over the Alpine Region in summer 2018. We argue that and similar algorithms can be successfully applied for mutually gapfilling land hydrology remote sensing data from diverse sources.**

I. MOTIVATION

Land-climate dynamics play an essential role in the climate system, and in particular affect the frequency and intensity of regional climate extremes such as heatwaves, droughts and heavy precipitation [1], [2], [3], [4]. Dependable observations of terrestrial water resources are thus essential for investigating land-climate dynamics in the context of climate change and associated changes in climate extremes, especially given the large spread of Earth system models with respect to the underlying processes and associated projections [5], [6].

In the past decades the rapid evolution of satellite-borne Earth system observations has allowed for a bird's-eye view on variations in terrestrial water storage and fluxes at continental to global scales [7], [8], [9]. New satellite-based products are available for land water variables such total water storage [10], [11], soil moisture [12], evapotranspiration [13], [14], [15] or land surface temperatures [16]. However, remotely-sensed observations suffer from missing data and inhomogeneities, limiting their wide spread use. Furthermore, efforts to consolidate observations from different platforms have usually focused on providing data for individual variables at a time (e.g. soil moisture [12] or precipitation [17]). Hence, although the past decades have seen massive advances in generating Earth System observations, the typical approach to only focus on one variable at a time and the inherent missingness of the datasets has led to a fragmentation of the observational record.

Corresponding author: V Bessenbacher, verena.bessenbacher@env.ethz.ch [1]Institute for Atmospheric and Climate Sciences, ETH Zurich

A. Gapfilling Methods in Literature

Recent research has shown that data-driven approaches can be used to tackle the challenge of fragmented observations in environmental and climate research. Several investigations have employed machine learning to upscale point-scale in-situ evapotranspiration [18], [19] and streamflow [20], [21] observations to regular grids covering continental to global regions. Other work has employed first-order process approximations, Bayesian inference and atmospheric reanalysis to reconstruct short remote sensing records of total water storage over the past century [22]. Finally, some approaches used advanced space-time interpolation techniques to fill in missing values in remotely-sensed time series of vegetation activity [23], [24]. A common feature of the aforementioned efforts is that they focus on one variable at a time and therefore cannot guarantee consistency in multivariate estimates. In other words, correlations between variables from the resulting data products may not reflect the true correlation structure. Furthermore, all of the above-mentioned approaches rely on gap-free predictor variables to infer missing values.

The proposed research aims at overcoming these limitations by using the mutual information content of Earth system observations to create reliable estimates of missing values in multivariable data sets.

The process of gap-filling multivariate datasets is often referred to as statistical imputation [25] or matrix completion [26]. Apart from trivial methods, like replacing missing values with the mean of the non-missing values, these imputation methods are based on identifying the dependence structure of multivariate data sets. Knowledge of this dependence structure is then used to infer missing values. The main difference to classical spatial interpolation methods lies in this exploitation of the covariance of different variables.

The methodological literature offers a large variety of algorithms for this task. One large group of algorithms is based on regression modelling including linear models [25] or Random Forest based techniques [27].

Another large group of algorithms is based on the idea that the (unknown) complete data matrix has a suitable low-rank representation. In climate science this is equivalent to the widespread notion that spatiotemporal fields can be sufficiently described by a limited number of principal components [28]. Algorithms in this group include different flavours of singular value decomposition (SVD) imputation [29], [30], [26] or neural network autoencoders [31], [32], [33], which are a non-linear generalization of an SVD.

B. Objective

In this paper, we describe a first step taken towards merging the fragmented observational record: we benchmark the candidate Random Forest based imputation method MissForest [27] to evaluate its performance for imputing missing values in large land hydrology datasets from an atmospheric reanalysis and to benchmark its performance against trivial methods.

II. Methods

For the purpose of methods development and benchmarking we fall back on atmospheric reanalysis, which provide gap-free estimates of essential climate variables. To benchmark the gap-filling method, we employ a ”perfect dataset approach”, where we assume the reanalysis dataset from the ERA5-Project [34] to be the ”true” state of the land-climate interactions. To evaluate the fidelity of the considered methods missing values are artificially introduced and subsequently imputed.

A. Dataset

We use reanalysis data from the ERA5-Project [34], a gap-free and coherent multivariate ”view” of the past atmosphere and land interaction. We confine our analysis to the Alpine Region (46°N to 52°N and 5°E to 8°E) from 2000 until 2018 in monthly time resolution. We include all land surface variables directly related to the terrestrial water cycle (precipitation, runoff, evapotranspiration and soil moisture in four soil layers), 2-meter temperature and leaf area index. Precipitation is log-transformed and all variables are normalized to span values $[0, 1]$. The so obtained dataset is subsequently called X_{orig}.

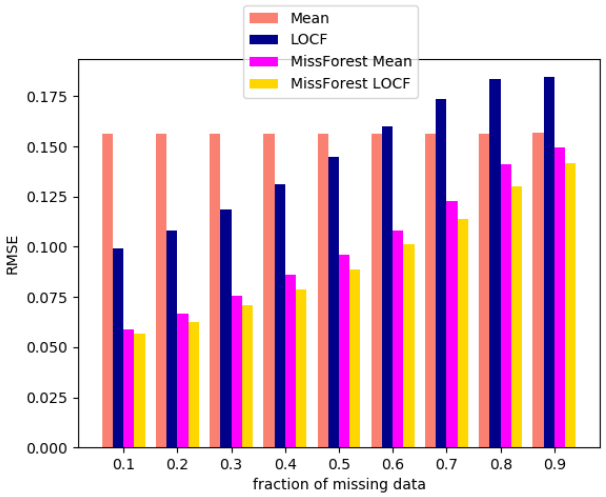


Fig. 1. RMSE per imputation method for different fractions of missing data.

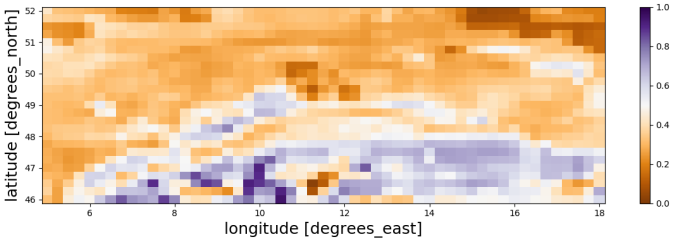


Fig. 2. Normalised ERA5 soil moisture of the uppermost soil layer over the Alpine Region in August 2018.

B. Imputation Method

The MissForest method [27] adapts the Random Forest regression [35] for cases where values are missing in both the predictand and the predictor. Since the Random Forest regression cannot deal with missingness, the missing values are first imputed with a simple imputation method and the result is iteratively improved by running a Random Forest regression over the imputed dataset and subsequently replacing the missing values from the learned regression model. This is done until the difference between the imputed data from two iterations is small. For details the reader is referred to [27]. Here, we employ an adapted convergence

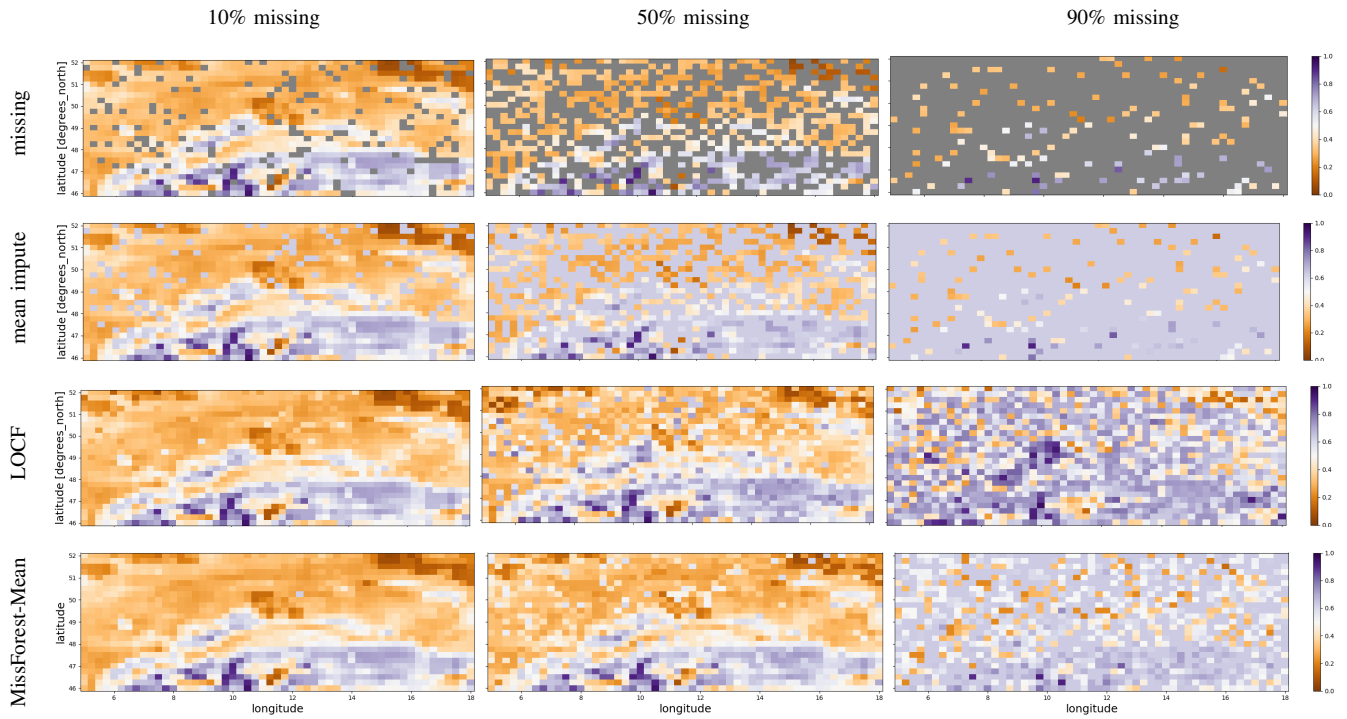


Fig. 3. Setup and results of the imputation methods: columns correspond to different fractions of missingness (10%, 50%, 90%). . The first row shows the original date with missing values (gray). The subsequent rows show different imputation methods (mean imputation, LOCF, MissForest-Mean).

criterion, i.e. stop the iteration if the difference between the current and the last iteration is is below 10^{-5}. Since the algorithm can only digest a data vector per variable, the dataset is flattened along time, latitude and longitude. Therefore all neighborhood relations and temporal information is lost. As hyperparameters for the MissForest regression, we found 100 trees with a maximum node depth of 100 to lead to the best results. We benchmark the MissForest regression with two simple imputation methods: (1) inserting the variable mean over the whole domain and (mean impute) (2) carry the last available observation forward until a new observation is available (last observation carried forward or LOCF). The second is conceptually the same as a persistence forecast.

C. Benchmarking Setup

Our benchmarking setup consits of three steps: In the first step, we delete a fraction of the data in the otherwise gap-free ERA5 dataset randomly. Subsequently, we fill the "lost" values with imputed values from the two simple imputation methods and the MissForest method and obtain X_{imp}. Finally, we evaluate the accuracy of the imputation method by calculating the root mean square error (RMSE) of the imputed values in X_{imp} and X_{orig}.

III. EVALUATION

Figure 1 shows the RMSE between X_{orig} and X_{imp} for all considered methods aggregated over all normalised variables, timesteps and locations for different fractions of missing data. The RMSE of the mean imputation stays relatively constant, since by definition it averages over all missing values. It converges to the standard deviation of the dataset. For the LOCF method, the RMSE steadily increases with increasing fraction of missing data. The larger the gaps are in between the observed points, the more variability is missed, therefore LOCF works quite well for small fractions of missing data, but deteriorates for larger fractions. The MissForest algorithm, once initialised with LOCF estimates (MissForest-LOCF) and once with mean imputation (MissForest-Mean) outperforms both simple imputation methods for all fractions of missingness. Note the impact of the inital gap-filling: MissForest-Mean outperforms MissForest-LOCF for all fractions of missingness. Initially gapfilling with the variable mean hence gives estimates closer to the "true" values and the MissForest algorithm converges more

quickly to more realistic estimates. This highlights the importance of the initial imputation and the definition of the convergence criterium.

Figure 2 shows a sample time slab of the ERA5 data. We arbitrarily chose the soil moisture values of the uppermost soil layer for August 2018 to show qualitatively the impact of different fractions of missing data and imputation methods in Figure 3.

For the case where 10% of the data is missing, LOCF and MissForest imputation perform similarly well and can give reasonable values. However, their results quickly deteriorate for larger amounts of missingness. The mean imputation, by definition, is not able to capture spatial and temporal variations in the data. LOCF gives noisy estimates for large fractions of missingness. Both methods have a severe impact on the statistical properties of the dataset: They shrink the variance significantly and destroy the multivariate covariance structure.

MissForest is able to keep the patterns in the dataset (mainly wetter soil in the Alpine Region) well for 10% and 50% of missing data. For the case where 90% of the data is missing, in the MissForest case lower imputed soil moisture values still appear more often outside the Alpine Region, i.e. the algorithm still catches some of the observed pattern, albeit with a lot of noise. This is likely stemming from the initial mean gapfill. We assume this happens especially at grid-points where little variables are observed at all, however this still needs to be investigated further.

Overall, the results indicate that complex imputation algorithms such as MissForest have the potential to perform well for variables related to the terrestrial water cycle. We argue that constitutes a first step towards the application of high level imputation algorithms to Earth system observations. Spatial neighborhoods and temporal autocorrelation are not yet exploited in the presented approach, but we plan to incorporate this in a next step, e.g. via using neighboring pixels as predictors in the MissForest regression, exploring the added value neural networks with suitable recurrent and convolutional features or benchmarking the presented method with gaussian process based methods.

ACKNOWLEDGMENTS

This work was supported by ETH Research Grant ETH-08 19-1 and the ESA's Climate Change Initiative for Soil Moisture 4000126684/19/I-NB.

REFERENCES

[1] S. I. Seneviratne, T. Corti, E. L. Davin, M. Hirschi, E. B. Jaeger, I. Lehner, B. Orlowsky, and A. J. Teuling, "Investigating soil moisture–climate interactions in a changing climate: A review," *Earth-Science Reviews*, vol. 99, pp. 125–161, May 2010.

[2] P. Greve, B. Orlowsky, B. Mueller, J. Sheffield, M. Reichstein, and S. I. Seneviratne, "Global assessment of trends in wetting and drying over land," *Nature Geoscience*, vol. 7, pp. 716–721, Oct. 2014.

[3] B. P. Guillod, B. Orlowsky, D. G. Miralles, A. J. Teuling, and S. I. Seneviratne, "Reconciling spatial and temporal soil moisture effects on afternoon rainfall," *Nature Communications*, vol. 6, p. 6443, Mar. 2015.

[4] Seneviratne Sonia I., Wartenburger Richard, Guillod Benoit P., Hirsch Annette L., Vogel Martha M., Brovkin Victor, van Vuuren Detlef P., Schaller Nathalie, Boysen Lena, Calvin Katherine V., Doelman Jonathan, Greve Peter, Havlik Petr, Humpenöder Florian, Krisztin Tamas, Mitchell Daniel, Popp Alexander, Riahi Keywan, Rogelj Joeri, Schleussner Carl-Friedrich, Sillmann Jana, and Stehfest Elke, "Climate extremes, land–climate feedbacks and land-use forcing at 1.5C," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 376, p. 20160450, May 2018.

[5] V. Humphrey, J. Zscheischler, P. Ciais, L. Gudmundsson, S. Sitch, and S. I. Seneviratne, "Sensitivity of atmospheric CO 2 growth rate to observed changes in terrestrial water storage," *Nature*, vol. 560, p. 628, Aug. 2018.

[6] M. M. Vogel, J. Zscheischler, and S. I. Seneviratne, "Varying soil moisture–atmosphere feedbacks explain divergent temperature extremes and precipitation projections in central Europe," *Earth System Dynamics*, vol. 9, pp. 1107–1125, Aug. 2018.

[7] G. Balsamo, A. Agusti-Panareda, C. Albergel, G. Arduini, A. Beljaars, J. Bidlot, E. Blyth, N. Bousserez, S. Boussetta, A. Brown, R. Buizza, C. Buontempo, F. Chevallier, M. Choulga, H. Cloke, M. F. Cronin, M. Dahoui, P. De Rosnay, P. A. Dirmeyer, M. Drusch, E. Dutra, M. B. Ek, P. Gentine, H. Hewitt, S. P. Keeley, Y. Kerr, S. Kumar, C. Lupu, J.-F. Mahfouf, J. McNorton, S. Mecklenburg, K. Mogensen, J. Muñoz-Sabater, R. Orth, F. Rabier, R. Reichle, B. Ruston, F. Pappenberger, I. Sandu, S. I. Seneviratne, S. Tietsche, I. F. Trigo, R. Uijlenhoet, N. Wedi, R. I. Woolway, and X. Zeng, "Satellite and In Situ Observations for Advancing Global Earth Surface Modelling: A Review," *Remote Sensing*, vol. 10, p. 2038, Dec. 2018.

[8] M. F. McCabe, M. Rodell, D. E. Alsdorf, D. G. Miralles, R. Uijlenhoet, W. Wagner, A. Lucieer, R. Houborg, N. E. C. Verhoest, T. E. Franz, J. Shi, H. Gao, and E. F. Wood, "The future of Earth observation in hydrology," *Hydrology and Earth System Sciences*, vol. 21, pp. 3879–3914, July 2017.

[9] D. P. Lettenmaier, D. Alsdorf, J. Dozier, G. J. Huffman, M. Pan, and E. F. Wood, "Inroads of remote sensing into hydrologic science during the WRR era," *Water Resources Research*, vol. 51, no. 9, pp. 7309–7342, 2015.

[10] M. Rodell, J. S. Famiglietti, D. N. Wiese, J. T. Reager, H. K. Beaudoing, F. W. Landerer, and M.-H. Lo, "Emerging trends in global freshwater availability," *Nature*, vol. 557, p. 651, May 2018.

[11] V. Humphrey, L. Gudmundsson, and S. I. Seneviratne, "Assessing Global Water Storage Variability from GRACE: Trends, Seasonal Cycle, Subseasonal Anomalies and Extremes," *Surveys in Geophysics*, vol. 37, pp. 357–395, Mar. 2016.

[12] W. Dorigo, W. Wagner, C. Albergel, F. Albrecht, G. Balsamo, L. Brocca, D. Chung, M. Ertl, M. Forkel, A. Gruber, E. Haas, P. D. Hamer, M. Hirschi, J. Ikonen, R. de Jeu, R. Kidd, W. Lahoz, Y. Y. Liu, D. Miralles, T. Mistelbauer, N. Nicolai-Shaw,

R. Parinussa, C. Pratola, C. Reimer, R. van der Schalie, S. I. Seneviratne, T. Smolander, and P. Lecomte, "ESA CCI Soil Moisture for improved Earth system understanding: State-of-the art and future directions," *Remote Sensing of Environment*, vol. 203, pp. 185–215, Dec. 2017.

[13] D. G. Miralles, C. Jiménez, M. Jung, D. Michel, A. Ershadi, M. F. McCabe, M. Hirschi, B. Martens, A. J. Dolman, J. B. Fisher, Q. Mu, S. I. Seneviratne, E. F. Wood, and D. Fernández-Prieto, "The WACMOS-ET project – Part 2: Evaluation of global terrestrial evaporation data sets," *Hydrology and Earth System Sciences*, vol. 20, pp. 823–842, Feb. 2016.

[14] B. Martens, D. G. Miralles, H. Lievens, R. van der Schalie, R. A. M. de Jeu, D. Fernández-Prieto, H. E. Beck, W. A. Dorigo, and N. E. C. Verhoest, "GLEAM v3: satellite-based land evaporation and root-zone soil moisture," *Geosci. Model Dev.*, vol. 10, pp. 1903–1925, May 2017.

[15] B. Mueller, M. Hirschi, C. Jimenez, P. Ciais, P. A. Dirmeyer, A. J. Dolman, J. B. Fisher, M. Jung, F. Ludwig, F. Maignan, D. G. Miralles, M. F. McCabe, M. Reichstein, J. Sheffield, K. Wang, E. F. Wood, Y. Zhang, and S. I. Seneviratne, "Benchmark products for land evapotranspiration: LandFlux-EVAL multi-data set synthesis," *Hydrology and Earth System Sciences*, vol. 17, pp. 3707–3720, Oct. 2013.

[16] I. F. Trigo, S. Boussetta, P. Viterbo, G. Balsamo, A. Beljaars, and I. Sandu, "Comparison of model land skin temperature with remotely sensed estimates and assessment of surface-atmosphere coupling: MODEL SKIN TEMPERATURE AND SATELLITE LST," *Journal of Geophysical Research: Atmospheres*, vol. 120, pp. 12,096–12,111, Dec. 2015.

[17] Q. Sun, C. Miao, Q. Duan, H. Ashouri, S. Sorooshian, and K.-L. Hsu, "A Review of Global Precipitation Data Sets: Data Sources, Estimation, and Intercomparisons," *Reviews of Geophysics*, vol. 56, pp. 79–107, Mar. 2018.

[18] M. Jung, M. Reichstein, and A. Bondeau, "Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model," p. 13, 2009.

[19] M. Jung, M. Reichstein, H. A. Margolis, A. Cescatti, A. D. Richardson, M. A. Arain, A. Arneth, C. Bernhofer, D. Bonal, J. Chen, D. Gianelle, N. Gobron, G. Kiely, W. Kutsch, G. Lasslop, B. E. Law, A. Lindroth, L. Merbold, L. Montagnani, E. J. Moors, D. Papale, M. Sottocornola, F. Vaccari, and C. Williams, "Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations," *Journal of Geophysical Research*, vol. 116, p. G00J07, Sept. 2011.

[20] G. Ghiggi, V. Humphrey, S. I. Seneviratne, and L. Gudmundsson, "GRUN: An observations-based global gridded runoff dataset from 1902 to 2014," *Earth System Science Data Discussions*, pp. 1–32, Mar. 2019.

[21] L. Gudmundsson and S. I. Seneviratne, "Observation-based gridded runoff estimates for Europe (E-RUNversion 1.1)," *Earth System Science Data*, vol. 8, pp. 279–295, July 2016.

[22] V. Humphrey and L. Gudmundsson, "GRACE-REC: a reconstruction of climate-driven water storage changes over the last century," *Earth System Science Data Discussions*, pp. 1–41, Feb. 2019.

[23] F. Gerber, R. d. Jong, M. E. Schaepman, G. Schaepman-Strub, and R. Furrer, "Predicting Missing Values in Spatio-Temporal Remote Sensing Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, pp. 2841–2853, May 2018.

[24] J. von Buttlar, J. Zscheischler, and M. D. Mahecha, "An extended approach for spatiotemporal gapfilling: dealing with large and systematic gaps in geoscientific datasets," *Nonlinear Processes in Geophysics*, vol. 21, pp. 203–215, 2014.

[25] S. v. Buuren, *Flexible Imputation of Missing Data, Second Edition*. Boca Raton: Chapman and Hall/CRC, 2 edition ed., July 2018.

[26] M. A. Davenport and J. Romberg, "An Overview of Low-Rank Matrix Recovery From Incomplete Observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, pp. 608–622, June 2016.

[27] D. J. Stekhoven and P. Bühlmann, "MissForest–nonparametric missing value imputation for mixed-type data," *Bioinformatics (Oxford, England)*, vol. 28, pp. 112–118, Jan. 2012.

[28] D. S. Wilks, *Statistical Methods in the Atmospheric Sciences, Volume 100*. Amsterdam ; Boston: Academic Press, 3 edition ed., June 2011.

[29] R. Mazumder, T. Hastie, and R. Tibshirani, "Spectral Regularization Algorithms for Learning Large Incomplete Matrices," p. 36.

[30] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman, "Missing value estimation methods for DNA microarrays," *Bioinformatics*, vol. 17, pp. 520–525, June 2001.

[31] L. Gondara and K. Wang, "MIDA: Multiple Imputation Using Denoising Autoencoders," in *Advances in Knowledge Discovery and Data Mining* (D. Phung, V. S. Tseng, G. I. Webb, B. Ho, M. Ganji, and L. Rashidi, eds.), Lecture Notes in Computer Science, pp. 260–272, Springer International Publishing, 2018.

[32] A. Nazabal, P. M. Olmos, Z. Ghahramani, and I. Valera, "Handling Incomplete Heterogeneous Data using VAEs," *arXiv:1807.03653 [cs, stat]*, July 2018. arXiv: 1807.03653.

[33] C. K. I. Williams, C. Nash, and A. Nazábal, "Autoencoders and Probabilistic Inference with Missing Data: An Exact Solution for The Factor Analysis Case," *arXiv:1801.03851 [cs, stat]*, Jan. 2018. arXiv: 1801.03851.

[34] A. Guillory, "ERA5." https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5, Nov. 2017.

[35] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, pp. 5–32, Oct. 2001.

DATA-DRIVEN VS. PHYSICALLY-BASED STREAMFLOW PREDICTION MODELS

Martin Gauch,[1] Juliane Mai,[2] Shervan Gharari,[3] Jimmy Lin[1]

***Abstract*—Climate change leads to more frequent and severe floods and droughts. Precise water flow forecasts for rivers and streams help mitigate damage and are direly needed. We evaluate physically-based and data-driven models on the task of streamflow prediction in the Lake Erie region: Physically-based models capture simplified representations of the physical processes that underlie streamflow, while purely data-driven models encode no such knowledge explicitly. Experiments show that data-driven approaches can provide more accurate predictions than a physically-based model, suggesting potential in hybrid approaches that combine hydrological understanding with high prediction accuracy.**

I. INTRODUCTION

Accurate prediction of *streamflow*—the amount of water that flows through a river at a certain time—plays a vital role in managing extreme floods and droughts. Due to climate change, such disasters have become increasingly frequent and impact the lives of people around the world. Hydrology has a long history of developing streamflow prediction models: for different watersheds, based on different datasets, and based on different evaluation criteria. However, it often remains unclear which model is best under which conditions. The ongoing *Great Lakes Runoff Inter-comparison Project for Lake Erie* (*GRIP-E*) compares hydrologic models in the largest Canadian effort yet to overcome these issues [1].

GRIP-E mostly considers *physically-based* hydrologic and land-surface models; by this, we mean models that replicate simplified representations of the underlying physical processes to predict streamflow. We, however, believe that *data-driven*, machine-learning models can in fact provide meaningful contributions towards understanding streamflow, too. Although researchers have been hesitant to adopt machine-learning models that are often black-box predictors, our work shows that data-driven models can aid hydrologists' advancement in explaining the physical processes that underlie streamflow. Purely data-driven models reveal how much streamflow information is extractable from the datasets that are used in physically-based models.

Our study compares a physically-based model with data-driven linear and tree-based models in their ability to accurately predict the water flow of a stream at a particular gauging station, given meteorological data of the surrounding area. We use a five-year meteorological dataset of the Lake Erie watershed to predict the streamflow at gauging stations in sub-watersheds around the lake. The purely data-driven approaches predict streamflow more accurately than the physically-based model. To us, this is good news, as it shows that there is sufficient signal in existing data to make accurate predictions, and points to potential hybrid models that are both useful for advancing hydrological understanding and making high-quality forecasts.

II. DATA AND METHODS

As streamflow ground truth, we use daily measurements at 46 gauging stations in the Lake Erie region from 2010 to 2014. These stations divide the watershed into sub-watersheds, each comprised of the area in which all water flows towards the gauging station. Figure 1 shows a map of the gauging stations used for GRIP-E and their corresponding sub-watersheds.

Both physically-based and data-driven models use gridded meteorological *forcing data* as input. In hydrology, forcing data are time-series datasets that are required to run, or *force*, the model. Many physically-based models additionally use geophysical inputs such as soil and elevation maps that change little over time. The forcing data used in this study include hourly meteorological variables of temperature, precipitation, pressure, wind speed, specific humidity, and short- and longwave radiation with a spatial resolution of around 15 km spanning five years (2010 to 2014). Table I

Corresponding author: Martin Gauch, mgauch@uwaterloo.ca
[1]David R. Cheriton School of Computer Science, University of Waterloo, ON, Canada; [2]Civil and Environmental Engineering, University of Waterloo, ON, Canada; [3]Coldwater Lab, University of Saskatchewan, AB, Canada

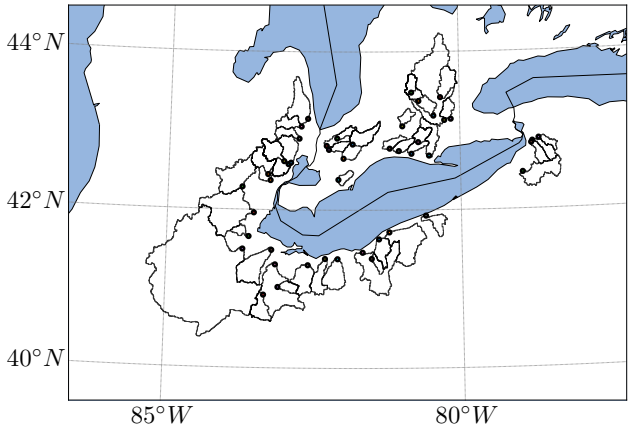


Fig. 1: Geographical outlines of the 46 sub-watersheds in our analysis, each draining towards a gauging station (black dots).

Variable	Explanation	Level	Unit
PR0	Quantity of precipitation	surface	m
TT	Air temperature	40 m	°C
FB	Downward solar flux	40 m	$\mathrm{W\,m^{-2}}$
FI	Surface incoming infrared flux	40 m	$\mathrm{W\,m^{-2}}$
P0	Surface pressure	surface	mbar
HU	Specific humidity	40 m	$\mathrm{kg\,kg^{-1}}$
UVC	Wind speed	40 m	kn

TABLE I: Meteorological forcing variables used in this study. Each variable covers the entire Lake Erie watershed at a resolution of around 15 km for the years 2010 to 2014 at an hourly resolution. The variables are available at the different vertical levels indicated.

summarizes more details about the variables. Figure 2 visualizes a snapshot of the temperature forcing data.

Formally, we aim to solve a regression problem. We predict the streamflow y_t^S at station S at time t, given the history of meteorological forcings $\mathbf{X}_{[1,t]}^S$:

$$\begin{aligned}\mathbf{X}_{[1,t]}^S &= \left[\mathbf{x}_1^1, \ldots, \mathbf{x}_t^1, \ldots, \mathbf{x}_1^{p_S}, \ldots, \mathbf{x}_t^{p_S}\right] \\ &= \left[\mathbf{x}_{[1,t]}^1, \ldots, \mathbf{x}_{[1,t]}^{p_S}\right] \in \mathbb{R}^{7\times(t\cdot p_S)} \end{aligned} \quad (1)$$

The superscripts $1, \ldots, p_S$ identify the p_S grid cells in the sub-watershed of station S, the subscripts $1, \ldots, t$ represent time steps, and each $\mathbf{x}_i^c$ is a vector of seven forcing variables (Table I). Since we use machine-learning models that operate on vectors rather than matrices, we introduce the following vectorization:

$$\mathbf{x}_{[1,t]}^S = \mathrm{vec}(\mathbf{X}_{[1,t]}^S) = \begin{bmatrix}\mathbf{x}_1^1 \\ \vdots \\ \mathbf{x}_t^{p_S}\end{bmatrix} \in \mathbb{R}^{7\cdot(t\cdot p_S)} \quad (2)$$

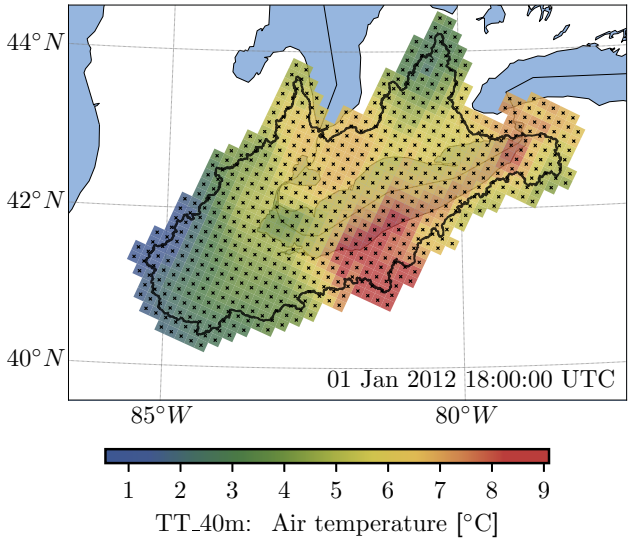


Fig. 2: Gridded forcing data for the Lake Erie watershed (black outline). As an example, the temperature of Jan 1, 2012 6pm (UTC) is depicted (colored tiles). Each tile is about $15 \times 15\,\mathrm{km}^2$ in size.

To obtain predictions, we train the parameters ω of a model f to output an estimate $\hat{y}_t^S = f(\mathbf{x}_{[1,t]}^S; \omega)$.

A. *Physically-Based Hydrologic Models*

Physically-based hydrologic models capture a simplified simulation of the underlying physical processes that result in streamflow and other hydrologic fluxes. These models often use various sources of data such as land cover maps, soil maps, or digital elevation models as their setup basis. Similar to data-driven models, the parameters ω of a hydrologic model are usually trained, or calibrated, against the observed streamflow time series at one or multiple gauges.

For this study, we make use of the *Variable Infiltration Capacity model based on Grouped Response Units* (*VIC-GRU*). This model originates from the *VIC* model [2], [3], which is a large-scale, semi-distributed hydrologic model that simulates each grid cell independently. VIC-GRU is a variant of VIC that processes spatial extents with similar characteristics as so-called *grouped response units* (GRUs). This grouping makes the simulation computationally more efficient, which allows us to use input data at a higher resolution.

We train one VIC-GRU model on all gauging stations, as the model already incorporates varying spatial characteristics through geospatial input information such as soil maps. As physically-based models approximate natural system states and fluxes, they need to attain realistic initial model conditions for the training

period before generating accurate output. To evaluate the model's goodness-of-fit, we discard the first year of model simulations (2010) as the so-called *warm-up period*, and only use the NSE coefficient for the training period 2011 to 2012. We use the parameter set that generates the best NSE values in the training period to predict the test period 2013 to 2014.

B. Machine-Learning Models

We use a cross-validated random search to find suitable parameters for each model. To reduce dimensionality, we only use the meteorological forcing data of the p_S grid cells in sub-watershed S. As the models we use neither naturally incorporate temporally-distributed nor spatially-distributed input, we flatten the data to a fixed history window of eight days and train one model per gauging station. We further aggregate the hourly forcing data into daily values to match the target streamflow data resolution. This aggregation uses the minimum and maximum temperature per day and total precipitation on that day. Preliminary experiments show that the remaining forcing variables do not improve prediction accuracy (results not shown). We therefore exclude them from the inputs for the data-driven models.

As a baseline, we train a linear ridge regression model to predict streamflow. Linear regression finds a parameter vector $\omega \in \mathbb{R}^{7\cdot(8\cdot p_S)}$ that minimizes the sum of squared residual differences between target values y_t^S and predicted values $\hat{y}_t^S = \mathbf{x}_{[t-7,t]}^S \omega$. As our problem involves high-dimensional data, ridge regression is an appropriate choice because it includes a weighted regularization term to reduce overfitting.

We also employ XGBoost as a more sophisticated approach that trains gradient-boosted regression trees (GBRTs) [4]. GBRTs iteratively train K regression trees f_k and generate an overall prediction $\hat{y}_t^S$ as the sum of their outputs. Additionally, GBRTs provide regularization parameters such as a maximum tree depth to control overfitting. For more details on the objective function minimization, see Chen and Guestrin [4].

C. Evaluation

We split the available data into a training period from 2010 to 2012 and a test period from 2013 to 2014. Our data-driven models are trained using mean squared error (MSE). After fitting a model during the training phase, we apply it to the test period and evaluate its prediction accuracy. Following common practice in hydrology, we use the *Nash-Sutcliffe efficiency coefficient* (NSE) to evaluate the simulated streamflow $\hat{y}^S$ compared to the observed streamflow time series y^S for station S, defined as follows:

$$\begin{aligned} \text{NSE} &= 1 - \frac{\sum_{t=1}^{T} (\hat{y}_t^S - y_t^S)^2}{\sum_{t=1}^{T} (y_t^S - \bar{y}^S)^2} \\ &= 1 - \frac{\text{MSE}}{\frac{1}{T}\sum_{t=1}^{T} (y_t^S - \bar{y}^S)^2} \end{aligned} \tag{3}$$

where $\bar{y}^S$ is the mean observed streamflow at station S across all T time steps. Hence, the denominator is the variance of the streamflow observations y^S. Equation 3 shows that NSE and MSE are strongly correlated [5]. NSE values range between $-\infty$ and 1, with 1 representing perfect predictions. A score of 0 is obtained when predicting $\bar{y}^S$ at every time step.

Statistic	VIC-GRU	Ridge regression	XGBoost
p_0	-6.302	-1.677	-0.206
p_{25}	0.184	0.298	0.412
p_{50}	0.328	0.380	0.522
p_{75}	0.376	0.469	0.561
p_{100}	0.597	0.585	0.666
$p_{75} - p_{25}$	0.191	0.170	0.149

TABLE II: Minimum p_0, maximum p_{100}, quartiles p_{25} and p_{75}, median p_{50}, and interquartile range $p_{75} - p_{25}$ of the NSE distributions for the physically-based model VIC-GRU and the two data-driven models (ridge regression and XGBoost).

III. RESULTS

Table II outlines characteristics of the three models' NSE distributions across all 46 gauging stations. The third row (p_{50}) shows the median NSE values, and the other rows list the measure at different percentiles as well as the interquartile range (see table caption for details). We see that a simple ridge regression model outperforms the physically-based VIC-GRU model, and that XGBoost provides the most accurate predictions of the three examined models. The XGBoost model outperforms VIC-GRU in 40 of the 46 stations by an average NSE difference of 0.473. Not only are the XGBoost predictions more accurate overall, but they also show fewer outliers (i.e., make terribly bad predictions) and a smaller variation of NSE coefficients for different stations.

In Figure 3, we provide an overview of results for three sample gauging stations: one where all models perform relatively well, one where VIC-GRU performs

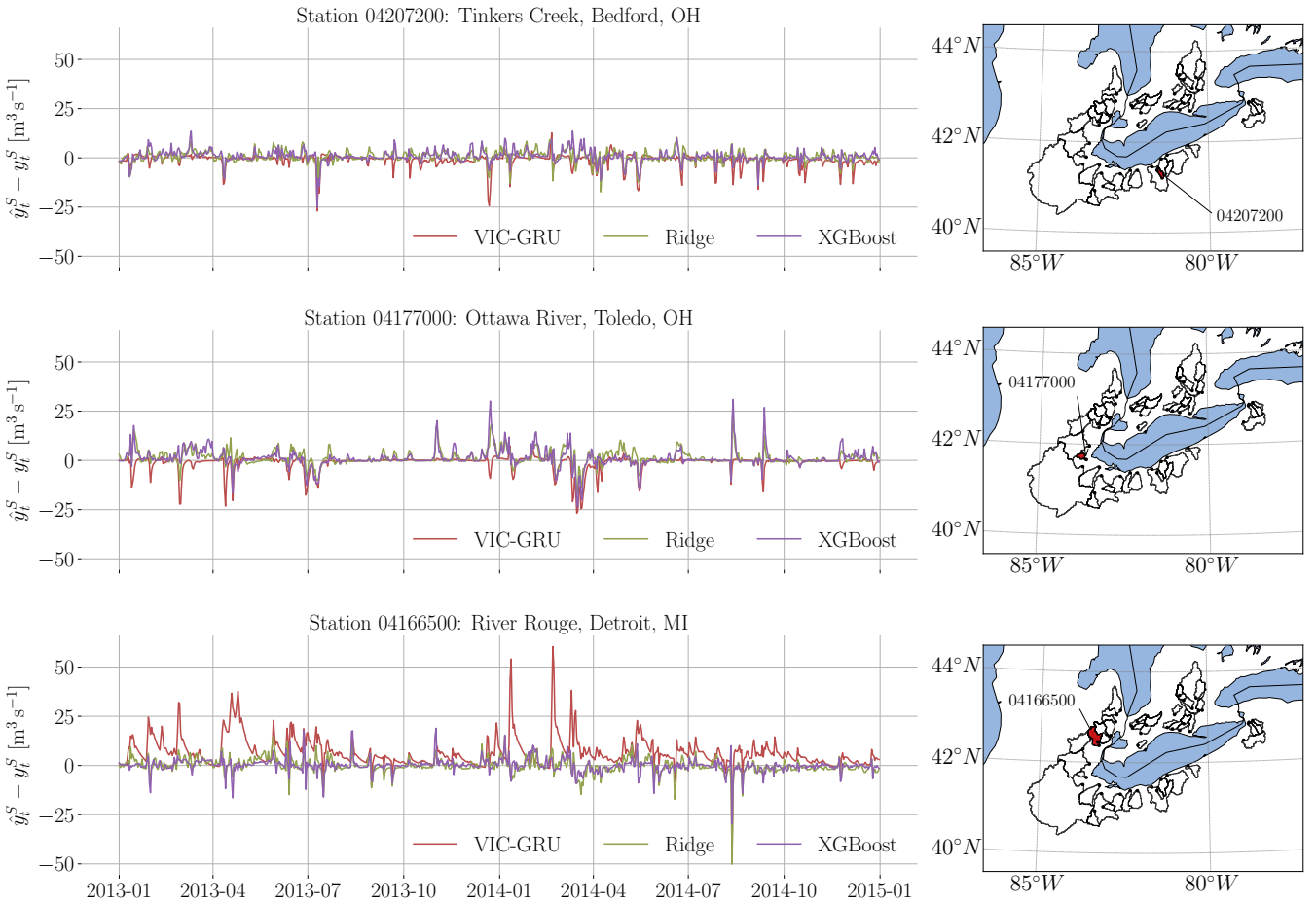


Fig. 3: Differences between actual streamflow y_t^S and predictions $\hat{y}_t^S$ for VIC-GRU (red), ridge regression (green), and XGBoost (purple) at gauging stations 04207200, 04177000, and 04166500 during the test period.

better than machine-learning models, and one where the machine-learning models outperform VIC-GRU. Each panel shows the differences between the actual and predicted streamflows at each gauging station based on the three models; the corresponding figures on the right highlight the stations' geographical locations on a map.

All three models provide rather good predictions for station 04207200 in Bedford, OH, USA (Figure 3, top), with NSE values between 0.44 (VIC-GRU) and 0.57 (XGBoost). For station 04177000 in Toledo, OH, USA (Figure 3, middle), VIC-GRU yields a better NSE than XGBoost. Largely, this seems to be due to a few overestimated streamflow spikes at the end of 2013 and around August 2014. In the third example, VIC-GRU makes very poor predictions for station 04166500 in Detroit, MI, USA (Figure 3, bottom), with an NSE score well below zero, while XGBoost and ridge regression provide far better results. VIC-GRU appears to struggle with the prediction of streamflow peaks, as it frequently incorrectly predicts high streamflows above $50\,\mathrm{m}^3\,\mathrm{s}^{-1}$ during the winter and spring months. XGBoost and ridge regression are more conservative and rarely predict streamflows above $30\,\mathrm{m}^3\,\mathrm{s}^{-1}$, likely because the station's training data contain few high-streamflow examples. This difference partly explains the more accurate predictions of the data-driven models, because the NSE calculation includes squared differences that emphasize outliers. Ridge regression often produces erratic predictions for periods of low streamflow, which explains the model's lower NSE coefficients compared to XGBoost.

It is further noteworthy that station 04166500 is located in the highly urbanized metropolitan area of Detroit. Such regions are more prone to human water regulation, which is often not included in the assumptions physically-based models make. In contrast, machine-learning techniques are able to implicitly capture water regulation policies.

IV. DISCUSSION

Our results show that, in relative terms, imprecise predictions by VIC-GRU cannot be solely explained by insufficient data, as the purely data-driven approaches outperform the physical model using only a subset of the data. In other words, it appears that the model does not yet fully exploit the available signals due to its design and the restrictive assumptions it makes. This is more pronounced in urban regions, where we envision great potential in augmenting physically-based models with machine-learning techniques.

We note, however, that the three-year time frame from 2010 to 2012 is a rather short training period for a physically-based model. As the GRIP-E project is still ongoing and awaiting an extended forcing dataset, we are unfortunately unable to train on a longer time period. Given the large differences in prediction accuracy, we however do not expect that additional data would fundamentally change our findings, especially since more data would benefit data-driven models also.

V. FUTURE WORK

Unfortunately, our data-driven models do not yet provide insights into *how* physically-based models could be improved. In future work, we therefore plan to study machine-learning models whose structures are more specifically targeted towards predictions on geospatial time series. Such models might allow for more interpretability as they more closely resemble the mechanics of physically-based models. With these models, we could further train one model to predict streamflow at arbitrary locations, which allows for more detailed comparisons to physically-based models.

Besides ridge regression and XGBoost, we have begun to explore predictions with neural networks, specifically simple long short-term memory cell (LSTM) architectures. In preliminary experiments, however, these neural models did not perform better than XGBoost. We assume that more sophisticated deep-learning approaches that are better suited for spatially-distributed input data would improve prediction accuracy.

VI. CONCLUSION

Our study shows that data-driven approaches predict streamflow more accurately than a physically-based model for the specific case study that we examined. The more rigid structure of a physically-based model prevents it from fully exploiting signals that are available in the input data. Especially for stations in highly urbanized regions, our data-driven models are better able to adapt to patterns of human water regulation.

From a high-level perspective, our project has the goal to both deliver more accurate streamflow predictions and advance hydrological understanding. We have taken only a small step in this direction, but are excited about future prospects.

ACKNOWLEDGMENTS

The authors would like to thank the Global Water Futures program and the Integrated Modeling Program for Canada (IMPC) for financial support of Juliane Mai, Shervan Gharari, and Martin Gauch. The authors wish to thank Nicolas Gasset and Vincent Fortin from Environment and Climate Change Canada for making the meteorological forcing dataset available for the GRIP-E project.

REFERENCES

[1] J. Mai, B. Tolson, H. Shen, E. Gaborit, V. Fortin, M. Dimitrijevic, N. Gasset, D. Durnford, Y. L. Shin, T. A. Stadnyk, L. M. Fry, T. Hunter, A. Gronewold, J. Smith, L. Mason, L. Read, K. FitzGerald, K. M. Sampson, A. F. Hamlet, F. Seglenieks, S. Gharari, S. Razavi, A. Haghnegahdar, D. G. Princz, and A. Pietroniro, "The Great Lakes Runoff Inter-comparison Project for Lake Erie (GRIP-E)," *AGU Fall Meeting Abstracts*, 2018.

[2] X. Liang, D. P. Lettenmaier, E. F. Wood, and S. J. Burges, "A simple hydrologically based model of land surface water and energy fluxes for general circulation models," *Journal of Geophysical Research*, vol. 99, no. D7, pp. 14415–14428, 1994.

[3] J. J. Hamman, B. Nijssen, T. J. Bohn, D. R. Gergel, and Y. Mao, "The Variable Infiltration Capacity model version 5 (VIC-5): infrastructure improvements for new applications and reproducibility," *Geoscientific Model Development*, vol. 11, no. 8, 2018.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 785–794, ACM, 2016.

[5] H. V. Gupta, H. Kling, K. K. Yilmaz, and G. F. Martinez, "Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling," *Journal of Hydrology*, vol. 377, no. 1-2, pp. 80–91, 2009.

DeepRainK: CONVLSTM NETWORK FOR PRECIPITATION PREDICTION USING HYBRID SURFACE RAINFALL RADAR DATA

Seongchan Kim[1], Ji-Sun Kang[2], Sa-kwang Song[1,3], Chang-Geun Park[4], Baek-Jo Kim[4]

***Abstract*—Precipitation forecasting is important because it has a considerable effect on economic and social activities. In this study, we propose a precipitation nowcasting model called *DeepRainK*. The model uses convolutional Long-Short Term Memory to predict the next 10 to 30 minutes of radar images from Hybrid Surface Rainfall (HSR) radar data of previous 2 hours. HSR is a radar composite, and it is characterized by the synthesis technique based on multiple altitude angles to produce near-terrestrial data, minimizing the effects of topographic shielding. For the experimental design, HSR data are used with input in a time series format in units of 10 minute divided into 12 records, to train the *DeepRainK*. The output is the predicted radar images for the next 30 minutes. The predicted radar data are translated into a precipitation amount by the quantitative prediction estimation method.**

I. INTRODUCTION

Precipitation information has a profound impact on everyday life, through management of the agriculture and construction industries for example. Owing to the importance of this task, meteorologists have been making great efforts toward building advanced forecasting models of weather and climate, mainly employing numerical models based on High Performance Computing (HPC).

In general, weather radar observations are the data used for numerical forecasting and hydro logic models to improve the accuracy of weather forecasts for hazardous weather events such as heavy rains and typhoons [1]. In particular, weather radar data refers to data represented by a radar image; the image is composed using the moving speed, direction, and strength of a signal transmitted by a radar transmitter into the atmosphere, and received after it has collided with water vapor or similar particles. The observational data are processed and synthesized into various composite indicators such as Plan Position Indicator (PPI), Constant Altitude Plan Position Indicator (CAPPI), BASE, ECHO TOP, CMAX, and Hybrid Surface Rainfall (HSR).

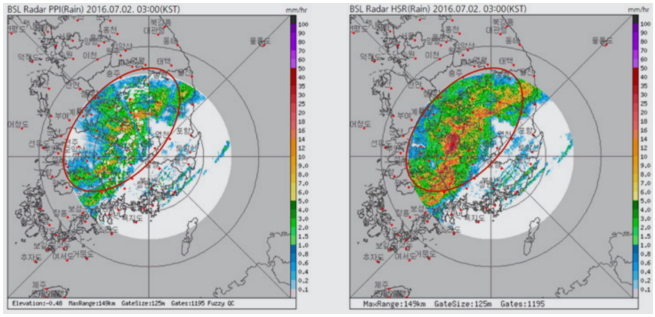


Fig. 1. Reduced rainfall amount estimation of PPI due to beam shielding and eliminating space in the terrain echo area (Left: BSL Radar PPI, Right: BSL Radar HSR, 2016.07.02 03:00 KST)

In Korea, since there are many mountainous areas, the data are often severely distorted due to lower observation levels from beam shielding, while non-meteoric echoes such as blue echoes and chaff occur frequently. Therefore, the KMA Radar Center[1] has developed a radar-based rainfall synthesis technique using the HSR method to minimize the radar's observation limits and error factors, and to produce a high accuracy radar rainfall field on the Korean peninsula scale [2]. HSR data are characterized by the synthesis technique based on multiple altitude angles to produce near-terrestrial data, minimizing the effects of topographic shielding. As a result, the accuracy of the radar estimated rainfall has been gradually improving. Figure 1 shows a comparison of PPI and HSR radar images.

In the study, we introduce *DeepRainK*, with a subsequent model of *DeepRain* ([3]) that targets the Korea peninsula to estimate the rainfall amount based on HSR

*Corresponding Author: Seongchan Kim, sckim@kisti.re.kr
[1]Research Data Sharing Center, KISTI, Korea; [2]Supercomputing Infra Center, KISTI, Korea; [3]Dept. of Data & HPC Science, UST-KISTI, Korea; [4]High Impact Weather Research Center, National Institute Meteorological Sciences, KMA

[1]https://radar.kma.go.kr

radar data. Compared to *DeepRain*, the contributions of this study for *DeepRainK* are as follows:

1) HSR radar data optimized for the mountainous areas, such as Korea, are used as input into *DeepRainK*.
2) The model architecture is renewed from many-to-one to many-to-many (encoder-decoder), to produce future radar data at the target times.
3) The quantitative rainfall estimation technique(Z-R relationship) is applied to determine rainfall amounts from radar data.

II. Related Work

In recent years, studies using deep learning techniques have been widely used to improve the accuracy of weather prediction. The convolution neural network (CNN) and recurrent neural networks (RNN), and the combination of both methods such as convolutional LSTM (ConvLSTM), are popular techniques for the prediction of weather-related events. Several studies [4], [5] have employed CNN for precipitation prediction, and [6], [7] have tried using combinations of these methods. Primarily, ConvLSTM, which is a variant of LSTM, was devised by Shi et al. [6] to embed the convolution operation inside the LSTM cell and model spatial data more accurately. The authors demonstrated that the ConvLSTM effectively predicts precipitation in their experiments. Kim et al. [3] adopted ConvLSTM for three-dimensional and four-channel radar data to predict the rainfall amount. They stacked the ConvLSTM cells for performance enhancement and confirmed the proposed method is more effective for predicting rainfall than linear regression. Souto et al. [8] suggested an ensemble method comprising a set of meteorological models for rain. The output of each model is given to a channel of ConvLSTM, and demonstrated that the combination method predictions are promising in a spatiotemporal context. Tan et al.[9] proposed the hierarchical ConvLSTM (FORECAST-CLSTM) model utilizing a new forecaster loss function. They predicted future satellite cloud image for cloudage nowcasting, and this model retained the uncertainty level of real atmospheric conditions better than the conventional loss function. Ayze et al. [10] used CNN for radar-based precipitation nowcasting. They verified the impact of different data preprocessing routines, loss functions, and convolutional kernel sizes. Their results show a comparable efficiency in comparison with the state-of-the-art optical flow based model and can be used as a reliable baseline for further development. However, none of the study utilized HSR radar data for their deep learning model yet. In this study, HSR is inputted first into the ConvLSTM model for precipitation nowcasting.

III. Data

This section describes data that were used for our experiment. The HSR data we foucus on are obtained from the KMA Radar Center[1]. They built the dataset for operational purposes and Table I lists the specification of the data. The data are radar observations from the Korea peninsula area.

TABLE I
DATA SPECIFICATION

Item	Spec
Time Resolution	5 min
Spatial Resolution	500 m
Projection Grid	2305 × 2881
Projection Method	Lambert conformal conic projection
Reference Diameter	N 38.0, E 126.0
Reference Grid Point	1121, 1682

The HSR file (binary), contains radar reflectivity (float, up to second decimal place), as well as header data including radar observation time, radar station code, and map information. We obtain total of three years of HSR dataset from May 2018 to July 2019, including 157,680 (=1095 days of 24 h 6 times, at 5 min interval) files. One compressed HSR file is about 3.42 Mb, and the whole set is about 540 Gb. The whole set is separated into training (60%), validation (20%), and test set (20%). From each split, we create successive time series data that have 15 time span (12 for encoding and 3 for decoder as labels) with 10 min intervals through incrementing of the time index by one. The successive data are used as input for the encoder of ConvLSTM.

IV. Method

DeepRainK was designed with a ConvLSTM cell, that is as a variant of LSTM. It replaces the matrix multiplication with convolution operations at each gate in the LSTM cell to capture spatial information from the input [6]. ConvLSTM cell keeps dimension of input(3-D) unlike data flows as 1-dimension vector in LSTM cell. ConvLSTM cell has four gates(forget gate, input gate, output gate) inside, which can help learn long-term dependencies in the network. Forget gate determines what information is thrown from the previous cell state, input gates decides to store what new information in the cell state, and output gate determines what will be the

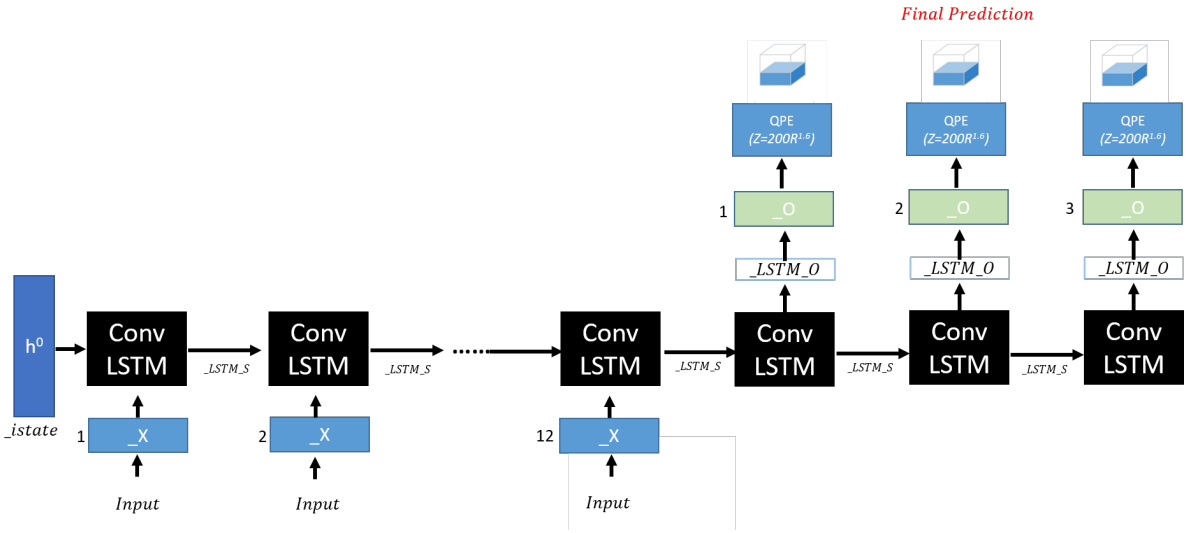


Fig. 2. *DeepRainK* architecture using ConvLSTM

output on the cell. state.[11] The equations of the gates (input, forget, and output) in ConvLSTM are as follows:

$$i_t = \sigma(W_{xi} * x_t + W_{hi} * h_{t-1} + b_i) \quad (1)$$

$$f_t = \sigma(W_{xf} * x_t + W_{hf} * h_{t-1} + b_f) \quad (2)$$

$$o_t = \sigma(W_{xo} * x_t + W_{ho} * h_{t-1} + b_o) \quad (3)$$

$$C_t = f_t \circ C_{t-1} + \tanh(W_{xc} * x_t + W_{hc} * h_{t-1} + b_c) \quad (4)$$

$$H_t = o - t \circ \tanh(c_t) \quad (5)$$

Figure 2 describes the architecture of *DeepRainK*. *DeepRainK* has a many-to-many structure (so called sequence-to-sequence model) that has an encoder and decoder to predict future radar images based on previous radar images. In other words, *DeepRainK* predicts how the radar image will be transformed over the next 10 to 30 minutes from the 2 hour input.

The input data X of the encoder receive 12 items of data according to the 10 minutes interval, and the shape of input data at each node is 10,201 ($101 \times 101 \times 1$) for radar reflection values (float) data. Input data X are cropped into 100×100 from he HSR original size of 2305×2881, to focus only on the Seoul area of the Korea peninsula. For height, we use one level of 0.5km. The decoder of the model produces three radar images at 10 to 30 min, with the same shape of the input. Note that this model predicts the next sequences of radar images. For estimating the amount of rainfall from the output radar data, we use quantative precipitation estimation (QPE). The *Z-R* relationship is adopted to estimate the rainfall amount. According to different types of precipitation, different *Z-R* relationships could be applied [12]. For evaluation, we compare the rainfall amounts from our model with ground truths, which are the actual observational data of amounts of rainfall corresponding to periods and times. The results are validated with a set of evaluation methods including correlation, accuracy, CSI, and bias. The obervational data are obtained from the KMA Data Portal [2].

V. EXPERIMENT

The proposed model (Figure 2) was trained with the Adam optimizer at a learning rate initially. We use random shuffling mini-batches for using the stochastic gradient descent learning method. The mean square error (MSE) between the predicted radar image and the actual image is used to measure the prediction loss. The testbed environment (workstation) configuration is as follows:

- CPU: Intel Xeon Gold 6138 (3.7GHz) Turbo 20C
- RAM: 128GB 2666MHz DDR4
- GPU: NVIDIA Quadro GV100 (x1), NVIDIA Titan RTX (x2)
- SSD: M.2 1TB PCIe NVMe Class 40
- HDD: 4TB 7200rpm Nearline SAS
- Framework: TensorFlow 1.13.1, Python 3.6.3, Keras 2.2.4

We build a cluster for a distributed training framework with one workstation and three PCs with horovod [3].

As a baseline, we will use MAPLE (McGill Algorithm for Precipitation nowcasting by Lagrangian

[2]https://data.kma.go.kr
[3]https://github.com/horovod/horovod

Extrapolation). MAPLE estimates the next radar image referring to motion vectors of radar precipitation echo based on variational methods, and then applies the QPE to produce a rainfall amount. KMA have been making precipitation forecasts by using MAPLE since 2008. We will compare the prediction results (rainfall amount) with our results. Furthermore, we will discuss the two models' strengths and drawbacks.

VI. Conclusion and Future Works

In this study, we proposed *DeepRainK* utilizing ConvLSTM to learn HSR radar data to predict the next sequence of radar images, then finally producing rainfall amounts at 10, 20, and 30 min. We will compare our experimental results with the ones of MAPLE. Future studies will utilize additional four-heights of HSR radar data from 0.5km to 3.5km for 3D convolution. 3D convolution may capture vertical information of the cloud in radar images, which can simulate more accurate cloud movements. Furthermore, it is under consideration to input the model dual polarization radar data to enhance performance. Dual polarization radar sends and receives both horizontally and vertically polarized electromagnetic waves. The two-dimensional wave foster measuring the sizes and shapes of clouds. Finally, we will extend the modeling area from Seoul to entire the Korean peninsula.

Acknowledgments

This research was supported by "Development of typhoon analysis and forecast technology (1365003070)" of the National Typhoon Center, and "Construction and Operation of National Research Data Platform (K-19-L01-C04-S01)" of Korea Institute of Science and Technology Information (KISTI).

References

[1] "Korean National Weather Radar Center, http://radar.kma.go.kr/eng/radar/composition.do."

[2] S. H. C. Young-a Oh, Mi-Kyung Suk, "Multi-altitude angle based radar estimation rainfall improvement and synthesis technology development," in *Proceedings of the Autumn Meeting of KMS, 2018*.

[3] S. Kim, S. Hong, M. Joh, and S.-k. Song, "DeepRain: ConvLSTM Network for Precipitation Prediction using Multichannel Radar Data," in *Climate Informatics Workshop*, nov 2017.

[4] W. D. Yong Zhuang, "Long-lead prediction of extreme precipitation cluster via a spatio-temporal convolutional neural network," in *Proceedings of the 6th International Workshop on Climate Informatics: CI 2016, NCAR Technical Notes NCAR/TN-529+PROC*, 2016.

[5] W. Zhang, L. Han, J. Sun, H. Guo, and J. Dai, "Application of Multi-channel 3D-cube Successive Convolution Network for Convective Storm Nowcasting," in *CVPR*, 2017.

[6] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *CVPR*, jun 2015.

[7] Q.-K. Tran and S.-k. Song, "Computer vision in precipitation nowcasting: Applying image quality assessment metrics for training deep neural networks," *Atmosphere*, vol. 10, no. 5, 2019.

[8] Y. M. Souto, F. Porto, A. M. Moura, and E. Bezerra, "A spatiotemporal ensemble approach to rainfall forecasting," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, July 2018.

[9] C. Tan, X. Feng, J. Long, and L. Geng, "Forecast-clstm: A new convolutional lstm network for cloudage nowcasting," in *2018 IEEE Visual Communications and Image Processing (VCIP)*, pp. 1–4, Dec 2018.

[10] G. Ayzel, M. Heistermann, A. Sorokin, O. Nikitin, and O. Lukyanova, "All convolutional neural networks for radar-based precipitation nowcasting," *Procedia Computer Science*, vol. 150, pp. 186–192, jan 2019.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[12] W. Wu, H. Zou, J. Shan, and S. Wu, "A Dynamical Z - R Relationship for Precipitation Estimation Based on Radar Echo-Top Height Classification," *Advances in Meteorology*, vol. 2018, pp. 1–11, aug 2018.

IMPACT OF SPARSE PROFILE SAMPLING ON THE RECONSTRUCTION OF SUB-SURFACE OCEAN TEMPERATURE FROM SURFACE INFORMATION

Natacha Galmiche[1,2], Julien Brajard[2,3], Anastase Charantonis[3,4,5] , Tsuyoshi Wakamastu[2,6]

***Abstract*—Projection of the dense surface measurement to subsurface using background information is a key approach in mapping the 3D ocean temperature. In the last years, a new inverse method combining self-organizing map (SOM) and hidden Markov model (HMM), PROFHMM, has been proposed for retrieving sub-surface ocean variables from the surface measurements. So far, this method requires a training dataset providing complete time-series of sets of surface and subsurface data. However, in reality, the spatiotemporal resolution of sub-surface data is sparser than the surface one and some subsurface profiles can be missing in a given training dataset.**

In this study, we extend PROFHMM to handle the sparsity in the subsurface vertical profile time-series and introduce a method to evaluate its robustness with respect to the proportion of profile available at an instant t. In our specific experiments where temperature profiles are reconstructed from sea surface temperature and shortwave radiation data at 5 days interval, we found that the method works as long as this proportion is higher than 70%.

I. MOTIVATION

Three dimensional (3D) mapping of ocean temperature is a challenging task due to the nature of sampling strategies in the current global ocean observing systems. Sea surface temperature (SST) measurement from space provides the richest information regarding spatial and temporal resolutions but its access is limited to the surface information. Autonomous floats networks represented by Argo program [1] are sampling subsurface profiles in global scale, but its target spatiotemporal coverage is sparse compared to the SST measurement. Projection of the dense surface measurement to subsurface using background information is a key approach in mapping the 3D ocean temperature. Recently, the new projection method based on the combination of a self-organizing map (SOM) and a hidden Markov model (HMM), PROFHMM, for retrieving sub-surface ocean variables from the surface measurement is proposed [2].

In this study, we have examined the robustness of PROFHMM under the case of temporally sparse subsurface sampling.

Corresponding author: N. Galmiche, natacha.galmiche@gmail.com [1]Ecole Nationale Supérieure d'Electrotechnique, d'Electronique, d'Informatique, d'Hydraulique et des Télécommunications (ENSEEIHT) [2]The Nansen Center, Bergen, Norway

[3]Sorbonne University, Paris, France

[4]Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise (ENSIIE), Evry, France [5]Laboratoire de Mathmatiques et Modélisation d'Évry, France

[6]Bjerknes Centre for Climate Research (BCCR), Bergen, Norway

II. METHOD

A. Self-Organizing Map (SOM)

A self-organizing map (SOM) is an artificial neural network (ANN) that performs dimensionality reduction and clustering of a given set of training data [3]. SOM can be considered as undirected graphs whose vertices are neurons (or classes) which are associated to a referent vector (or weights). A class is represented by an index defining its position within the SOM and a referent vector is a vector of the same size as the data we want to classify. The set of classes are representative of a whole cluster of these data.

In PROFHMM [2], a set of two SOMs have been trained: one denoted M_{obs} for the surface data and one denoted M_{dis} for the subsurface.

In our experiments we have:

- Surface SOM M_{obs} containing $N_{\text{obs}} = 2 \times 70 = 140$ classes.Referent vectors of dimension 2 represent the sea-surface temperature and the incoming shortwave radiation.
- Vertical profile SOM M_{dis} containing $N_{\text{dis}} = 4 \times 70 = 280$ classes. Referent vectors of dimension 26 represent the temperature at 26 different levels of depth from $0m$ to $92m$ with a higher vertical resolution near the surface.

Note that the purpose of this study is not to find the optimal PROFHMM configuration for this problem, but to extend the method to the case with missing data and to present a method to evaluate its robustness. For the same dataset, a better configuration could be found and provide better results.

The topologies of both M_{obs} and M_{dis} are rectangular maps. The coordinates of a particular class of one of these maps, denoted c^t ($t \in \mathbb{N}$, arbitrary here), is given by a couple of integers (x_t, y_t). The distance $D(c^1, c^2)$ between 2 classes c^1 and c^2 is the Euclidean distance

$$D(c^1, c^2) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}. \quad (1)$$

As an ANN, there are two phases in the SOM construction process: a training phase and a mapping phase. The training phase builds the map. For each training iteration, all the referent vectors are updated as following [3]:

1) Finds the Best Matching Unit (BMU): class whose referent vector is the closest to the input (in the input data space)
2) Update every referent vector:
   a) compute the distance between the class associated with the referent vector and the BMU (in the map)
   b) update the referent vector so that the closer their associated class is to the BMU (in the map) the more their referent vector will be modified towards the input data.

After training, the mapping phase classifies a new input vector into the class associated with the nearest referent vector.

One of the main advantages of applying a SOM to oceanographic data, which are conditioned and constrained by physical laws is that referent vectors associated to topologically neighbouring classes will have similar physical properties and behaviours (e.g. they represent the same period of the year).

B. Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is a Markov model in which the system has unobservable (hidden) states. In simpler Markov models, the state is always visible and depends solely on the previous state. In an HMM, the current observable state depends only on the current hidden state (emission) and the current hidden state depends only on the previous hidden state (transition). The parameters of a HMM are the probabilities of every emission and transition, they are stored in 3 matrices: Tr, Em and $\boldsymbol{\pi}$:

$$\text{Tr}(i,j) = p(c^t_{\text{dis}} = j | c^{t-1}_{\text{dis}} = i) \quad \begin{cases} i,j \in [1..N_{\text{dis}}] \\ c^t_{\text{dis}}, c^{t-1}_{\text{dis}} \in M_{\text{dis}} \end{cases},$$

$$\text{Em}(i,j) = P(c^t_{\text{obs}} = j | c^t_{\text{dis}} = i) \quad \begin{cases} i \in [1..N_{\text{dis}}] \\ j \in [1..N_{\text{obs}}] \\ c^t_{\text{dis}} \in M_{\text{dis}} \\ c^t_{\text{obs}} \in M_{\text{obs}} \end{cases},$$

$$\boldsymbol{\pi}(i) = p(c^0_{\text{dis}} = i) \quad \begin{cases} i \in [1..N_{\text{dis}}] \\ c^0_{\text{dis}} \in M_{\text{dis}} \end{cases},$$

called transition matrix, emmission matrix and the initial probabilities, respectively [4] [5] [6].

These matrices are initialized by counting transitions and emissions that occurred in the training dataset [7]. The elements of the matrices are then optimized using the Baum-Welch algorithm (B-W) [5].

From the setting of SOMs in our experiments, we have

- 140 classes of M_{obs}: 140 referent surface data (observable states),
- 280 classes of M_{dis}: 280 referent vertical profiles (hidden states).

III. Contribution

A. *Initializing* Em *and* Tr

Extension of PROFHMM to the sparse sub-surface training dataset is achieved by implementing a new counting strategy in initializing Em and Tr with sparse time-series as follows:

- An emission at instant t is counted only if the vertical profile at this instant t is not missing
- A transition at instant t is counted only if both the vertical profile at instants t and $t-1$ are not missing

Thus if one profile (instant t) is missing we lose 2 transitions ($t-1 \to t$ and $t \to t+1$) and 1 emission (instant t).

Our choice of the counting strategy for sparse time series is the simplest way we could imagine, but other methods of counting emission and transition could be implemented and compared in future work. For instance provided a good enough time resolution, a transition between a state at instant t and $t+2$ could be counted with some weight if $t+1$ state is missing.

Note also that not only surface data which are associated with a non-missing vertical profile, but all the surface observations can be used in the B-W algorithm.

B. Radius of neighborhood function

In practice, there is not enough training data to estimate properly the elements of Em and Tr (and especially Tr) even with a complete set of surface and subsurface time-series. To overcome this problem, Charantonis (2015) [2] suggested using a neighbouring function that exploits the topological properties of the SOM. We used a similar yet slightly different neighbourhood function to propagate the probability of a class to its neighbour. We first calculate $\mathrm{Tr}^{\mathrm{b-w}}$ and $\mathrm{Em}^{\mathrm{b-w}}$ as described previously, using the counting procedure and the B-W algorithm (the subscript $\mathrm{b-w}$ stands for "B-W algorithm"), then we improve the values by calculating Tr^{s} and Em^{s} (the subscript s stands for "smooth") using the following expressions:

$$\mathrm{Tr}^{\mathrm{s}}(i,j) = S_{\mathrm{Tr}}^{-1} \sum_{k=1}^{N_{\mathrm{dis}}} exp(-\frac{D(j,k)}{\sigma})\, \mathrm{Tr}^{\mathrm{b-w}}(i,k) \quad (2)$$

$$\mathrm{Em}^{\mathrm{s}}(i,j) = S_{\mathrm{Em}}^{-1} \sum_{k=1}^{N_{\mathrm{dis}}} exp(-\frac{D(j,k)}{\sigma})\, \mathrm{Em}^{\mathrm{b-w}}(i,k) \quad (3)$$

Where $D(j,k)$, see Eq. (1), is the distance between the class j and k within M_{dis}, σ is the so-called radius of the neighborhood function and S_{Tr}, S_{Em} are normalizing factor such as $\sum_{j=1}^{N_{\mathrm{dis}}} \mathrm{Tr}^{\mathrm{s}}(i,j) = \sum_{j=1}^{N_{\mathrm{dis}}} \mathrm{Em}^{\mathrm{s}}(i,j) = 1$. σ is a regularizing term: if σ is small, the smooth quantities will be very close to the original B-W computation, if σ is large, Tr^{s} and Em^{s} will not vary much throughout the map. In our case, σ is computed according to the actual transitions that occurred while training M_{dis} Therefore, σ represents the typical distance of a transition in the training time-series. This typical distance is strongly linked to the nature of the data used and the shape of the SOM, thus using the same radius as [2] used could have been irrelevant. Instead, we defined a criterion that could be used for any other application and configuration of PROFHMM.

C. Penalize long distance transitions

In the case of complete times-series, counting and smoothing in initializing Em and Tr give good enough results and B-W does not have a great impact on the accuracy of the estimates. However, since B-W is an expectation-maximization algorithm, it only needs as input the observation time-series. Then the fewer vertical profiles we had compared to surface data during the initialization the greater the impacts of B-W and observable time-series are. This might cause some discontinuity in the vertical profile reconstructed. To prevent this problem we used another neighbourhood function that decreases the probability to transit from i to j if they are far from each other in M_{dis}. Once again we used the same radius of σ in this neighbourhood function.

$$\mathrm{Tr}^{\mathrm{l}}(i,j) = S^{-1} exp(-\frac{D(i,j)}{\sigma})\, \mathrm{Tr}^{\mathrm{s}}(i,j)$$

where the subscript l stands for "local" and S is a normalizing factor such as $\sum_{j=1}^{N_{\mathrm{dis}}} \mathrm{Tr}^{\mathrm{s}}(i,j) = 1$.

IV. Evaluation

A. Experiments setup

Time series of temperature profile data are produced by one-dimensional numerical ocean model GOTM configured with atmospheric forcing at the Bermuda Atlantic Time-series Study (BATS) site [8]. Starting from the initial state in 1990, the model was integrated up to 2007. The model states and the surface atmospheric forcing are saved every 12 hours. The training and testing sets are then generated by temporally averaging hidden state: temperature profile and observable states: sea surface temperature (SST) and downward shortwave radiation (DWSR) within 5 days window. The temperature profile is further averaged in the vertical axis over 5 model levels. Removing the first two years record as model spinning-up period, we define training and testing datasets as follows:

- Training dataset: 1992-2003: total length of time-series: 879
- Testing dataset: 2004-2007: total length of time-series: 292

The choice of DWSR as an observable state in addition to SST is made based on the dynamical understanding that DWSR is a key forcing driver that introduces seasonal asymmetry in the sub-surface temperature profile in Spring and Autumn.

To emulate missing profiles we used a Bernoulli distribution of parameter p at each instant t. To have a first overview of the behaviour of the method applied with missing data, we drew 10 subsets of the complete data set for each p decreasing from 1 (no missing data) to 0.50 with 0.05 steps. According to the results we then focused on $0.65 < p \leq 1$, drawing this time 50 subset for each p.

The same SOM M_{obs} and M_{dis} were used for mapping all the vertical profiles and they were trained

with a complete time-series. In the ideal case, a new SOM should have been trained for each sample and each probability. This would have been extremely time-consuming and would not have had a great impact on the results since the SOM needs fewer vertical profiles to give satisfying results and it is not the main cause of the error in the final reconstruction. However, different values of Tr and Em are estimated for each p and each subset.

The quality of the profile reconstruction on the testing dataset is evaluated using the root-mean-square error (RMSE) with respect to each depth as a metric. Another metric could have been selected if the study called for it (such as the maximum error in time for example).

The results are compared to a lower-bound which is the climatology (the average profile of each given date over the years 1992-2003). Nevertheless, note that this lower-bound is computed using the complete dataset without missing data, and by consequence should be considered as a reference value for the performance of the algorithm.

B. *Results and interpretation*

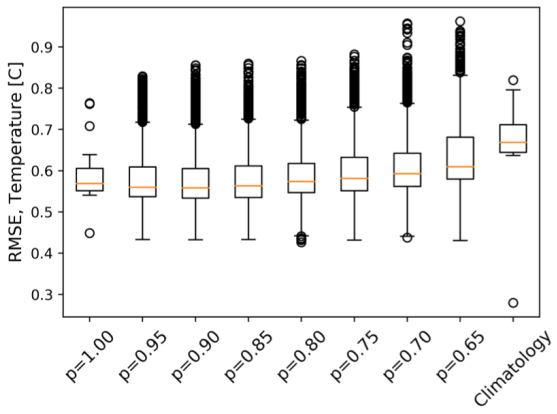


Fig. 1. Boxplot comparison of performance of PROFHMM for different values of p (probability of missing a vertical profile).

For values of p down to 0.7, the reconstructions obtained are close to the reconstruction obtained with $p = 1$, as seen in Fig. 1. The RMSE observed remain consistently bellow those obtainable if simply using a climatology.

In Fig. 2, it can be seen that, the performance is a lot better for the first 40 meters of depth, which is where the results matter the most in this problem (See Figure 2).

Finally, the method outperforms the climatological reconstruction as long as the probability of getting a

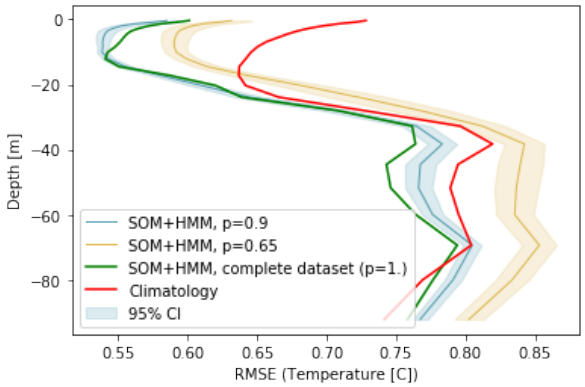


Fig. 2. RMSE with respect to depth on the testing data set.

vertical profile is higher than 0.70 (See Figure 1). This is highly encouraging, especially given that the configuration of PROFHMM has not been optimized and that further improvement could be done for initializing the matrices Em and Tr. It suggests that, for this particular problem of temperature reconstruction, PROFHMM could be applied in a realistic in-situ dataset with missing profiles such as Argo floats measurements. Further investigations remain to be done to generalize these results to other problem (e.g. chlorophyll-a profiles) or other time and spatial scales.

ACKNOWLEDGMENTS

This work was supported by the strategic institute program (SIS) of the Nansen Center funded by the Norwegian Ministry of Climate and Environment (KLD).

REFERENCES

[1] S. C. Riser, H. J. Freeland, D. Roemmich, S. Wijffels, A. Troisi, M. Belbéoch, D. Gilbert, J. Xu, S. Pouliquen, A. Thresher, *et al.*, "Fifteen years of ocean observations with the global argo array," *Nature Climate Change*, vol. 6, no. 2, p. 145, 2016.

[2] A. A. Charantonis, F. Badran, and S. Thiria, "Retrieving the evolution of vertical profiles of chlorophyll-a from satellite observations using hidden markov models and self-organizing topological maps," *Remote Sensing of Environment*, vol. 163, pp. 229–239, 2015.

[3] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[4] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[5] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains," *Ann. Math. Statist.*, vol. 41, pp. 164–171, 02 1970.

[6] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford university press, 1995.

[7] N. C. Laan, D. F. Pace, and H. Shatkay, "Initial model selection for the baum-welch algorithm as applied to hmms of dna sequences," *Queens University, Kingston, ON, Canada*, 2006.

[8] M. Butenschön, J. Clark, J. N. Aldridge, J. I. Allen, Y. Artioli, J. Blackford, J. Bruggeman, P. Cazenave, S. Ciavatta, S. Kay, *et al.*, "Ersem 15.06: a generic model for marine biogeochemistry and the ecosystem dynamics of the lower trophic levels," *Geoscientific Model Development*, vol. 9, no. 4, pp. 1293–1339, 2016.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

RECONSTRUCTION OF THE PALEOCLIMATE FROM PROXIES RECORDS: A MACHINE LEARNING INVESTIGATION

Marie Déchelle[1], Anastase Charantonis[1,2,3], Beyrem Jebri[1], Myriam Khodri[1], Sylvie Thiria[1]

***Abstract*—In the present work, we investigate the capacity of machine learning to reconstruct simulated large-scale surface temperature anomalies given a sparse observation field. Several methods are combined: self-organizing maps and recurrent neural networks. To evaluate our global scale reconstruction, we base our validation on global climate indices time series. In our experiments, the reconstructions of the global surface temperature anomalies obtained provide a good correlation (over 90%) with the target values when considering scarce available observations sampling about 0.5% of the globe.**

I. MOTIVATION

To understand the various processes governing the climate system, we need to rely on available observations (field and satellite data) and general circulation models (GCMs). Whereas the first ones are limited to the last 70 years or so, the second ones are still improvable. Relying on past climate reconstructions deduced from the analyses of natural archives or proxies (such as sediment cores, corals, tree-rings, etc.) can help expanding the window of observations beyond the instrumental period over the past millennia and allow understand the processes governing natural climate variability before the industrial era [1]. Available proxies records provide a spatio-temporally scattered and sparse data set of climate variables such as temperature and precipitation anomalies, with various time resolutions, uncertainties in amplitude and temporal positioning, that can be translated into time series of varying length and confidence [2].

As such information is patchy, statistical methods are needed to allow bridging the global and regional scales [3]. On the one hand, previous studies such as [3] use state-space model based on maximum likelihood methods to estimate models parameter for proxy-based climate reconstruction. On the other hand, self-organizing maps have long been used to look into seasonal cycles in GCM, such as done in [4].

In our present work we investigate the ability of neural networks to reconstruct climate anomalies from spatially sparse temperature time sequences. As a first step, we rely on a perfect model approach and attempt to reconstruct global surface temperature anomalies for the pre-industrial climate as simulated by a state-of-the-art GCM. In section II, we detail the data used, as well as our methodology regarding the neural networks deployed and the validation approach. In section III we present and discuss the obtained results.

II. MATERIALS AND METHOD

A. Materials

This study is based on the simulated surface temperatures in a steady 2500 years-long pre-industrial control run by the coupled ocean-atmosphere IPSL-CM5A2 model [5].

1) Initial data set: It consists of 2750 gridded maps composed of 96*96 pixels, representing monthly surface temperature with a resolution of 3.75°in longitude and 1.875°in latitude on land. The ocean resolution is 2°. In the tropical region, the latitudinal resolution decreases to 1/2°[5]. To be consistent with the length of available instrumental observations, we separate our data set into 3 parts: a training set of 150 years, comprising a validation set of 30 years, and a test set consisting of 100 years.

2) Sparse observations: To be consistent with the scarce coverage of available proxies records, we build a test case by selecting grid points randomly on the globe in both hemispheres, over land and ocean. We however sample more land grid points since paleoclimatic data over the last 2000 years are predominantly located on land and in the northern hemisphere [2].

Corresponding author: Marie Déchelle, marie@dechelle.net
[1]Laboratoire d'Océanographie et du Climat : Expérimentations et Approches Numériques [2]Laboratoire de Mathématiques et Modélisation d'Évry [3]École nationale supérieure d'informatique pour l'industrie et l'entreprise

3) Preprocessing: The data are preprocessed to remove the mean seasonal cycle and focus on annual climate anomalies. For this purpose, a 30 years mean climatology is removed and a low-pass filter (sliding window average with a width of 12 months) is applied on the monthly anomalies to remove the intra seasonal variability.

B. Methods

In order to reconstruct climate anomalies from spatially sparse temperature time series, we combine a variety of algorithms. They are implemented in Python and MATLAB, and the codes are accessible at: https://github.com/MarieDchelle/CI_2019.

1) Self-Organizing Map (SOM): SOM is an unsupervised multidimensional clustering method based on neural networks. Each cluster is represented by a referent vector in the data space and an index on a topological map. This one corresponds to an organized discrete space in two dimensions, such that two nodes that are nearby on the topological map have referent vectors that are close in the data space [6].

In this study, SOM algorithm is initially applied to spatially define N geographical areas based on the similarity of the time-series of temperature over the training data set, as discussed in III-A.

2) ItCompSOM: As introduced by [7], ItCompSOM is an iterative completion method drew from SOM, applicable when completing large, highly correlated, multidimensional databases containing a significant part of missing data. When reconstructing missing values of a vector through analogue methods, the equivalent on which the reconstruction is grounded is usually selected by computing the Euclidean distance over the non-null elements of the observation and their potential referents.

ItCompSOM relies on the inherent correlation between the variables, aiming at assigning a vector to a referent based on the correlation between the observed part of the vector and its missing values. This is done by using a similarity function between a vector X in the data space containing missing and non-missing components and a referent vector ref^c of the SOM class c, denoted $s^c(X, ref^c)$. It is defined as [8] :
$s^c(X, ref^c) = \sum_i (1 + \sum_j (cor^c_{i,j})^2) * \sqrt{(X_i - ref^c_i)^2}$
where $cor^c_{i,j}$ is the local correlation between the missing (j) and non-missing (i) variables computed over the data attributed to the class c during the training phase of the SOM algorithm.

In this study, ItCompSOM is used to complete the temperature anomalies missing values of the masked clusters from the N different geographical regions determined by the spatial clustering. It is done by using the information from the unmasked clusters, as discussed in III-B.

3) Recurrent Neural Network (RNN): RNNs are a family of shared-weights neural networks for processing sequential data [9], exhibiting temporal dynamic behavior. Such a network can use the coherence of the climate evolution to correct a sequence of time series initially reconstructed with ItCompSOM which does not explicitly take into account the temporal aspects of the data.

In this work, a Long Short Term Memory (LSTM) network is trained over the 150 years-long training period with an internal memory of 2 years. It takes as inputs the reconstructions through ItCompSOM of the N clusters obtained with SOM (II-B1) of the artificially masked training set, and as targets the clusters' actual values. The model learns the regions' dynamic behavior for 2 years and is then applied over the whole test period (1200 time steps) to obtain the N reconstructed sequences of temperature anomalies. The network is structured as follows : two bidirectional layers with respectively 30 and 20 units, followed by two dense layers with 90 and 191 neurons. It contains 87,311 weights in total. During the training, a L1_L2 kernel regularization is applied and we use the Adam optimizer.

C. Evaluation

When investigating the climate system, we validate the developed methodology on the basis of the Root Mean Square Error (RMSE) between the reconstructed surface temperature and the simulated truth. To verify the reconstruction, the RMSE in °C between the reconstructed and target images is computed over time and space such that
$RMSE = \sqrt{\frac{1}{9216*T} \sum_t \sum_p (x^r_{t,p} - x^o_{t,p})^2}$
where $x^r_{t,p}$ denotes the reconstructed pixel p at time t and $x^o_{t,p}$ the original one.

As an averaged error, the RMSE does not sufficiently represent the global climate evolution. Thus, we also base our validation on coupled ocean-atmosphere phenomenons inducing anomalies in the sea surface temperature from inter-annual to decadal timescales. We investigate three global variability modes reconstructions based on sea surface temperature anomalies in the Pacific and Atlantic Ocean [10], and by calculating the Pearson correlation between the target and reconstructed indices for:

1) The Atlantic Multidecadal Oscillation (AMO) in the north Atlantic Ocean [11].

2) The El Niño-Southern Oscillation (ENSO) in Pacific Ocean which is the main climate mode at inter annual timescale [12].
3) The Interdecadal Pacific Oscillation (IPO) in the Pacific Ocean [13].

To better indicate the reconstructions accuracy, in addition to their correlations, we include the averaged error computed between reconstructed and target climate indices.

III. RESULTS

A. Reduced model

Applying a SOM, we create a regionalization based on the temporal correlations between temperature anomalies evolution over the 150 years of the training data set. From 9216 time series of 1800 time steps each, we obtain a spatial encoding of 191 regions (clusters) representing the 9216 input grid points, as shown in 1.

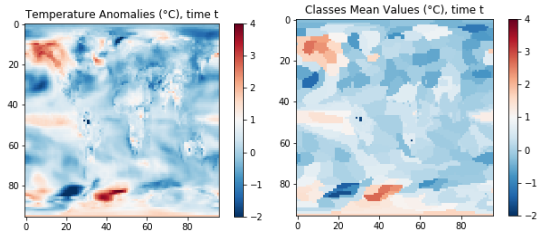
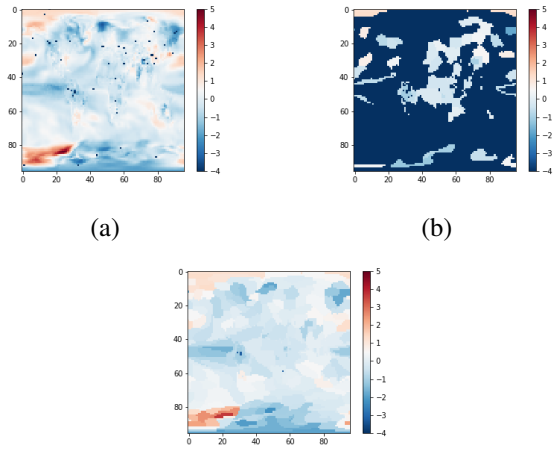


Fig. 1: Reduced model regionalization. Each regions' grid point is assigned the mean value of its spatial cluster.

When evaluating this clustering on the test set, we obtain a RMSE = 0.360°C and over 0.98 correlation between original and reconstructed indices time series. 95% of the reconstructed indices have an absolute error less than 0.05°C for AMO and 0.09°C for both ENSO and IPO. The maximum absolute errors are respectively 0.09°C, 0.12°C, 0.16°C. In III-B, we use this reduced climate model to reconstruct temperature anomalies maps from spatially sparse temperature time series.

B. Surface temperature anomalies reconstruction

To reconstruct climate anomalies beyond the 150 years simulating the observations period, we first train a model that reconstruct the unknown values of the regions obtained in III-A, using the available data and ItCompSOM. We then refine the reconstruction, using the temporal relationship between the sampled anomalies time series and the reduced model clusters, learnt by a RNN.



(a) (b)

(c)

Fig. 2: Clusters reconstruction from sparse information, at time t. (a) Temperature anomalies (°C) and sampled grid points (represented as blue dots). (b) Clusters mean values (°C) associated to the sampling (deep blue: missing values) (c) ItCompSOM clusters reconstruction (°C).

As discussed in II-A3, we pseudo-randomly select the grid points at which we consider observations of temperature anomalies (2(a)). They are located in the northern hemisphere with a probability of 0.8 and on land with a probability of 0.84. A cluster containing at least one selected grid point is considered to be known (2(b)).

To reconstruct the masked clusters through ItCompSOM, only the first 150 years are supposed completely known. The clustering of the SOMs used by ItCompSOM in this application has the 191 regions (III-A) as variables.

We evaluate the ItCompSOM reconstructions on the test set with declining number of known clusters as input, as shown in I, the Obs. referring to the % of observed grid points.

The strength of our method lies in the fact that we can greatly decrease the number of sampled grid points while preserving information, thanks to the implicit temporal coherence of the spatial clusters learnt in III-A. While 0.5% of available data represents 43 grid points on the globe, if well distributed, we can suppose that we have knowledge for 43 values among the 191 regions and thus 22% of global information (in terms of regions).

The method's performance becomes unreliable when sampling only 0.05% of total information (2% in terms of regions). Even though a correlation to indices over

Obs.	5%	0.5%	0.05%	0.05%
RMSE	0.370°C	0.383°C	0.420°C	0.397°C
AMO	0.98 (0.02)	0.97 (0.04)	0.89 (0.09)	0.92 (0.07)
ENSO	0.99 (0.05)	0.99 (0.07)	0.96 (0.15)	0.97 (0.12)
IPO	0.99 (0.05)	0.99 (0.06)	0.97 (0.13)	0.98 (0.11)

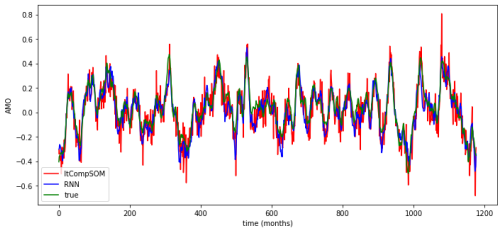
TABLE I: Error computation after completion with ItCompSOM (three firsts columns) and improvement by RNN (last column). The values correspond to the correlations between the reconstructed and target indices and the associated averaged error between parenthesis.

0.89 seems significant, time series are no longer efficiently reconstructed. Whereas the RMSE over climate indices reaches 0.05°C for IPO indice with 5% input data, it is worth 0.13°C with 0.05% available data. It is noticeable that the correlation between AMO indices decreases first. This is due to the computation of AMO over the entire north Atlantic Ocean which represents more zones of the reduced model than the two others.

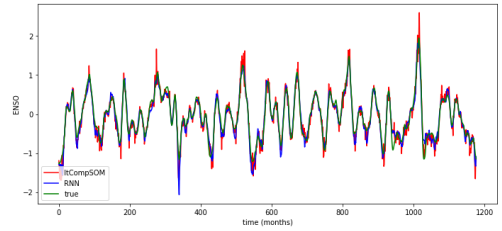
We enhance this reconstruction using a RNN. To explicitly use the temporal dimension of climate evolution, a model is trained on the 191 clusters values over a 150 years-long period. The training data set is composed of the reconstruction with ItCompSOM from 0.05% input data, using as label data the regionalization computed in III-A. This way, we improve our results, as figured on 3, to reach over 92% of correlation to the climate indices and reducing the RMSE to less than 0.4 °C. As shown in 3, the RNN is particularly helpful to lessen errors made on peak cycles. Those peaks are important when analyzing global climate as, for instance, extreme ENSO values can indicate extreme weather events such as droughts, floodings and tropical storms [12].

IV. Conclusion

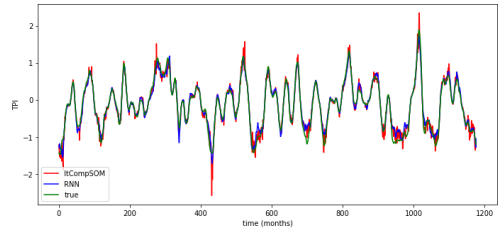
In this paper we show the capacity of machine learning to reconstruct global temperature anomalies evolution given a sparse observation field. Our methodology, combining iterative self-organizing map completions and a recurrent neural network correction of the temporal trajectory, is applied to reconstruct the surface temperature anomaly fields from 0.05% of total number of data points. We obtain an RMSE of 0.398°C. We further validate the quality of the results calculating a correlation of 0.92, 0.97 and 0.98 between the reconstructed and target indices of AMO, ENSO and IPO. These results are highly encouraging and open the way to apply the developed methodology to real proxies records and deal with new issues such as spatio-temporal uncertainties.



(a) AMO time series (°C).



(b) ENSO time series (°C).



(c) IPO time series (°C).

Fig. 3: Reconstruction evaluation. The green curve depicts the indices computed on the original data set, the red curve the indices computed on the reconstruction through ItCompSOM and the blue one the indices reconstruction enhanced by RNN.

Acknowledgments

Funding for the authors was provided by CNES.

References

[1] M. Khodri, D. Swingedouw, J. Mignot, M.-A. Sicre, E. Garnier, V. Masson-Delmotte, A. Ribes, and L. Terray, "Le climat du dernier millénaire," *La Météorologie*, 2015.

[2] D. L. Hartmann, *Global physical climatology*, vol. 103. Newnes, 2015.

[3] T. C. Lee, M. Tsao, and F. W. Zwiers, "State-space model for proxy-based millennial reconstruction," *Canadian Journal of Statistics*, vol. 38, no. 3, pp. 488–505, 2010.

[4] J. P. Main, *Seasonality of circulation in southern Africa using the Kohonen self-organising map*. PhD thesis, University of Cape Town, 1997.

[5] A. Caubel and P. Sepulchre, "IPSL-CM5A2-VLR configuration." http://forge.ipsl.jussieu.fr/igcmg_doc/wiki/DocHconfigAipslcm5a2, 2018. [Online; accessed 02-July-2019].

[6] B. Hewitson and R. Crane, "Self-organizing maps: applications to synoptic climatology," *Climate Research*, vol. 22, no. 1, pp. 13–26, 2002.
[7] A. A. Charantonis, P. Testor, L. Mortier, F. Dortenzio, and S. Thiria, "Completion of a sparse glider database using multi-iterative self-organizing maps (itcomp som)," *Procedia Computer Science*, vol. 51, pp. 2198–2206, 2015.
[8] C. Chapman and A. A. Charantonis, "Reconstruction of subsurface velocities from satellite observations using iterative self-organizing maps," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 5, pp. 617–620, 2017.
[9] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.
[10] C. Deser, M. A. Alexander, S.-P. Xie, and A. S. Phillips, "Sea surface temperature variability: Patterns and mechanisms," *Annual review of marine science*, vol. 2, pp. 115–143, 2010.
[11] T. L. Delworth and M. E. Mann, "Observed and simulated multidecadal variability in the northern hemisphere," *Climate Dynamics*, vol. 16, no. 9, pp. 661–676, 2000.
[12] E. Guilyardi, A. Wittenberg, A. Fedorov, M. Collins, C. Wang, A. Capotondi, G. J. Van Oldenborgh, and T. Stockdale, "Understanding el niño in ocean–atmosphere general circulation models: Progress and challenges," *Bulletin of the American Meteorological Society*, vol. 90, no. 3, pp. 325–340, 2009.
[13] B. J. Henley, J. Gergis, D. J. Karoly, S. Power, J. Kennedy, and C. K. Folland, "A tripole index for the interdecadal pacific oscillation," *Climate Dynamics*, vol. 45, no. 11-12, pp. 3077–3090, 2015.

CAUSAL LINK ESTIMATION UNDER HIDDEN CONFOUNDING IN ECOLOGICAL TIME SERIES

Violeta Teodora Trifunov[1,2], Maha Shadaydeh[1], Jakob Runge[2], Veronika Eyring[4,5], Markus Reichstein[3,6], Joachim Denzler[1,3]

***Abstract*—Understanding the causes of natural phenomena is a subject of continuous interest in many research fields such as climate and environmental science. We address the problem of recovering nonlinear causal relationships between time series of ecological variables in the presence of a hidden confounder. We suggest a deep learning approach with domain knowledge integration based on the Causal Effect Variational Autoencoder (CEVAE) which we extend and apply to ecological time series. We compare our method's performance to that of vector autoregressive Granger Causality (VAR-GC) to emphasize its benefits.**

I. INTRODUCTION

Many research fields such as climate and environmental sciences [1], [2] are continuously striving to understand the causes of natural phenomena. The complex nature and the continuously changing climate system both contribute to the slow advances in this field. This issue was shown to be amenable through the development of data-driven methodologies that are guided by theory to produce more accurate models [3]. We propose to mitigate the problem of recovering nonlinear causal relationships between time series of ecological variables in the presence of a hidden confounder. We suggest a deep learning approach with domain knowledge integration in the form of the ground-truth causal graph, shown in Fig. 1, for mending this issue. Our approach is based on the Causal Effect Variational Autoencoder (CEVAE) [4] which we extend by modelling an intervention for time series of confounded ecological variables. We compare our method's performance to that of the vector autoregressive Granger Causality (VAR-GC) [5], [6] and find that our approach is indeed capable of recovering nonlinear causal relationships under hidden confounding in contrast to VAR-GC baseline. In the study of ecosystems for example, considering confounders is important when trying to determine a causal link between the air temperature (T_{air}) and the ecosystem respiration (R_{eco}). Since both of these variables are influenced by the global radiation (R_g), one cannot know with certainty that the causal link between T_{air} and R_{eco} is not affected by R_g. Therefore, not taking confounding into consideration may lead to erroneous conclusions. Two variables, W and Y, are said to be *confounded* if there exists another variable Z that is a cause for both W and Y. In order to confirm if the confounder is influencing the link between W and Y, one needs to intervene on W in the sense of *do*-Calculus [7] and thereby remove any influence of Z on W. If the intervention on W induces no change in the outcome Y, it is evident that the causal link between W and Y is solely influenced by the hidden confounder Z itself. In the case of an observed confounder, a conventional approach for accounting for its effect is to "control" for it. This is done for instance

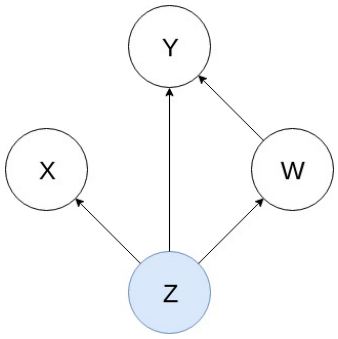


Fig. 1. Graphical model portraying hidden confounding with one proxy. *Y* denotes an outcome, *W* an intervention variable, *Z* an unobserved confounder and *X* denotes a proxy variable providing noisy views on the hidden confounder *Z*.

Corresponding author: Violeta Teodora Trifunov, violetateodora.trifunov@uni-jena.de [1]Computer Vision Group, Friedrich Schiller University Jena, Jena, Germany [2]Climate Informatics Group, Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institute for Data Science, Jena, Germany [3]Michael Stifel Center Jena for Data-Driven and Simulation Science, Jena, Germany [4]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany [5] University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany [6]Max Planck Institute for Biogeochemistry, Jena, Germany

by covariate-adjusted regression or propensity score regression [8]. However, if a confounder is hidden, it is impossible to estimate the effect of the intervention on the outcome without making further assumptions [7]. Figure 1 depicts a form of this problem when there is a single proxy variable. A proxy is an observed variable that describes the unobserved confounder and it is then used in causal link estimation between the confounded variables instead of the hidden confounder itself. For the cases of more general proxy models and the conditions under which they can be identified, please refer to [9].

II. Related work

A standard causality analysis method is Granger Causality (GC) [5] applied in the setting of no hidden confounding [10]. The main assumption of this concept is that causes always come before their effects in time. This means that if one time series causes another series, knowing the former series should be helpful for predicting future values of the latter series after influences of all other variables have been considered. Ecological variables often contain trends or periodic components such as diurnal or seasonal cycles which act as a hidden confounder. In [11], [12] authors have used parametric spectral representation for inferring the cause-effect relationships between ecological variables, assuming no hidden confounder. They have shown that time domain causality analysis of ecosystem variables based on VAR-GC [6], may result in spurious causal links due to the above-mentioned periodic components of ecological variables and thus proposed to use the parametric frequency domain representation of VAR-GC instead. In [13] it was further shown using a deep learning approach for causal inference based on a Causal Effect Variational Autoencoder (CEVAE) [4], that cause-effect analysis can be done in the presence of a periodic component acting as a hidden confounder in the time domain.

In regards to other deep learning methods for causal inference, several approaches have been suggested recently. One such is applied to inference of interactions between variables while learning the dynamics in an unsupervised manner [14]. Moreover, Causal Effect Network (CEN) [15] has been proposed for assessing causal relationships of time series, as well as their time delays between different processes. However, most causal inference methods cannot be applied if hidden confounders are present [2]. Another research branch dealing with the modelling of the latent variable space using deep graphical models was introduced in the recent years. Autoencoders are a class of deep learning methods that can be combined with directed probabilistic graphical models for efficient inference in the presence of continuous latent variables with intractable posterior distributions, such as the Variational Autoencoder (VAE) [16]. Moreover, VAE represents the fundamental building block of the CEVAE [4], which allows for the estimation of the unknown latent space and inference of causal links between the confounded variables. Our work extends the capabilities of the CEVAE to time series that are based on real observations of global radiation and provides a comparison to VAR-GC.

III. Methodology

In the first part of this section, we describe the main deep graphical model our method relies on. Then, we explain how our method builds upon that graphical model. Finally, we provide a brief introduction to the VAR-GC method, which we use as a baseline.

A. *Causal effect variational autoencoder*

Based on the VAE [16] and the TARnet [17] generative model structure, the CEVAE [4] is a deep learning method that addresses hidden confounding by estimating the latent space and summarizing the causal effect of discrete or continuous, non-sequential variables. This is accomplished through the use of a noisy proxy related to the confounder, as shown in Fig. 1. In its original application to medical data, W from Fig. 1 denotes treatment, Y an outcome of the treatment, whereas a hidden confounder Z represents the socio-economic status of each patient. Its proxy X represents patient's income for the previous year and a place of residence. The main objective was, therefore, recovering the Individual Treatment Effect (ITE) and the Average Treatment Effect (ATE) defined in (1) and (2), respectively:

$$\begin{aligned} ITE(x) :=& \mathbb{E}(Y|X=x, do(W=w^1)) - \\ & \mathbb{E}(Y|X=x, do(W=w^0)) \end{aligned} \tag{1}$$

$$ATE := \mathbb{E}(ITE(x)) \tag{2}$$

The metrics from Eq. 1 and 2 are defined for each individual value x of variable X, and by w^1 we denote the provided treatment, while w^0 denotes the values of W when no treatment is provided. To obtain the ITE, we need to recover the joint probability $p(Z, X, W, Y)$, as shown by Theorem 1 in [4]. Obtaining this joint distribution is done through a model network of the CEVAE by estimating the true posterior over Z which depends on X, W and Y, where Z is modelled by the standard normal distribution since it is unobserved and

an assumption about its distribution has to be made. The estimate of the posterior is then inferred via TARnet [17] by calculating it for each intervention group in W. It is then possible to construct a single objective for the inference and model networks, i.e. the *variational lower bound*

$$\mathcal{L} = \sum_{i=1}^{N} \mathbb{E}_{q(z_i|x_i,w_i,y_i)} \Big(\log p(z_i) - \log q(z_i|x_i, w_i, y_i) + \log p(x_i, w_i|z_i) + \log p(y_i|w_i, z_i) \Big), \tag{3}$$

of the graphical model from Fig. 1. By x_i we denote an input data point, by w_i each treatment assignment, by y_i the outcome of the specific treatment, by z_i the latent confounder and by q we denote estimates of the true probability distributions p which are computationally intractable. Finally, since it is necessary to know the intervention assignment w along with its outcome y before inferring the posterior distribution over Z, two auxiliary distributions are introduced, helping to predict w_i and y_i for new samples, so the variational lower bound becomes

$$\mathcal{F}_{\text{CEVAE}} = \mathcal{L} + \sum_{i=1}^{N} \Big(\log q(w_i = w_i^*|x_i^*) + \log q(y_i = y_i^*|x_i^*, w_i^*) \Big), \tag{4}$$

where x_i^*, w_i^*, y_i^* are, respectively, the observed values for the input, intervention and outcome variables in the training set.

B. CEVAE for ecological time series

Ecological time series often encompass nonlinearly related variables, as well as influence by a hidden confounder. In contrast to the conventional CEVAE setting, our intervention variable W, as well as variables X, Y and Z, are time series. Doing an intervention in real-world climate data is most often not feasible, so an alternative approach has to be followed. One such approach could be separating the values of an intervention variable W during day and night time or during winter and summer, for instance. As intervention we consider those values of W for which the influence of the confounder Z, the global radiation, is the weakest and thus arguably negligible. We provide more detail on modelling the intervention in Section IV. We also adjust several probability distributions to accommodate our problem setting. Namely, we model a conditional distribution of W given Z as follows:

$$p(W|Z) = \mathcal{N}(\mu_w, \sigma_w^2), \quad [\ \mu_w, \sigma_w] \ = f_1(Z). \tag{5}$$

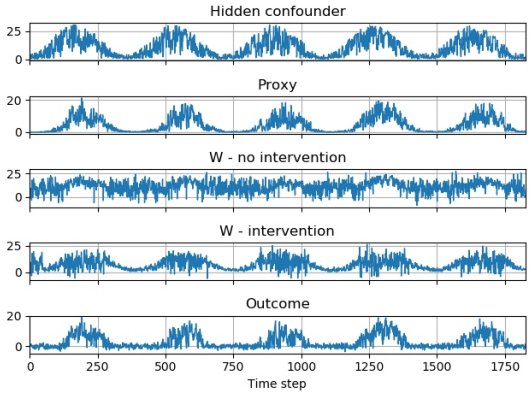


Fig. 2. Synthetic data. The first row shows real observation of global radiation $R_{g_{obs}}$ as the hidden confounder Z; the second row shows the noisy GPP as the proxy X for $c = 0.03196$, $e = 0.3$, $\tau_3 = 23$ and $\tau_4 = 14$; the third row shows the variable T_{air} as W for $a = 0.4632$, $b = 0.48078$, $\tau_1 = 22$ and $\tau_2 = 17$ without intervention; the fourth row shows variable W with intervention; the fifth row shows R_{eco} as Y for $e = 0.198087$, $\alpha = 0.61078$, $\tau_5 = 21$ and $\tau_6 = 21$.

Estimation of this distribution is obtained through the use of the proxy X:

$$q(W|X) = \mathcal{N}(\hat{\mu}_w, \hat{\sigma}_w^2), \quad [\ \hat{\mu}_w, \hat{\sigma}_w] \ = f_2(X). \tag{6}$$

Functions f_1 and f_2 are feedforward neural networks with three layers. To measure the intervention effect of W on Y, we extend ITE (Eq. 1) to the case of a sequential intervention and define the Interval Intervention Effect (IIE) and the Average Intervention Effect (AIE):

$$\text{IIE}(I_i) := \mathbb{E}(Y|X \in I_i, do(W = w^1)) - \mathbb{E}(Y|X \in I_i, do(W = w^0)) \tag{7}$$

$$\text{AIE} := \ \mathbb{E}(\text{IIE}(I_i)_{i=0,\dots,m-1}) \tag{8}$$

By interval $I_i = [x_i, x_{i+1}]$ we denote the i-th uniform quantization level of X, with x_i and x_{i+1} being its limits, for $i = 0, ..., m-1$ with $m = 256$. In Eq. (8), by $\text{IIE}(I_i)_{i=0,\dots,m-1}$ we denote an m-dimensional vector whose elements are $\text{IIE}(I_i)$ for each $i = 0, \dots, m-1$. In this manner we have extended the CEVAE to the setting of a continuous intervention variable, since intervals of W and Y are taken into account for calculating the causal effect between them, instead of the individual values, as seen in Eq. (1) and Eq. (2).

C. Vector autoregressive Granger causality

The main assumption of Granger causality (GC) [5] is that causes precede their effects and can be used for their prediction. Let $u_i, i = 1, \cdots, N$ be the time series of N ecological variables. Each time series $u_i(t), t =$

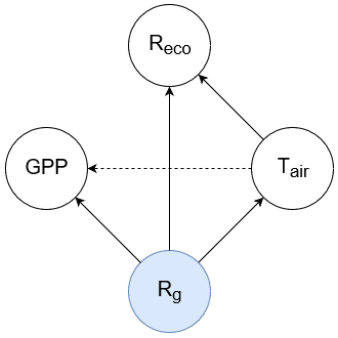


Fig. 3. Causal graphical model portraying causal relationships between ecological variables defined in equations (11)-(14). Respiration of the ecosystem R_{eco} denotes the outcome, air temperature T_{air} denotes an intervention variable, the global radiation R_g is assumed to be the hidden confounder and GPP denotes a proxy variable providing noisy views on the unobserved confounder R_g.

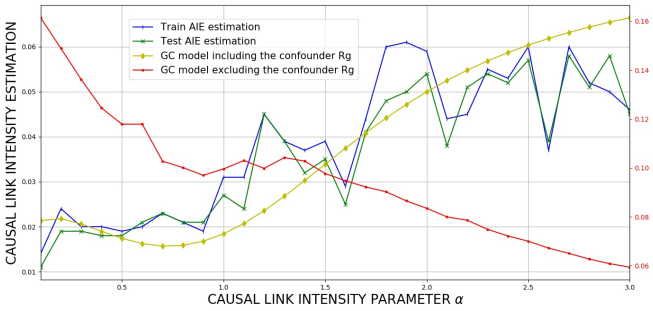


Fig. 4. Causal link estimation results of our method in comparison to the vector autoregressive Granger causality (VAR-GC). Blue and green curves show our methods estimation of the AIE during training and test, respectively. The yellow curve represents the VAR-GC method's estimate $\gamma_{T_{air}\to R_{eco}}$ when the confounding variable R_g is included along with T_{air}, R_{eco} and GPP. The red curve represents the VAR-GC method's estimate $\gamma_{T_{air}\to R_{eco}}$ when the proxy variable GPP is used along with T_{air} and R_{eco}, instead of the confounder itself. We note that in real ecological data $\alpha < 1$.

$1, \cdots, k$ is a realization of length k real valued discrete stationary stochastic process $U_i, i = 1, \cdots, N$. These N time series can be represented by a pth order vector autoregressive model (VAR(p)) of the form

$$\begin{bmatrix} u_1(t) \\ \vdots \\ u_N(t) \end{bmatrix} = \sum_{r=1}^{p} A_r \begin{bmatrix} u_1(t-r) \\ \vdots \\ u_N(t-r) \end{bmatrix} + \begin{bmatrix} \epsilon_1(t) \\ \vdots \\ \epsilon_N(t) \end{bmatrix}. \tag{9}$$

The residuals $\epsilon_i, i = 1, \cdots, N$ form a white noise stationary process with covariance matrix Σ. The model parameters at time lags $r = 1, \cdots, p$ comprise the matrix $A_r = [a_{ij}(r)]_{N\times N}$. Let Σ_j be the covariance matrix of the residual ϵ_j associated to u_j using the model in (9), and let Σ_j^{i-} denote the covariance matrix of this residual after excluding the ith raw and column in A_r. The time domain VAR-GC of u_i on u_j conditioned on all other variables is defined by [6]

$$\gamma_{i\to j} = \ln \frac{|\Sigma_j^{i-}|}{|\Sigma_j|}. \tag{10}$$

IV. Experimental results

We have applied the proposed method to a synthetic data generated from 1825 real observations of the global radiation ($R_{g_{obs}}$) measured at the flux tower in Heinich National Park - Germany, over the period of five years as suggested by [18]:

$$R_g(t) = R_{g_{obs}}(t) \tag{11}$$

$$T_{air}(t) = a \cdot T_{air}(t-\tau_1) + b \cdot R_g(t-\tau_2) + \eta_1(t) \tag{12}$$

$$GPP(t) = c \cdot R_g(t-\tau_3) \cdot T_{air}(t-\tau_4) + \eta_2(t) \tag{13}$$

$$R_{eco}(t) = e \cdot R_g(t-\tau_5) \cdot \alpha^{\frac{T_{air}(t-\tau_6)}{20}} + \eta_3(t). \tag{14}$$

In equations (11)-(14), by η_1, η_2, η_3 we denote mutually uncorrelated Gaussian noise with mean $\mu = 0$ and variance $\sigma = 1$, by $\tau_i \in \mathbb{N}$ for $i = 1, \ldots, 6$ we denote time lags, and by $a, b, c, e \in (0, 1)$ we denote constants. The plots of all above-mentioned variables along with the exact parameters used is shown in Fig. 2. The causal link intensity parameter is denoted by $\alpha \in \mathbb{R}$ and it represents the value of our highest interest. Namely, throughout our experiments, we increase the parameter α and observe the AIE estimation results in accordance to this increase. The portion of data we consider spans from April to September of each year, which results in 900 samples of real global radiation observations, corresponding to the hidden confounder Z. Synthetically generated variables T_{air}, GPP and R_{eco}, correspond to W, X and Y, respectively, as shown in Fig. 3. We note that even though there is a causal link from T_{air} to GPP, it does not change any of the original problem settings, nor any of the probability distributions involved. During those months the influence of T_{air} to R_{eco} should be more pronounced as the temperatures are higher. We model the intervention by considering a certain threshold of the proxy mean instead of the hidden confounder. More precisely, we consider values of the intervention variable T_{air} corresponding to the daily values of the proxy GPP that are smaller than the threshold of 0.7 of its mean. To create w^1 from Eq. (7), we use values of W corresponding to the proxy values smaller than the said threshold. This alone would lead to w^1 having missing values for time steps that correspond to the proxy values greater than the threshold. In order to prevent that, we concatenate all

present values and replicate them successively until the sample size is reached. We define w^0 from Eq. (7) in an analogue fashion. This type of intervention was chosen to simulate the properties of *do*-calculus as closely as possible. Moreover, it allows for a more easily adaptable application of our method to real data. Namely, after intervention, causal link between T_{air} and its parent R_g should either be removed or so small, that it can be neglected. This naturally occurs during winter or night time.

As far as the neural network architecture is concerned, we closely followed [4]. We used feedforward neural networks, namely f_1 and f_2 with 3 hidden layers and the ELU [19] nonlinearity. We note, however, that more hidden layers as well as different types of networks can be used. We modelled variable Z as normally distributed with 20 dimensions, due to its latency. We used a small weight decay term for all parameters with $\lambda = 0.0001$. For optimization, Adamax [20] was utilized with a learning rate of 0.01. Furthermore, early stopping according to the lower bound on a validation set was performed. For obtaining the outcomes $p(y|x_i \leq X \leq x_{i+1}, do(W = w^1))$ and $p(y|x_i \leq X \leq x_{i+1}, do(W = w^0))$ we averaged over 100 samples from the approximate posterior $q(Z|X) = \sum_w \int q(Z|w, y, X)q(y|w, X)q(w|X)dy$. As domain knowledge in this work, we consider the ground truth causal relationships between time series of ecological variables which are used to verify our method's results. When the causal graph is not partly or entirely known, methods like [1] should be used for obtaining it to a greater extent. Our goal is to estimate the causal link intensity between T_{air} and R_{eco} in the presence of the unobserved confounder R_g. We do so by running our method for different values of parameter α from Eq. (14), function of which is proportional to the causal link intensity between the variables in question. The results for each value of α from 0.1 to 3 are obtained as the average of the outputs of ten different realizations of the data. In real ecological time series, usually $\alpha < 1$. We note that the increase of α, yields the increase of the estimate of the causal link strength measured by the absolute value of AIE, as shown by the blue and green curves in Fig. 4. We also note that the curves describing the relation between the causal link intensity parameter α and the AIE estimation during training and testing is nonlinearly increasing. This means we are able to estimate the nonlinear causal relationship between T_{air} and R_{eco} from Eq. (14) under hidden confounding. Furthermore, for the purpose of a fair comparison, we applied the VAR-GC method to all four variables $u_1 = T_{air}$, $u_2 = R_{eco}$, $u_3 = \text{GPP}$ and $u_4 = R_g$ from Eq. (9), over the entire 1825 data samples. More specifically, we included the otherwise hidden confounder R_g. This way, we could also reproduce the nonlinear causal link between T_{air} and R_{eco}, as seen in Fig. 4. To test if VAR-GC can detect the increase of the nonlinear causal link's intensity between T_{air} and R_{eco} without using R_g, we omitted it and observed as shown by a decreasing red curve in Fig. 4 that the desired causal link intensity increase could not be detected.

V. CONCLUSION AND FUTURE WORK

The goal of this work was to infer nonlinear causal relationships between time series of ecological variables in the presence of a hidden confounder using an extended version of the CEVAE. We provided a comparison to the baseline VAR-GC method with and without the use of a hidden confounder and concluded that our method is much more suitable for the task when the confounder is unobserved. This is since we use the proxy in estimating the AIE metric instead of the confounder, influence of which is removed after the intervention on T_{air}. We intend to enhance our method by doing a quantitative evaluation of the said comparison and incorporating recurrent neural networks into the current architecture.

REFERENCES

[1] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018.

[2] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, and C. Glymour, "Inferring causation from time series of earth system sciences," *Nature Communications*, vol. 10, p. 2553, 2019.

[3] J. H. Fahmous and V. Kumar, "A big data guide to understanding climate change: The case fortheory-guided data science," in *Big Data*, vol. 2, pp. 155–163, 2014.

[4] C. Louizos, U. Shalit, J. Mooij, Z. Sontag, D., R., and M. Welling, "Causal effect inference with deep latent-variable models," in *Advances in Neural Information Processing Systems 30*, pp. 6446–6456, 2017.

[5] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica - Journal of the Econometric Society*, vol. 37, no. 3, pp. 424–438, 1969.

[6] J. Geweke, "Measurement of linear dependence and feedback between multiple time series," in *Journal of the American statistical association*, vol. 77, pp. 304–313, 1982.

[7] J. Pearl, *Causality*. Cambridge University Press, 2009.

[8] L. Li, K. Kleinman, and M. W. Gillman, "A comparison of confounding adjustment methods with an application to early life determinants of childhood obesity," *Journal of developmental origins of health and disease*, vol. 5, no. 6, pp. 435–447, 2014.

[9] W. Miao, Z. Geng, and E. Tchetgen Tchetgen, "Identifying causal effects with proxy variables of an unmeasured confounder," in *arXiv preprint arXiv:1609.08816*, 2016.
[10] M. Eichler, *Causal inference in time series analysis*, pp. 327–354. Wiley Series in Probability and Statistics, United States: John Wiley & Sons, 2012.
[11] M. Shadaydeh, Y. Guanche, M. Mahecha, M. Reichstein, and J. Denzler, "Causality analysis of ecological time series: a time-frequency approach," in *Climate Informatics Workshop 2018*, 2018.
[12] M. Shadaydeh, J. Denzler, Y. Guanche, and M. Mahecha, "Time-frequency causal inference uncovers anomalous events in environmental systems," in *German Conference on Pattern Recognition (GCPR)*, 2019.
[13] V. T. Trifunov, M. Shadaydeh, J. Runge, V. Eyring, M. Reichstein, and J. Denzler, "Nonlinear causal link estimation under hidden confounding with an application to time series anomaly detection," in *German Conference on Pattern Recognition (GCPR)*, 2019.
[14] T. Kipf, E. Fetaya, K.-C. Wang, M. Welling, and R. Zemel, "Neural relational inference for interacting systems," in *Internvational Conference on Machine Learning 2018 (ICML), arXiv:1802.04687v2 [stat.ML]*, 2018.
[15] M. Kretschmer, D. Coumou, J. F. Donges, and J. Runge, "Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation," *Journal of Climate*, vol. 29, pp. 4069–4081, 2016.
[16] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proceedings of the 2nd International Conference on Learning Representations (ICLR), arXiv: 1312.6114 [stat. ML]*, 2014.
[17] U. Shalit, F. Johansson, and D. Sontag, "Estimating individual treatment effect: generalization bounds and algorithms," in *arXiv:1606.03976v5 [stat.ML]*, 2016.
[18] C. Krich, M. Mahecha, J. Runge, M. Reichstein, and D. G. Miralles, "Revealing causal dependencies between land-surface fluxes and meteorological variables," *Geophysical Research Abstracts*, vol. 20, 2018.
[19] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv:1511.07289 [cs.LG]*, 2016.
[20] D. P. Kingma and J. B. Adam, "A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

CAUSALITY ANALYSIS IN CLIMATE TIME SERIES USING WINDOWED REGRESSION

Ali Gorji[1], Mohammad Gorji[2]

***Abstract*—An important problem in climate studies is to use climate model simulations and observations to find evidence for effect of natural and anthropogenic forces on climate change. Granger Causality is a common data-driven regression-based technique to test the hypothesis of causality between time series. In this work, we have utilized the windowed Granger analysis which is an extended method of regression to reveal the causality in multiple time series and different time lags. The results show the good performance of the approach on synthetic data. Finally, we have tested the method with real climate time series.**

I. MOTIVATION

An important topic of research in climate informatics is to investigate the causes of climate change and variability. One common method is to use numerical methods and physical equations in the Global Climate Models (GCMs). However, these dynamical models are complex and it is difficult to extract cause and effect relationship from them. In addition, the climate models do not include all the processes in the climate system, which leads to a higher uncertainty of inference. In contrast, data-driven approaches can build some statistical models which can be used for assessing the cause and effect relationships between the external/internal forcings of the climate system and the climate variables [1], [2], [3].

A common technique for analyzing causality in time series is the *Granger Causality (GC)* which was essentially developed for econometric studies, and also common in climate data science [4], [5]. Generally speaking, a cause occurs prior to its effect. Formally, For two time series x and y generated from a system, it is defined that x Granger causes y if future values of y can be better predicted using the past values of x and y rather than only the past values of y. To test for causality, a regression model is fitted between the time series and the corresponding coefficients about the $x \longrightarrow y$ interaction are statistically tested. The GC is essentially defined using bi-variate time series with linear models, but also extended to multivariate time series or nonlinear models.

One challenge in GC analysis is that the history length of time series are necessary for modeling and should be known. Often, statistical methods e.g. Akaike Information Criterion are being used to estimate the lag. However, the cause-effect interaction among two entities might happen in multiple delays. The Granger analysis based on sliding window regression developed by [6] reveals the causal network in multiple time series. It also considers multiple time delays and the effects of several causes at multiple delays on a target are detected. The basic idea of Windowed Granger method is that a sliding windows creates many sensitive, but noisy, inference models which are aggregated finally into a more stable and accurate result.This method is a framework that can use various linear or nonlinear regression methods. It was empirically shown that using windowed approach has better performance in recovering the causal links than applying the regression model to all the length of the multiple time series.

In [6], the Windowed Granger method was tested on multiple biological networks with LASSO, random forest and partial least square. In this work, we have extended the method to also use the linear regression with Wald and Likelihood ratio tests. The p-values of linear regression provides a metric that show how elements of a multiple time series are related to each other at different time lags. We tested the method on a synthetic nonlinear model and real climate data and compare our results to other studies on the same data.

II. METHOD

A. Granger Causal Network

The main concept of Granger causality test is to predict the future values of a time series from past values of another time series. For example consider two univariate stationary time series x , y. If there

Corresponding author: A Gorji aligorjis@gmail.com [1] Department of Computer Engineering, Guilan University, Rasht, Iran
Syntelli Solutions, Charlotte, USA

is a statistically significant improvement of prediction between restricted linear model (Eq. 1) and unrestricted linear model (Eq. 2), x is GC of y. In the unrestricted model, y is only predicted by its past values. In contrast in restricted model, y is predicted by its past value and past values of x. It is common to apply t-test on coefficients c_i, $i \in [1, s]$ or likelihood ratio test on variance of residuals for restricted/unrestricted model to reject the null hypothesis of y does not GC x.

$$\begin{aligned} y_t = \beta + a_1 y_{t-1} + a_2 y_{t-2} + ... + a_s y_{t-s} \\ + c_1 x_{t-1} + c_2 x_{t-2} + ... + c_p x_{t-p} + \epsilon(t) \end{aligned} \quad (1)$$

$$y_t = \gamma + b_1 y_{t-1} + b_2 y_{t-2} + ... + a_s y_{t-s} + \epsilon(t) \quad (2)$$

The linear Granger causality analysis is also extended to P-dimensional multivariate time series represented where a linear *Vector Autoregressive Model (VAR)* is used for modeling purpose.

$$X(t) = \sum_{l=1}^{L} A^{(l)T} X(t-l) + \epsilon(t) \quad (3)$$

Here, $A^{(l)}$ is the matrix of coefficients that models the effect of the time series with l lags. Usually the minimum value for l is 1, however in the case of *Instantaneous Causation*, $l = 0$. The instantaneous causation occurs when the length of measurement interval is too long so that the cause/effect action happened in shorter time than the length of a single time interval. If all of the l VAR coefficient matrices are lower triangular, then x fails to GC y. This can be tested using Wald hypothesis test [7]. While GC was first introduced for linear regression models, but there are nonlinear extensions using Kernel methods or nonparametric regression.

B. Windowed Granger Analysis

The method of Sliding Windowed Regression analysis has been developed by [6] to discover Granger causality. This method provides a framework to extend the regular multivariate regression analysis to model the relationship between time series at multiple delays. In regular Granger analysis, a regression model is fitted to total samples of a multiple time series $X(t) = [x_1(t), ..., x_N(t)]$. Instead in windowed approach, a sliding window traverses through the $X(t)$ creating several overlapping time windows $\boldsymbol{W} = \{W_1, ..., W_Q\}$ (Fig. 1). Here, the number of windows Q is related to length of each window w as $Q = (T - w + 1)$. A regression model is fitted to samples of these windows where one dimension of a windowed X is considered as the target and multiple lagged-values of windowed X are the the predictors (Fig. 2). Let the minimum and maximum delays for modeling the time series are τ_{min} and τ_{max}. The fitted regression model would have w samples for target and $(N \times L) - 1$ features where $L = \tau_{max} - \tau_{min} + 1$. This regression model provides us a *Confidence Metric* between any two windows of X and these metrics are recorded in a matrix $A_{(N.L)\times(N.L)}$. Each element in A shows the confidence metric of regression between two dimensions of X in various windows and delays. For example the cells of (x_3^2, x_1^4) in Fig. 3 marked with a green triangle shows the confidence of regression model with predictor as second window of $x_3(t)$ and target as fourth window of $x_1(t)$. Many elements of A are empty, as we have assumed a lower and upper bound on the τ values, and there are smaller number of predictors for $\tau < \tau_{max}$. The elements of A for $\tau = 0$ with self-regulation (marked with stars) are not considered. After completing the A, the confidence metrics between any two x_i, $i \in \{1, ..., N\}$ at different lags τ are calculated by taking the average of the related confidence metrics over the windows. Finally, a table is created showing the causal links ranked by their average confidence metrics.

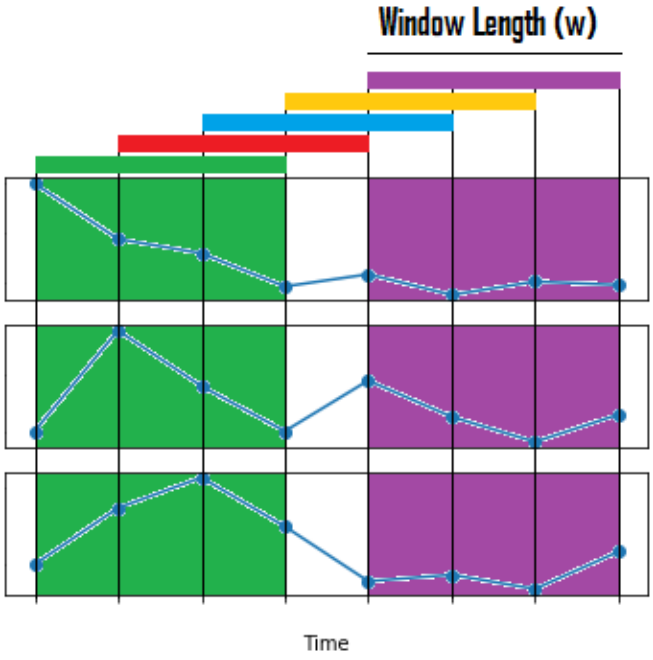


Fig. 1. A sliding window move over time series to create overlapping time partitions

III. CONFIDENCE METRICS

As mentioned above, the Windowed Granger analysis uses a confidence metric to show the strength of link

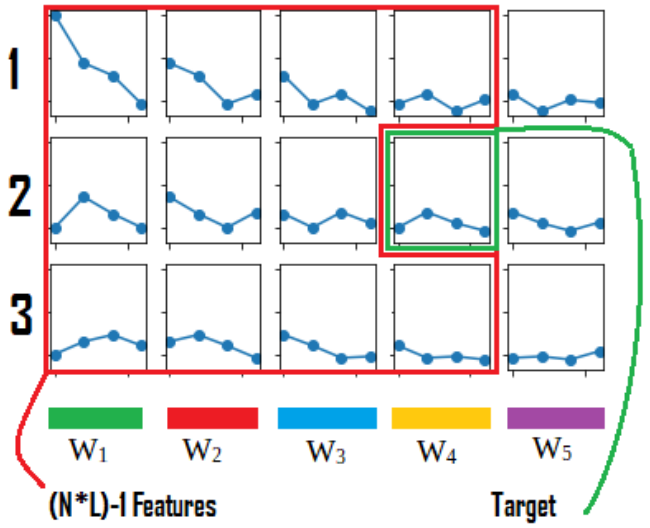


Fig. 2. One dimension of time series in each windows is assumed as target and delayed windows are considered as predictor. The linear regression model is created and p-values of coefficients are extracted

between two windows. Beyond stability of selection in LASSO and importance score in Random Forest utilized in [6], we have added the Wald and Likelihood ratio tests with linear regression to the analysis. The $1 - p$-value of these two tests are considered as confidence metrics. The lower the p-value, the more confidence about the link between the predictor and target. Similarly, the *feature importance* is used for confidence metric in random forest.

Stability of LASSO Estimation: The *Least Absolute Shrinkage and Selection Operator (LASSO)* is a regression method which performs feature selection using L_1 regularization. *Least Angle Regression (LARS)* is an approach to find the path of LASSO based on the regularization parameter. In the LARS algorithm, first all the coefficients are assumed to be zero and then the predictors are added to the model in a direction equiangular to each one's correlations with the residual. The main issues of LARS algorithm is that first it is very sensitive and unstable when in feature selection multicollinear features. Also, it doesn't provide a metric to quantify the amount that a feature has effect on the target variable.

In the method of LASSO stability selection, a confidence metric is found for each predictor in the LASSO model based on number of times this predictor appears in the model during the LARS path [8]. In this method, the training dataset is randomly split to two equal halves and re weighted by a random number $0 < \alpha < 1$. Then the LARS algorithm is applied to this data to find the LASSO path. The features which are selected in the LARS algorithm are recorded. This procedure is

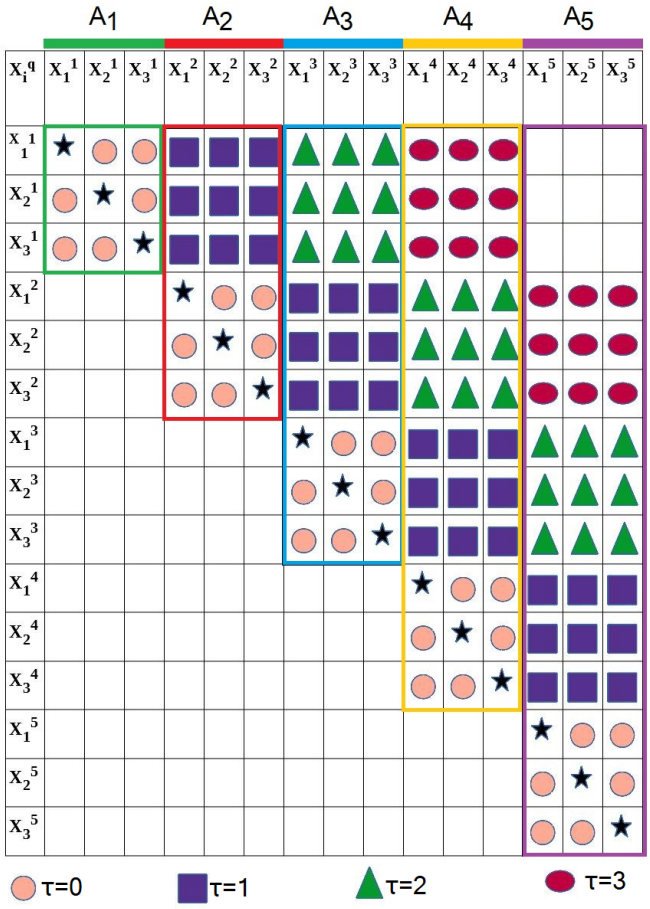


Fig. 3. The confidence metric of regressions associated with various delays or dimensions are placed in the Matrix A. For example, the triangles show the loci of confidence metrics for delay $\tau = 2$ between predictors and targets

repeated M times. Finally the features can be scored based on number of times they were selected of L LARS steps in M number of repetition. Final score of each feature is calculated by *Area Under Curve (AUC)* of the graph of selection frequency.

IV. EVALUATION

In this section we test the Windowed Granger method with four confidence metrics on two multivariate time series.

Experiment 1: In this simulation, we test our method on a time series generated from a example that Rossler system (x_1, x_2, x_3) driving the Lorenz system (y_1, y_2, y_3) [9]:

$$
\begin{aligned}
\dot{x_1} &= -6(x_2 + x_3) \\
\dot{x_2} &= 6(x_1 + 0.2x_2) \\
\dot{x_3} &= 6(0.2 + x_3(x_1 - 5.7)) \\
\dot{y_1} &= 10(-y_1 + y_2) \\
\dot{y_2} &= 28 - y_1 - y_2 - y_1 y_2 + 2x_2^2 \\
\dot{y_3} &= y_1 y_2 - \tfrac{8}{3} y_3
\end{aligned}
$$

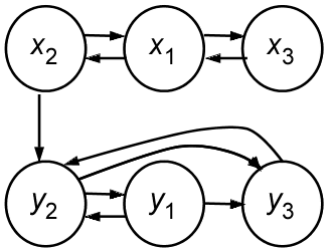


Fig. 4. True network for coupled Rossler-Lorenz system

A total number of 2000 data points were generated then first 1000 were thrown away, and the next 1000 point used to infer the network. The starting point is selected as [0,0,0.4,0.3,0.3,0.3]. The goal is to recover the network of Fig. 4 and especially $x_2 \longrightarrow y_2$ showing the coupling between these two nonlinear subsystems. The parameter of our methods in this simulation was $\tau_{min} = 1$ and $\tau_{max} = 3$, $w = T/2$ where T is the length of time series. By applying the Windowed-GC method, the causal network is recovered and its corresponding adjacency matrix is calculated. Then, the estimated and true adjacency matrices are compared using F1 score metric. For the four regression methods we found that the best F1 score between recovered and true network as LASSO:0.66 , likelihood ratio test: 0.81, Wald test: 0.81 and random forest: 0.77. However, the ranking of the coupling link in these methods are LASSO: 37, likelihood ratio:50, Wald: 51, and random forest: 12. The random forest method has a highest confidence in the coupling link between Lorenz and Rossler subsystems.

Experiment 2: Here, we test the algorithm on a real dataset from Climate domain. Causal relationships among four prominent modes of atmospheric low-frequency variability in boreal winter—namely, the Western Pacific Oscillation (WPO), Eastern Pacific Oscillation (EPO), Pacific–North America (PNA) pattern, and North Atlantic Oscillation (NAO). These modes, also known as *atmospheric tele-connections* are characterized by synchronized low-frequency (longer than typical synoptic time scale of a week) fluctuations in the sea level pressure (SLP) or geo-potential height fields at different geographical locations [10]. The data used here consists of a time series of daily index value for each of the four modes for the period 1 June 1948–1 July 2019 [1]. similar to [10], the analysis was focused on the December–February (DJF) time period with $T = 6467$. We performed a temporal analysis based on daily values of WPO, PNA, EPO, NAO in DJF period. The parameter of our methods in this simulation was $\tau_{min} = 1$, $\tau_{max} = 10$, and $w = T/2$.

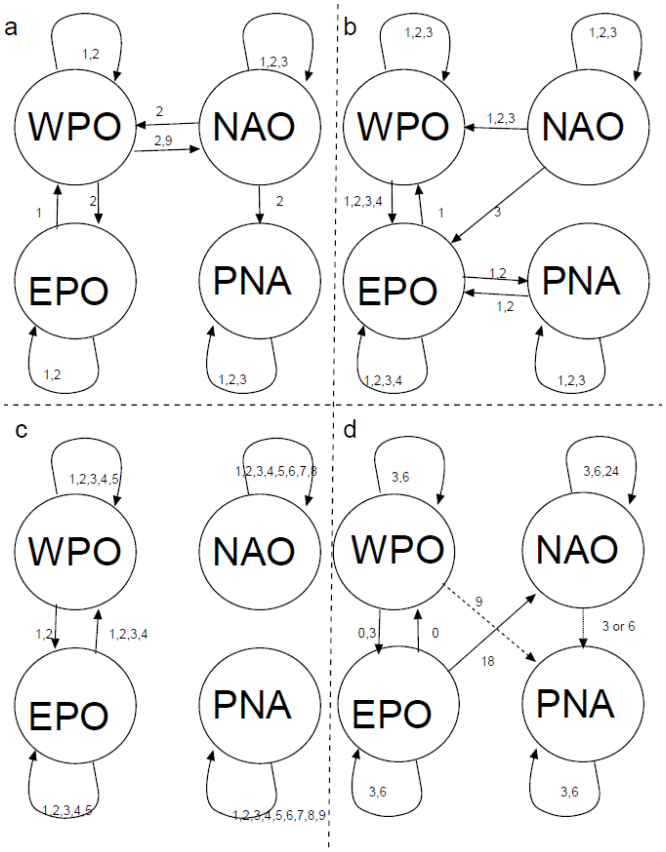


Fig. 5. Four Reconstructed network a) Lasso b) Likelihood ratio and Wald tests c) Random Forest d) network in [10] where the medium confidence link is marked with dashed line. Numbers on arrows represent the corresponding time delays.

In order to select the threshold of selected edges from, we use a method similar to L-curve common in inverse methods. We look for a threshold where the confidence metrics of the regression after that regression has a smaller rate of change. We leave a more quantitative approach of selecting the threshold to our future studies.

Comparing the results of Windowed method with four regression approaches with the [10] (which used the Bayesian Network method) in Fig. 5, we can see the two-way edges between WPO and EPO are recovered in both of the methods. Also the NAO $\longrightarrow$ PNA

[1] https://www.esrl.noaa.gov/psd/forecasts/reforecast2/teleconn/

link is found in LASSO approach. Random forest is the only nonlinear method while Wald , Likelihood ratio test and Lasso are linear.All of our four method discovers self regulation in all four entitiy these are also seen in [10]. There are similarity between all of the model in case of relationship between WPO and EPO, and there is similarity between WPO and NAO in Wald and likelihood ratio and Lasso and we can see reveres relationship between EPO and NAO in Wald and likelihood ratio and graph of [10].

REFERENCES

[1] A. Attanasio, A. Pasini, and U. Triacca, "Granger causality analyses for climatic attribution," *Atmospheric and Climate Sciences*, vol. 3, no. 04, p. 515, 2013.

[2] F. Estrada and P. Perron, "Causality from long-lived radiative forcings to the climate trend," *Annals of the New York Academy of Sciences*, vol. 1436, no. 1, pp. 195–205, 2019.

[3] D. I. Stern and R. K. Kaufmann, "Anthropogenic and natural causes of climate change," *Climatic Change*, vol. 122, pp. 257–269, Nov. 2013.

[4] S. Samarasinghe, M. McGraw, E. Barnes, and I. Ebert-Uphoff, "A study of links between the arctic and the midlatitude jet stream using granger and pearl causality," *Environmetrics*, vol. 30, no. 4, p. e2540, 2019.

[5] T. Gries, M. Redlin, and J. E. Ugarte, "Human-induced climate change: the impact of land-use change," *Theoretical and Applied Climatology*, vol. 135, no. 3-4, pp. 1031–1044, 2019.

[6] J. D. Finkle, J. J. Wu, and N. Bagheri, "Windowed granger causal inference strategy improves discovery of gene regulatory networks," *Proceedings of the National Academy of Sciences*, vol. 115, no. 9, pp. 2252–2257, 2018.

[7] M. H. Pesaran, *Time series and panel data econometrics*. Oxford University Press, 2015.

[8] A.-C. Haury, F. Mordelet, P. Vera-Licona, and J.-P. Vert, "Tigress: Trustful inference of gene regulation using stability selection," *BMC Systems Biology*, vol. 6, p. 145, Nov 2012.

[9] A. Krakovská, J. Jakubík, H. Budáčová, and M. Holecyová, "Causality studied in reconstructed state space. examples of uni-directionally connected chaotic systems," 2015.

[10] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, pp. 5648–5665, Sept. 2012.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

GAUSSIAN MIXTURE MODELING DESCRIBES THE GEOGRAPHY OF THE SURFACE OCEAN CARBON BUDGET

Daniel C. Jones[1], Takamitsu Ito[2]

***Abstract*—We use an unsupervised classification technique (i.e. Gaussian mixture modeling or GMM) to identify ocean regions with similar balances between processes that determine the surface budget of dissolved inorganic carbon. GMM objectively locates sub-populations in the distribution of carbon budget terms. We use a simple four-class description and find regimes that are broadly consistent with classical theoretical frameworks. Class 1 covers 24% of ocean surface area and corresponds to highly productive areas with strong vertical mixing, wind-driven open ocean upwelling, and absorption of atmospheric carbon dioxide. Class 2 covers 8% of ocean surface area and corresponds to regions of especially weak productivity. Class 3 covers 16% of ocean surface area and corresponds to wind-driven coastal and equatorial upwelling. Finally, class 4 covers the remaining 52% of ocean surface area and corresponds to the relatively unproductive subtropical gyres, which are typically characterized by downwelling and low surface nutrient concentrations. We argue that GMM may be a useful method for comparing biogeochemical regimes between climate models.**

I. MOTIVATION

The global ocean is a critical part of Earth's climate system, in part because it absorbs atmospheric carbon dioxide from fossil fuel burning, cement production, and biomass burning, thereby slowing the rate of surface warming. At present, the ocean absorbs between 20 to 35 percent of anthropogenic CO_2 emissions [1], [2], [3]. The ocean's ability to transport carbon from the near-surface ocean into the deep interior, where it is out of contact with the atmosphere, is sometimes referred to as the ocean carbon pump. Broadly speaking, this pump consists of two components: the solubility pump and the biological pump. The solubility pump is a consequence of the global overturning circulation, whereby atmospheric carbon is more readily absorbed by cold, high latitude waters and subducted into the interior ocean via deep convection. The biological pump is a consequence of ocean ecology, by which carbon is transferred from a dissolved inorganic carbon (DIC) pool in the surface ocean to an organic carbon pool. Some fraction of this organic carbon is respired back to DIC throughout the water column, and a small fraction ultimately reaches the seabed. The net result is a vertical transfer of DIC away from the surface ocean into the deep interior, where it is out of contact with the atmosphere and unable to directly affect surface climate.

The processes that govern the surface carbon budget display considerable spatial variability. For example, air-sea gas exchange is highly nonuniform, due to spatial variability in mixed layer depths, near-surface winds, and carbonate chemistry parameters [4], [5]. Biological productivity varies based on the nutrient distribution and other ecological factors [6]. Physical transport, which controls the ocean solulbility pump, is also spatially variable as the ocean's global overturning circulation is set by bathymetry, surface forcing, and internal dynamics, which all have their own spatial patterns. The rate of change of the surface carbon concentration is set by the residual of these processes. Our present understanding of the surface carbon budget relies on classical theoretical frameworks that describe balances between these processes. As a complement to existing expertise-driven approaches, it may be useful to develop a suite of alternative methods by which we can characterize the surface carbon budget. Unsupervised learning may offer such a possibility. In unsupervised learning, one applies a classification algorithm to an unlabeled dataset, and the algorithm attempts to identify sub-populations in the data distribution [7]. To the extent that such methods can be shown to be robust and objective, they could be useful for comparing

Corresponding author: D. Jones, dannes@bas.ac.uk [1]British Antarctic Survey, NERC, UKRI, Cambridge, UK [2]School of Earth and Atmospheric Sciences, Georgia Institute of Technology, Atlanta, GA, USA

biogeochemical regimes in different climate models, which are sometimes difficult to compare directly due to systematic biases [8].

In this note, we apply Gaussian mixture modeling, an unsupervised classification method, to the surface carbon budget derived from a numerical circulation and biogeochemistry model. We find that the surface carbon budget can be described using four different classes that roughly correspond to regimes found in classical theoretical frameworks. We briefly discuss the possibility of using unsupervised learning to compare climate models.

II. METHODS

Here we describe the ocean circulation and biogeochemistry model that we used to evaluate the surface carbon budget. We also describe the unsupervised learning method (i.e. Gaussian mixture modeling) that we applied to the surface carbon budget.

A. Ocean biogeochemistry model

To quantify the steady state global ocean surface carbon budget, we use a coarse resolution ocean circulation and biogeochemistry model [9]. The numerical model is an instance of MITgcm (http://mitgcm.org/, [10], [11]) with a simple biogeochemistry component [12]. The biogeochemistry package uses six tracers: DIC, alkalinity, PO_4, dissolved organic phosphorous, oxygen, and iron. The export of biological carbon out of the surface is calculated as a function of available light, PO_4, and iron. The model uses a fixed grid with a horizontal resolution of $2.8^\circ \times 2.8^\circ$ in latitude-longitude and 23 vertical levels with gradually increasing cell thickness, with relatively thin cells in the the rapidly-changing surface and thicker cells in the relatively quiescent interior. Unresolved transport is parameterized using an isopycnal thickness diffusion scheme with a uniform diffusivity of 1000 m^2/s [13]. We also impose along-isopycnal diffusion at the same rate [14], and mixed-layer processes are parameterized using the K-Profile Parameterization (KPP) scheme [15]. Vertical diffusivity is set to 0.3×10^{-4} m^2/s in the upper 2000 m and increases to 10^{-4} m^2/s in the interior ocean following an arctangent profile [16]. The Arctic is not included in this model, in part due to convergence issues with latitude-longitude grids near the poles. The model was spun up for 1000 years and then run for another 100 years for evaluation. We average the last 10 years of the simulation in order to construct the steady state budget.

We evaluate the steady state surface carbon budget in the top 185 m of the model domain. The budget can be expressed as follows:

$$\begin{aligned} 0 = &- \mathbf{u} \cdot \nabla_H C - -w\frac{\partial C}{\partial z} \\ &+ \nabla_H(\mathbf{K}\nabla_H C) + \frac{\partial}{\partial z}\left(K_z \frac{\partial C}{\partial z}\right) \\ &+ V_P(1-f)K_H \Delta p\mathrm{CO}_2 \\ &+ \mathrm{FWF} + \mathrm{Bio}, \end{aligned} \tag{1}$$

where C is the dissolved inorganic carbon concentration, $\mathbf{u}$ is the horizontal velocity vector, ∇_H is the horizontal component of the gradient operator, w is the vertical velocity, z is the depth coordinate, $\mathbf{K}$ is the diffusivity tensor, K_z is the vertical component of the diffusivity tensor, and $\Delta p\mathrm{CO}_2$ is the air-sea difference in partial pressure of CO_2. The terms on the RHS of equation (1) represent the processes of horizontal advection, vertical advection, horizontal diffusion, vertical diffusion, air-sea gas exchange, freshwater flux, and biological sources and sinks of DIC, respectively. In terms of model diagnostics, the unresolved, parameterized fluxes are contained in the diffusive terms of the budget.

B. Gaussian mixture modeling

Gaussian mixture modeling (GMM) attempts to represent the density of data in an abstract space as a linear combination of multi-dimensional Gaussian functions [17]. A GMM is "trained" by adjusting the means and covariances of the Gaussian functions. GMM has been applied to ocean temperature and salinity data in order to identify different "profile types" in different ocean regions [18], [19].

We follow the method of [20], wherein each term of the steady-state, two-dimensional barotropic vorticity budget equation is used as a feature for unsupervised classification; each term/feature represents a different physical process. The result of their classification analysis is a robust, algorithmically defined global geography of ocean dynamical regimes [20]. In our application, we use each term of the surface carbon budget in equation (1) as a feature for classification analysis; in doing so, we represent the distribution of data in a seven-dimensional abstract feature space. At every $2.8^\circ \times 2.8^\circ$ model grid cell, there is a value for each of the seven terms of equation (1). The vector of budget term values from a grid cell is used as a single seven-dimensional "observation" in the clustering analysis. We weight each grid cell by its ocean surface area. We

do *not* standardize the budget term values beforehand, as we want the terms to retain their relative magnitudes. This should not affect the GMM fitting procedure, as the covariances of the Gaussian functions are generally allowed to scale as needed to fit the data.

The total number of classes N is a free parameter in GMM. Although one can use statistical tests to estimate the value of N with the highest likelihood relative to an overfitting "penalty" term (e.g. Bayesian Information Criterion or BIC), one can also use N as a description of the complexity of the statistical model. A simple statistical model with small N will likely be easier to interpret than a complex statistical model with large N. In this way, a range of GMM models with different N constitutes a model hierarchy, and one may be able to learn about the system under investigation based on how it changes as one adds or removes sources of complexity [21].

We use the Scikit-learn machine learning package in Python to carry out the classification analysis [22]. We use the expectation-maximization algorithm to determine the means and covariances of the Gaussians that have the highest probability of correctly representing the steady-state budget data as a linear combination of multi-dimensional Gaussian functions. We use the "full" covariance type to allow the Gaussians to change their orientations and covariances in any way that increases the overall probability of the distribution. We use every ocean grid cell from the numerical model, which is 4447 data points (or "observations") in total. Once the GMM has been trained, we use it to assign a class label to each grid cell. Specifically, GMM assigns to each grid cell a probability distribution across all of the classes, and it assigns each grid cell to the class with the maximum posterior probability.

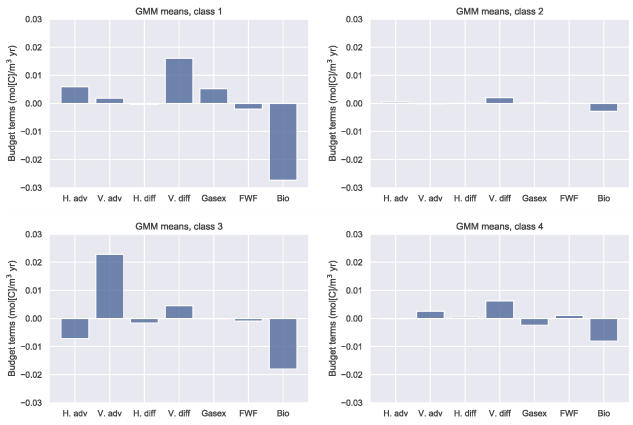


Fig. 1. Each class features a different balance distribution across the terms of the surface carbon budget. Here we plot the means of the four multi-dimensional Gaussian functions used to statistically model the data density. The terms are horizontal advection (H. Adv.), vertical advection (V. Adv.), horizontal diffusion (H. Diff.), vertical diffusion (V. Diff.), air-sea gas exchange (Gasex.), freshwater flux (FWF), and biological sources and sinks of carbon (Bio).

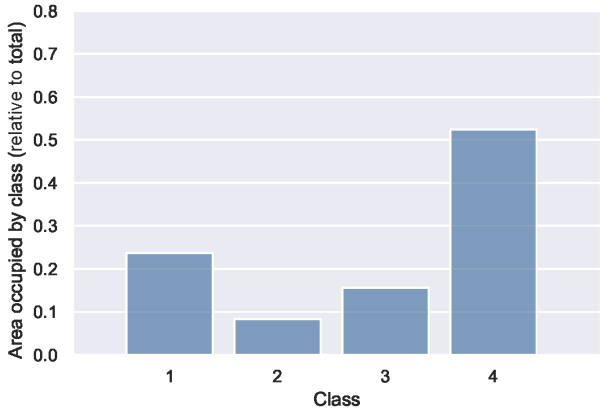


Fig. 2. The surface area occupied by each class, relative to the total ocean surface area.

III. RESULTS

In our implementation of GMM, each class broadly represents a different distribution of balances in the terms of equation (1). Specifically, each of the four seven-dimensional Gaussians can be described by a set of means (a seven-dimensional vector) and covariances (a tensor) across the different processes. For simplicity, we only show the means of the classes (Figure 1). We see that there are classes with less biological productivity (e.g. class 2), and classes with more biological productivity (e.g. class 1), corresponding to a large export of DIC from the surface waters. Note that these mean values do not necessarily represent every observation in a given class; they are means of the Gaussian functions used to represent the data. The surface area occupied by each class is shown in Figure 2.

Here we describe the GMM classes and the classical theoretical frameworks to which they approximately correspond. Despite the fact that GMM was not given any information about the latitude-longitude locations of the grid cells, it is still able to identify spatially coherent regimes in the surface carbon budget (Figure 3). Along with the distribution across processes (Figure 1), the spatial distribution of the labels helps in our attempt to interpret the classes. Class 1 corresponds to the highly productive open ocean, which is dominated by wintertime convection and mixing, seen in the term balance as vertical diffusion (Figure 1). Class

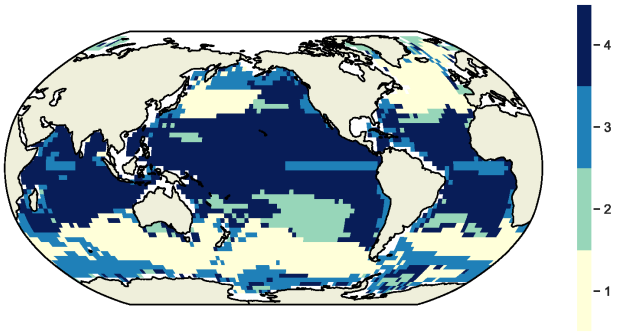


Fig. 3. GMM produces spatially coherent regimes for the surface carbon budget, despite the fact that it is not given any information about the location of the grid cells. Here we show the labels assigned by GMM to each grid cell. Regions with no data are masked out in white.

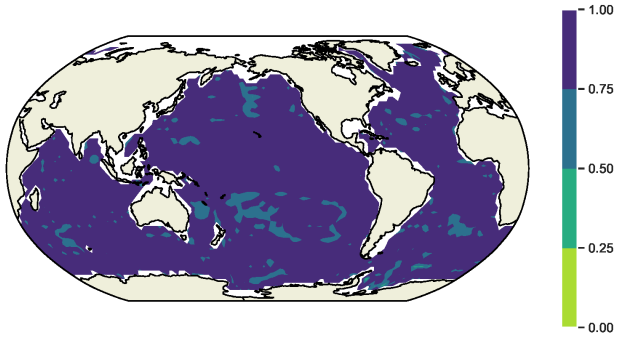


Fig. 4. Posterior probabilities can be used to identify transition regions between classes or where simple characterization of the class type is difficult. Regions with no data are masked out in white.

2 corresponds to relatively isolated patches of low productivity. Class 3 features wind-driven upwelling, as evidenced by the larger values of the advective terms. Both coastal and open ocean wind-driven upwelling can bring nutrients to the surface, where they can enable primary productivity and encourage the export of carbon out of the surface layer. This class is possibly the high nutrient low chlorophyll regime, which features relatively shallow mixed layer depths and (typically) iron limited productivity [23]. Finally, class 4 corresponds to the relatively unproductive gyres and tropics. The subtropical gyres are typically characterized by low local values of productivity, although they may still contribute significantly to the global carbon budget due to their large size [24]. Low productivity in the subtropical gyres is typically explained as a result of large-scale downwelling due to the wind-driven convergence of surface waters, which prevents productivity-enabling nutrients from reaching the surface waters [25]. Overall, the carbon distribution appears to reflect the nutrient budget, i.e. biological productivity is high in regions where upwelling can supply nutrients to the surface and low in regions where downwelling suppresses surface nutrient availability.

For each seven-dimensional observation of the steady state budget terms at a grid cell, GMM calculates a probability distribution across all four Gaussians. It labels each grid cell based on the Gaussian with the highest posterior probability. The maximum posterior probability is a measure of confidence in GMM's assignment of a grid cell to a class and can be used to characterize boundaries between classes. In this application, we find that the maximum posterior probability values are high ($\geq 90\%$) in the tropics and subtropics, with somewhat lower values between classes in the Southern Ocean. This may simply reflect the complex spatial structure of the classes in the Southern Ocean, which features numerous transition regions.

IV. DISCUSSION

In general, using budget terms as inputs to an unsupervised learning algorithm allows us to interpret clustering results in terms of balances between processes, representing an important link between data-derived results and the process-based physical and biogeochemical understanding that underpins much of modern oceanography. This approach may offer a viable bridge between machine learning methods and more traditional approaches.

We use four classes in this example implementation of GMM for ease of interpretation. Based on the BIC score, we could improve the overall likelihood of the GMM by increasing the number of classes to somewhere between 14-19 (see appendix). Although this would enhance the ability of the GMM to statistically describe the data density, it could decrease our ability to understand the results in terms of existing conceptual frameworks. The tradeoff between accuracy of representation and interpretability is a familiar contrast in ocean modeling. One strategy for dealing with this contrast is to use a hierarchical approach, in which we try to learn about a system by comparing models with different levels of complexity [21]. In terms of GMM, this would amount to changing the maximum number of classes and comparing results.

GMM as applied to budget terms may be a useful method for comparing different climate models, for example the ensemble members of the Climate Model Intercomparison Project [8]. These models often features biases with respect to each other, but they display

similar physical and biogeochemical regimes characterized by balances between processes. Unsupervised learning may offer a set of methods for objectively identifying these regimes in different models; the properties of the objective regimes could be compared, as opposed to comparing different geographical regions, which are often chosen using crude and somewhat arbitrary latitude-longitude boxes.

One limitation of this study is the relatively coarse resolution of the model; an application of GMM to a high-resolution biogeochemical state estimate like B-SOSE would be a welcome extension to this study [26]. We have also not thoroughly explored the many alternative unsupervised classification methods avalable, including DBSCAN and variational Bayesian approaches.

APPENDIX

Here we present additional information about the GMM classification results. BIC tends to increase as the likelihood of the statistical model increases with the total number of classes N, but that tendency is offset by a penalty term which discourages overfitting. Usually, one would choose the value of N with the minimum value of BIC, if the goal is to create a detailed statistical description of the dataset. The BIC mean score reaches a minimum at 19 classes, although the error suggests that the minimum could be between 14-19 (Figure 5(a)). In order to examine the distinctiveness of the clusters, we use two complementary dimensionality reduction techniques. First, we use principal component analysis (PCA) to project the data onto three PC axes that together explain 91% of the variance (52% PC1, 29% PC2, and 10% PC3). Projections of the principal components into 2D space show that class 1, which corresponds to the highly productive open ocean, is reasonably distinct from the others (Figure 5(b-d)). Class 2 is tightly clustered around the origin, which is consistent with the low values of the flux terms that characterize this class. Next, although classes 3 and 4 have some overlap around the origin, they do have distinct structures in PC space, with class 3 showing a larger spread along the PC1 axis.

For an alternative view on the distinctiveness of the classes, we employ t-SNE, a technique for exploring structures in high-dimensional data [27]. The t-SNE technique creates two-dimensional "maps" from high-dimensional data using non-linear transformations. It has a tunable parameter called "perplexity" which roughly corresponds to the attention paid to local versus global aspects of the data in feature space (see https://distill.pub/2016/misread-tsne/ for details). As we

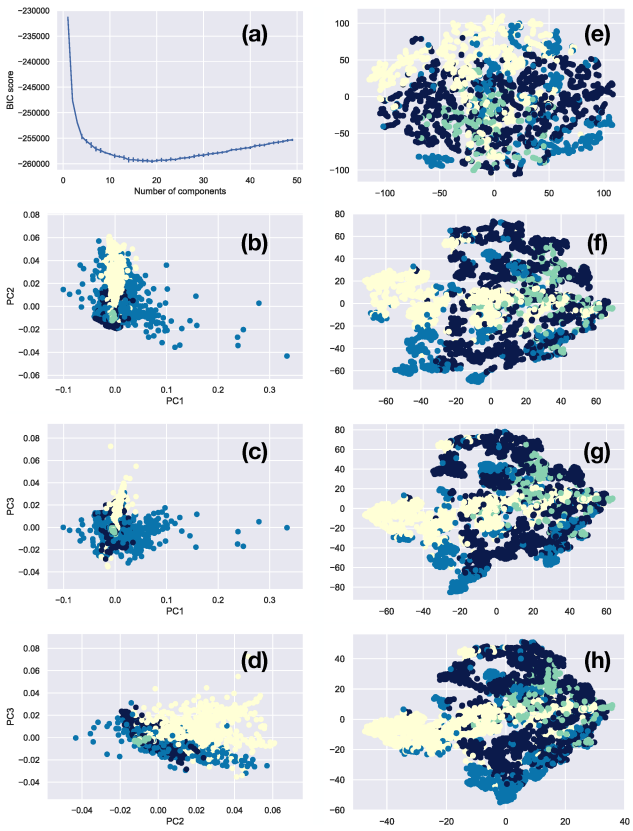


Fig. 5. Additional clustering diagnostics. (a) Mean and standard deviation of BIC scores from 25 independent instances of GMM for each value of N. (b-d) Reduced dimensionality view using a three-component PCA, viewed as three different projections onto 2D space. (e-h) Reduced dimensionality view using t-SNE for perplexity values of 5, 30, 50, and 100, respectively. The axes are the arbitrary t-SNE dimensions. Color values correspond to those in Figure 3.

increase perplexity, we see class 1 emerge as a distinct feature. Classes 2-4 have some considerable regions of overlap, but classes 3 and 4 have some distinct lobes above and below the class 1 cluster.

ACKNOWLEDGMENTS

DJ is supported by the Natural Environment Research Council, UKRI (grant NE/N018028/1). TI is supported by the US National Science Foundation (OCE1737188, OPP-1744755).

REFERENCES

[1] C. Sabine, R. Feely, N. Gruber, R. Key, K. Lee, J. L. Bullister, R. Wanninkhof, C. S. Wong, D. W. Wallace, B. Tilbrook, F. J. Millero, T.-H. Peng, A. Kozyr, T. Ono, and A. F. Rios, "The oceanic sink for anthropogenic CO2," *Science*, vol. 305, no. 367, 2004.

[2] R. Houghton, "Balancing the global carbon budget," *Annual Review of Earth and Planetary Sciences*, vol. 35, no. 1, pp. 313–347, 2007.

[3] S. Khatiwala, F. Primeau, and T. Hall, "Reconstruction of the history of anthropogenic CO2 concentrations in the ocean," *Nature*, vol. 462, pp. 346 EP –, 2009.

[4] T. Takahashi, S. Sutherland, R. Wanninkhof, C. Sweeney, R. A. Feely, D. W. Chipman, B. Hales, G. Friederich, F. Chavez, C. Sabine, A. Watson, D. C. E. Bakker, U. Schuster, N. Metlz, H. Yoshikawa-Inoue, M. Ishii, T. Midorikawa, Y. Nojiri, A. Kortzinger, T. Steinhoff, M. Hoppema, J. Olafsson, T. S. Arnarson, B. Tilbrook, T. Johannessen, A. Olsen, R. Bellerby, C. S. Wong, B. Delille, N. R. Bates, and H. J. W. d. Barr, "Climatological mean and decadal change in surface ocean pCO2, and net sea-air CO2 flux over the global oceans," *Deep Sea Research II*, vol. 56, pp. 554–577, 2009.

[5] D. C. Jones, T. Ito, Y. Takano, and W.-C. Hsu, "Spatial and seasonal variability of the air-sea equilibration timescale of carbon dioxide," *Global Biogeochemical Cycles*, vol. 28, no. 11, pp. 1163–1178, 2014.

[6] C. R. McClain, S. R. Signorini, and J. R. Christian, "Subtropical gyre variability observed by ocean-color satellites," *Deep Sea Research Part II: Topical Studies in Oceanography*, vol. 51, no. 1, pp. 281 – 301, 2004. Views of Ocean Processes from the Sea-viewing Wide Field-of-view Sensor (SeaWiFS) Mission: Volume 1.

[7] C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[8] A. Anav, P. Friedlingstein, M. Kidston, L. Bopp, P. Ciais, P. Cox, C. Jones, M. Jung, R. Myneni, and Z. Zhu, "Evaluating the Land and Ocean Components of the Global Carbon Cycle in the CMIP5 Earth System Models," *Journal of Climate*, vol. 26, no. 18, pp. 6801–6843, 2013.

[9] Y. Takano, T. Ito, and C. Deutsch, "Projected centennial oxygen trends and their attribution to distinct ocean climate forcings," *Global Biogeochemical Cycles*, vol. 32, no. 9, pp. 1329–1349, 2018.

[10] J. Marshall, A. Adcroft, C. Hill, and L. Perelman, "A finite-volume, incompressible Navier Stokes model for studies of the ocean on parallel computers," *Journal of Geophysical Research*, vol. 102, pp. 5753–5766, 1997.

[11] J. Marshall, C. Hill, L. Perelman, and A. Adcroft, "Hydrostatic, quasi-hydrostatic, and nonhydrostatic ocean modeling," *Journal of Geophysical Research*, vol. 102, pp. 5733–5752, 1997.

[12] S. Dutkiewicz, M. J. Follows, and J. G. Bragg, "Modeling the coupling of ocean ecology and biogeochemistry," *Global Biogeochemical Cycles*, vol. 23, no. 4, p. GB4017, 2009.

[13] P. Gent and J. Mcwilliams, "Isopycnal mixing in ocean circulation models," *Journal of Physical Oceanography*, vol. 20, pp. 150–155, 1990.

[14] M. Redi, "Oceanic isopycnal mixing by coordinate rotation," *Journal of Physical Oceanography*, vol. 12, pp. 1154–1158, 1982.

[15] W. Large, J. Mcwilliams, and S. Doney, "Oceanic vertical mixing: A review and a model with a nonlocal boundary layer parameterization," *Reviews of Geophysics*, vol. 32, no. 4, pp. 363–403, 1994.

[16] K. Bryan and L. J. Lewis, "A water mass model of the world ocean," *Journal of Geophysical Research: Oceans*, vol. 84, no. C5, pp. 2503–2517, 1979.

[17] D. Reynolds, *Gaussian Mixture Models*, pp. 659–663. Boston, MA: Springer US, 2009.

[18] G. Maze, H. Mercier, R. Fablet, P. Tandeo, M. L. Radcenco, P. Lenca, C. Feucher, and C. Le Goff, "Coherent heat patterns revealed by unsupervised classification of Argo temperature profiles in the North Atlantic Ocean," *Progress in Oceanography*, vol. 151, pp. 275–292, 2017.

[19] D. C. Jones, H. J. Holt, A. J. S. Meijers, and E. Shuckburgh, "Unsupervised Clustering of Southern Ocean Argo Float Temperature Profiles," *Journal of Geophysical Research - Oceans*, vol. 40, no. 2, pp. 1556–13, 2019.

[20] M. Sonnewald, C. Wunsch, and P. Heimbach, "Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions," *Earth and Space Science*, vol. 6, p. 784 794, 2019.

[21] I. Held, "The gap between simulation and understanding in climate modeling," *Bulletin of the American Meteorological Society*, vol. 86, no. 11, pp. 1609–1614, 2005.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[23] J. Pitchford and J. Brindley, "Iron limitation, grazing pressure and oceanic high nutrient-low chlorophyll (HNLC) regions," *Journal of Plankton Research*, vol. 21, pp. 525–547, 03 1999.

[24] W. J. Jenkins and S. C. Doney, "The subtropical nutrient spiral," *Global Biogeochemical Cycles*, vol. 17, no. 4, 2003.

[25] R. G. Williams and M. J. Follows, *Ocean Dynamics and the Carbon Cycle: Principles and Mechanisms*. Cambridge University Press, 2011.

[26] A. Verdy and M. R. Mazloff, "A data assimilating model for estimating Southern Ocean biogeochemistry," *Journal of Geophysical Research - Oceans*, vol. 122, no. 9, pp. 6968–6988, 2017.

[27] L. v. d. Maaten and G. Hinton, "Visualizing Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

WEATHER TYPES PREDICTION AT MEDIUM-RANGE FROM ENSEMBLE FORECASTS

Gabriel Jouan[1,3], Anne Cuzol[2], Valerie Monbet[3], Goulven Monnier[1]

***Abstract*—Medium-range weather forecasts can be of high economic value in many fields: agriculture, renewable energy production, maintenance operations planning. Such forecasts can be based on ensembles derived from weather models, and the postprocessing of such ensembles is an active research problem in the statistical weather community. In this work, we try to face the problem of long forecasting horizons, and focus on the multivariate case where different meteorological variables interact. The prediction problem is simplified and defined as the prediction of a weather type, which is a categorical variable defined by the interaction of the meteorological variables. We use machine learning techniques to predict this weather type from the multivariate ensemble forecasts. The algorithms are applied to a 5 to 10 days weather forecasting in the north-west of France, based on wind and precipitation data from the ECMWF ensemble system.**

I. MOTIVATION

Nowadays, meteorological institutes provide ensemble forecasts like, for instance, the European Center for Medium-range Weather Forecast (ECMWF) ensemble. However such ensemble forecasts of surface weather parameters are known to be under-dispersed and often biased [1],[2],[3],[4].

To improve the accuracy of such forecasts, statistical postprocessing has been studied these last years. One of the most common approach to calibrate the ensemble for one given variable is based on a regression model which helps to predict observations of the variable given a description of the ensemble as input. For example, the state-of-the-art method, referred to as Ensemble Model Output Statistics (EMOS) [5], is based on an heteroscedastic linear regression. More recently, non-parametric algorithms have been proposed [6], [7], [8]. Multivariate calibration techniques have also been developed in order to reproduce dependencies between variables [9].

Corresponding author: G Jouan, gabriel.jouan@scalian.com
[1]Scalian Alyotech Rennes; [2]Univ. Bretagne-Sud, LMBA; [3]Univ. Rennes, IRMAR

In this article, we focus on multivariate forecasting for horizon higher than 3 days. Such medium to long-range forecasts are of high value, for instance for maintenance operations in many fields, but this problem is known to be difficult. In this work, the goal will be to predict weather types from ensemble forecasts, instead of performing a calibration of the whole multivariate distribution of the meteorological variables of interest. The weather type is a categorical variable, described for instance by ”good”, ”windy”, ”rainy”. Such qualitative information is sufficient for many applications.

A natural approach to predict such weather types is to apply direct classification algorithms. A state-of-the-art non-linear method is the random forest classifier (RFC) [10] based on the aggregation of tree classifiers [11]. To compare random forest results on multiple weather types classification, a linear approach can be used: the multiple logistic regression, also called multinomial lasso regression (MLR) [12].

Such direct classification algorithms applied to the ensemble forecasts jointly perform a correction and a classification, which can be difficult. An other way to solve the weather type prediction problem is to perform a multivariate calibration, followed by a transformation of the output into weather types. The multivariate calibration consists in applying independent univariate calibrations for each variable, followed by a reordering method [13]. This approach will be used as comparison.

The paper is organized as follows. In Section II, we describe the classification algorithms used for the prediction of weather types. In Section III part A, the considered data are introduced and the weather types are defined. In part B, the performances of the proposed methods are compared for a forecasting range of 5 days and 10 days, for a chosen location in the north-west of France. The paper ends with some concluding remarks in Section IV.

II. METHODOLOGY FOR THE PREDICTION OF WEATHER TYPES

Our aim is to calibrate multivariate medium-range forecast ensembles for one location. Since qualitative information is sufficient for some applications, we propose to tackle the problem of weather types prediction, each weather type being defined from several meteorological variables. For instance, the weather types can be defined as "good", "windy", "rainy", "windy and rainy". Our contribution will be to propose classification algorithms where the inputs are given by a multivariate ensemble and the output is the weather type.

To solve the classification problem, two approaches will be considered. The first one is based on a direct application of machine learning classification algorithms (Section A). The second one consists in applying the weather type definition to the output of classical calibration methods (Section B).

The ensemble members can not be used directly as inputs of the machine learning algorithm because they are exchangeable [14]. This means that ensemble members are invariant under permutation, and consequently can not be used as predictors. Then, following [6], [15], the considered features are some statistics of the ensembles of the precipitation and wind speed, namely means, standard deviations, kurtosis, skewness, first and ninth deciles, interquartile range and precipitation probabilities. It is standard to also add the control and the high resolution members to the features set. The month and the hour, considered as factor inputs, allow to take into account the daily and yearly cycles existing in the data. Finally, since we consider (observed) weather types as output, we decide to also add to the inputs the corresponding weather types computed from the raw (uncalibrated) ensembles.

A. *Direct classification*

Two classical machine learning algorithms are considered: random forest and penalized multinomial regression. The random forest is known to be more flexible and multinomial regression more robust.

Random forest classifier (RFC) was proposed by [11] and [10]. It combines elementary classification trees, learned on random samples generated from the data, to estimate the probability of each weather type. The principle of each tree is to infer a partition of the input space by a greedy algorithm. Each part is called a leaf. At each step of the algorithm, the current leaf is splitted into two parts if it improves the Gini impurity. At the end, the probabilities of weather types are obtained from the mean of all trees.

Penalized multinomial regression (MLR) is described for instance in [12]. Each multinomial regression model predicts the probability of one of the K weather types against the others so that $K-1$ models are fitted. We expect that all input variables are not of the same importance to help to discriminate the weather types. So, a Lasso penalty is introduced in the estimation task which helps to select the most discriminant variables. The idea is to penalize the log-likelihood by a the sum of the absolute values of the coefficients of the regression. It has the consequence to shrink to zero the coefficients of useless inputs and to lead to a more robust prediction tool.

B. *Classification from multivariate calibration*

The direct machine learning classification algorithms proposed in this paper has to be compared to other machine learning solutions proposed in the ensemble forecast calibration literature, in particular [6] which also used random forests. In [6] and reference therein, the authors perform univariate calibration with quantile regression forests [16].

Here, we propose to apply a quantile regression forest separately for the calibration of each meteorological variable (wind speed and precipitation for instance). Then the two independent calibrated ensembles are combined by a Schaake Shuffle (SS) algorithm [13] to reproduce the dependent structure existing between variables. In this reordering algorithm, the marginal postprocessings are combined to reproduce the empirical copula estimated from past observations. One recent improvement, referred to as SimSchaake ([17], [9]), proposes to combine the SS algorithm with analog approaches. It allows to select past observations from meteorological configurations close to the current one and it reduces the bias in the estimation of the dependence structure. After applying this reordering procedure, the multivariate output is transformed into the predefined weather types.

Methods have been applied using the R software with the "randomForest" [18], "glmnet" [19] and "quantregforest" [20] packages for the RFC, MLR and quantile regression forests algorithms.

III. APPLICATION

The classification algorithms are now applied to data from the north-west of France (city of Rennes), described in Section A. The performances of the proposed methods are then compared in Section B.

A. Data description and weather types definition

Ensemble forecast data of the ECMWF [21] are collected from the Thorpex interactive grand global ensemble archive (TIGGE) [2], [1]. The TIGGE archive includes a minimum of 10 ensemble forecasting systems on a time-period from 2008 to 2018.The ECMWF ensemble system is composed of 50 exchangeable ensemble members generated from the Ensemble of Data Assimilations (EDA) based perturbations with singular vectors in the initial conditions and stochastic physics models [22], [23].

Collected data are composed of observations and ensemble forecasts of precipitation (Precip, mm) and wind speed (WS, m.s^{-1}) at forecasting range 5 days and 10 days, two runs (6 am and 6 pm) for the French city of Rennes.

As mentioned earlier, the continuous variables are transformed to define weather types. $K = 4$ balanced weather types are chosen. The data contains approximately the same number of observations in each weather type avoiding an unbalanced classification problem. For an observation vector $y = (y^{Precip}, y^{WS})^\top$, and the set of thresholds $\{0.02, 2.8, 4\}$, the ϕ thresholding function is defined as:

$$\phi(y) = \begin{cases} 1 & \text{if} \quad y^{Precip} < 0.02, y^{WS} < 2.8 \\ 2 & \text{if} \quad y^{Precip} < 0.02, y^{WS} \geq 2.8 \\ 3 & \text{if} \quad y^{Precip} \geq 0.02, y^{WS} < 4 \\ 4 & \text{if} \quad y^{Precip} \geq 0.02, y^{WS} \geq 4 \end{cases} \quad (1)$$

The four weather types are referred to as "good" if $\phi(y) = 1$, "windy" if $\phi(y) = 2$, "rainy" if $\phi(y) = 3$ and "rainy and windy" if $\phi(y) = 4$. Other thresholds could be chosen depending of the application in mind.

B. Results

The classification algorithms performances are evaluated and compared using classical scores like the accuracy, the precision and the recall [24]. All the scores are computed by cross-validation. For that, the data set is randomly splitted into 2 subsets. The validation subset contains 912 days, randomly extracted from the period 2014-2018. The learning subset is composed of all remaining days over the period 2008-2018. This is repeated 30 times in order to approximate the distribution of the scores.

A k-fold cross-validation has been performed ($k = 10$), and the penalization hyperparameter of the MLR model has been fixed to 0.03.

The reference result that we seek to improve is the forecast obtained from the uncalibrated multivariate ensemble transformed into weather types following rules (1). This forecast will be referred to as "Raw" in the sequel.

Figure 1 shows the accuracy score on the left panel. It is an overall criteria which is close to 1 if the weather types are correctly classified. The accuracy is close to 0.6 for all the methods, but the direct classification algorithms (MLR and RFC) slightly improve both the "Raw" result and the classification based on multivariate calibration.

The right panel of Figure 1 shows the precision and the recall which allow to analyze results per weather type. Precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. We want to achieve a good compromise between precision and recall for each weather type.

We can see on Figure 1 that the "Raw" forecast at 5 days leads to the highest recall for the "good" weather type, but with the lowest precision. On the other hand, it leads to the lowest recall for the "rainy" class, while having the highest precision.

Compared to the "Raw" result, the RFR/SimSS method improves the recall of the "rainy", but deteriorates the precision. On the other hand, the direct classifications (MLR and RFC) show an improvement of recall for the "rainy" and "windy and rainy" classes, while maintaining a precision close to the "Raw" result for these two weather types. Note that this gain in recall is higher for RFC, but with a slight decrease in precision compared to MLR.

It can be observed that all methods lead to a comparable precision for the "windy" class, but none of the methods is able to improve the recall obtained by the "Raw" ensembles for this weather type.

Figure 2 shows the forecast results for a horizon of 10 days. A global decrease of all classification scores can be observed, due to the increase of the uncertainties of the numerical weather prediction system. However, the RFC method is still leading to the highest accuracy for this long-range forecasting problem. One interesting point is that the classification obtained from the multivariate calibration (RFR/SimSS) is not able to improve the "Raw" result obtained from uncalibrated ensembles.

The results of the precision and recall for 10 days ensemble forecast display higher variations between the weather types than the 5 days ensemble forecast.

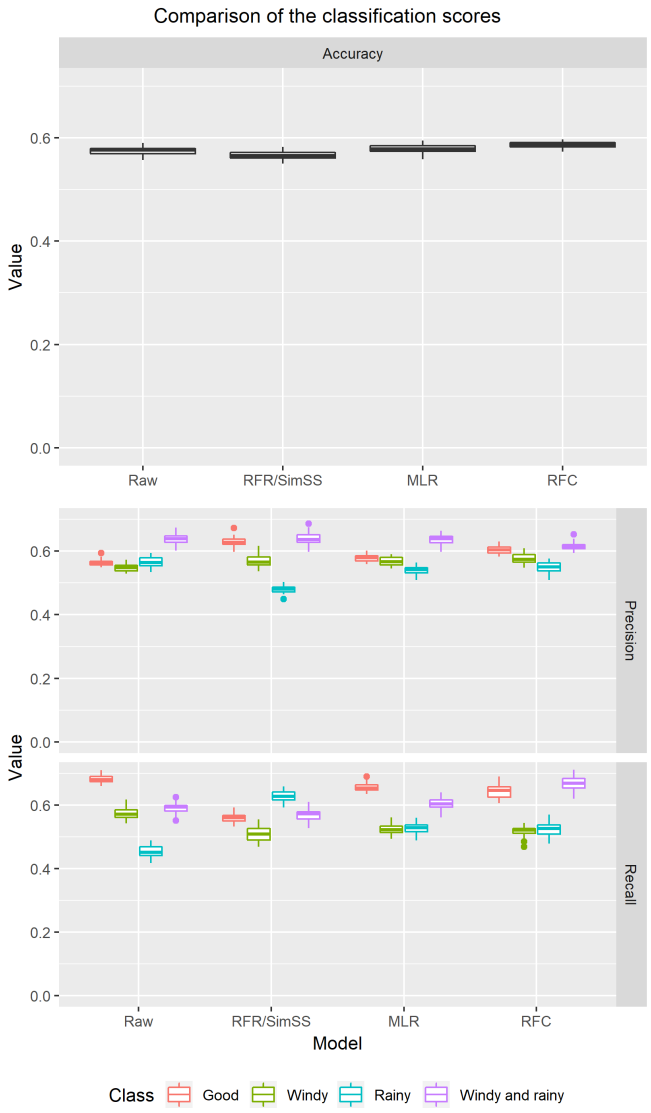


Fig. 1. Classification scores for a horizon of 5 days. Raw: Forecast from uncalibrated multivariate ensemble; RFR/SimSS: Classification obtained from a multivariate calibration; MLR: Multinomial lasso regression; RFC: Random forest classifier. Left: Accuracy scores; Top-right: Precision scores for each weather type; Bottom-right: Recall scores for each weather type.

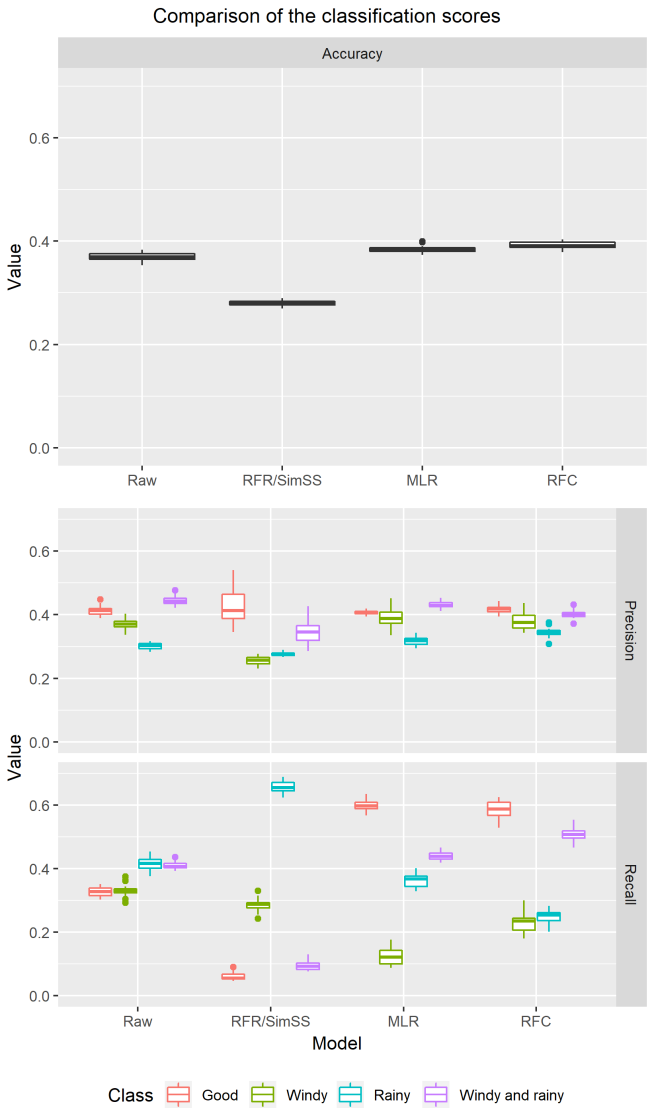


Fig. 2. Classification scores for a horizon of 10 days. Raw: Forecast from uncalibrated multivariate ensemble; RFR/SimSS: Classification obtained from a multivariate calibration; MLR: Multinomial lasso regression; RFC: Random forest classifier. Left: Accuracy scores; Top-right: Precision scores for each weather type; Bottom-right: Recall scores for each weather type.

In the right panel of the Figure 2, RFR/SimSS model is highly overestimating the ”rainy” type and shows a poor probability of detection of ”good” and ”windy and rainy” types. Direct classification algorithms (MLR and RFC) lead to the best recall for ”good” and ”windy and rainy”, while maintaining a comparable level of precision. However, there is a decrease of performance over the ”Raw” result for the ”windy” and ”rainy” weather types.

IV. CONCLUDING REMARKS

Compared to the reference result obtained from the uncalibrated multivariate ensemble of forecasts, direct classification models lead to a small improvement in prediction accuracy. This is not the case for the classification obtained after a multivariate calibration. However, the study of precision and recall scores show that this improvement is not observed for all weather types. For instance, direct classification models increase the probability of detection of ”rainy” and ”windy and rainy” weather types for a horizon of 5 days.

For a longer horizon (10 days), these models lead to better detections for "good" and "rainy and windy" weather types. A study of variables importance in the classification models (not shown in this paper) can help understanding the differences in performance between weather types.

In [6] and [15], the quantile regression forest has been compared to linear approaches for short-range ensemble forecast. A comparison of linear approaches on the wind speed and cumulative rainfall with ensemble forecast at 5 days and 10 days is needed. The recent EMOS models of [25] for calibration of wind speed and [26] for calibration of precipitation will be applied.

Other ensembles with a shorter medium-range (3 days) will be tested and compared to assess the classification results obtained at 5 days and 10 days. Also, the weather types prediction problem needs to be investigated on other spatial locations.

ACKNOWLEDGEMENTS

This research was supported by funding from Scalian group and IRMAR.

REFERENCES

[1] P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D. H. Chen, B. Ebert, M. Fuentes, T. M. Hamill, K. Mylne, *et al.*, "The thorpex interactive grand global ensemble," *Bulletin of the American Meteorological Society*, vol. 91, no. 8, pp. 1059–1072, 2010.

[2] Y.-Y. Park, R. Buizza, and M. Leutbecher, "Tigge: Preliminary results on comparing and combining ensembles," *Quarterly Journal of the Royal Meteorological Society*, vol. 134, no. 637, pp. 2029–2050, 2008.

[3] T. M. Hamill and S. J. Colucci, "Verification of eta–rsm short-range ensemble forecasts," *Monthly Weather Review*, vol. 125, no. 6, pp. 1312–1327, 1997.

[4] T. M. Hamill and J. S. Whitaker, "Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application," *Monthly Weather Review*, vol. 134, no. 11, pp. 3209–3229, 2006.

[5] T. Gneiting, A. E. Raftery, A. H. Westveld III, and T. Goldman, "Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation," *Monthly Weather Review*, vol. 133, no. 5, pp. 1098–1118, 2005.

[6] M. Taillardat, O. Mestre, M. Zamo, and P. Naveau, "Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics," *Monthly Weather Review*, vol. 144, no. 6, pp. 2375–2393, 2016.

[7] S. Scher and G. Messori, "Predicting weather forecast uncertainty with machine learning," *Quarterly Journal of the Royal Meteorological Society*, vol. 144, no. 717, pp. 2830–2841, 2018.

[8] J. B. Bremnes, "Constrained quantile regression splines for ensemble postprocessing," *Monthly Weather Review*, vol. 147, no. 5, pp. 1769–1780, 2019.

[9] R. Schefzik and A. Möller, "Ensemble postprocessing methods incorporating dependence structures," in *Statistical Postprocessing of Ensemble Forecasts*, pp. 91–125, Elsevier, 2018.

[10] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[11] L. Breiman, J. Friedman, R. Olshen, and C. Stone, "Classification and regression trees. wadsworth int," *Group*, vol. 37, no. 15, pp. 237–251, 1984.

[12] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[13] M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, "The schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields," *Journal of Hydrometeorology*, vol. 5, no. 1, pp. 243–262, 2004.

[14] M. Courbariaux, *Contributions statistiques aux prévisions hydrométéorologiques par méthodes densemble*. PhD thesis, Université Paris-Saclay, 2017.

[15] M. Taillardat, A.-L. Fougères, P. Naveau, and O. Mestre, "Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting," *Weather and Forecasting*, no. 2019, 2019.

[16] N. Meinshausen, "Quantile regression forests," *Journal of Machine Learning Research*, vol. 7, no. Jun, pp. 983–999, 2006.

[17] R. Schefzik, "A similarity-based implementation of the schaake shuffle," *Monthly Weather Review*, vol. 144, no. 5, pp. 1909–1921, 2016.

[18] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[19] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of Statistical Software*, vol. 39, no. 5, pp. 1–13, 2011.

[20] N. Meinshausen, *quantregForest: Quantile Regression Forests*, 2017. R package version 1.3-7.

[21] E. Directorate, "Describing ecmwfs forecasts and forecasting system," *EcMWF Newsletter*, vol. 133, pp. 11–13, 2012.

[22] R. Buizza, M. Leutbecher, and L. Isaksen, "Potential use of an ensemble of analyses in the ecmwf ensemble prediction system," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 134, no. 637, pp. 2051–2066, 2008.

[23] R. Buizza, "Weather prediction in a world of uncertainties: should ensembles simulate the effect of model approximations?," in *ECMWF/WWRP Workshop: Model Uncertainty*, (ECMWF, Reading), 05/2016 2016.

[24] C. Rijsbergen, "v.(1979)," *Information retrieval*, vol. 2, 1979.

[25] S. Baran and S. Lerch, "Mixture emos model for calibrating ensemble forecasts of wind speed," *Environmetrics*, vol. 27, no. 2, pp. 116–130, 2016.

[26] M. Scheuerer and T. M. Hamill, "Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions," *Monthly Weather Review*, vol. 143, no. 11, pp. 4578–4596, 2015.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

A DIRECT APPROACH TO DETECTION AND ATTRIBUTION OF CLIMATE CHANGE

Enikő Székely[1,*], Sebastian Sippel[2], Reto Knutti[2], Guillaume Obozinski[1], Nicolai Meinshausen[2]

***Abstract*—We present here a novel statistical learning approach for detection and attribution (D&A) of climate change. Traditional optimal D&A studies try to directly model the observations from model simulations, but practically this is challenging due to high-dimensionality. Dimension reduction techniques reduce the dimensionality, typically using empirical orthogonal functions, but as these techniques are unsupervised, the reduced space considered is somewhat arbitrary. Here, we propose a supervised approach where we predict a given external forcing, e.g., anthropogenic forcing, directly from the spatial pattern of climate variables, and use the predicted forcing as a test statistic for D&A. We want the prediction to work well even under changes in the distribution of other external forcings, e.g., solar or volcanic forcings, and therefore formulate the optimization problem from a distributional robustness perspective.**

I. INTRODUCTION

Traditional Detection and Attribution (D&A) methods quantify the connection between observations and model simulated responses to different external forcings [1], [2], [3]. While detection aims to find if there is a change in the observations that cannot be explained by internal variability alone, attribution tries to assign the detected change to a particular external forcing or a combination of forcings. D&A studies first reduce the dimensionality by projecting onto the space spanned by the first few empirical orthogonal functions (EOFs)/principal components (PCs), and then regression is used in this reduced space to estimate the scaling factors [1]. One of the issues with the dimension reduction is that the procedure is unsupervised, and the resulting reduced space depends on the precise form of the dimension reduction technique. EOFs/PCs reduce the dimension by finding the few leading eigenvectors/fingerprints that maximize the variance, but the choice of the number of eigenvectors remains subjective.

[1]Swiss Data Science Center, ETH Zürich and EPFL, Switzerland [2]ETH Zürich, Switzerland, Corresponding author: E. Székely, eniko.szekely@epfl.ch

We propose here a proof of concept in a perfect model scenario where we predict directly the radiative forcing, e.g., anthropogenic forcing, in a supervised way, and use the predicted radiative forcing as a test statistic for D&A. The supervised setting allows us to find the projection of interest that best explains the radiative forcing, and avoids the arbitrariness of unsupervised dimension reduction as preprocessing step. To ensure that the results are robust to changes in the distribution of other external forcings, e.g., solar or volcanic forcing, we formulate the optimization problem from a distributional robustness perspective. We aim to find a robust estimator for a whole class/set of distributions, not only for the target population distribution. The set of distributions will be given by climate interventions in model simulations, e.g., control runs, Representative Concentration Pathways (RCPs), and the class of shift interventions on the external forcings, e.g., natural, solar, or volcanic forcing. This work fits into the emerging framework of data-driven approaches for detection and attribution using data assimilation [4] or statistical and machine learning [5], [6].

II. METHODOLOGICAL FRAMEWORK

A. Traditional Detection and Attribution

Traditional D&A studies first extract the fingerprint of external forcings from model simulations driven with the respective forcing by averaging across a large number of runs to reduce the influence of internal variability [7], [8]. In addition, the so-called optimal D&A studies project both the model simulations and the observations onto the space spanned by the first few EOFs of a set of (unforced) control simulations that feature only internal climate variability [1].

Let $X_{obs}^{EOF} \in \mathbb{R}^{T\times d}$ and $X_{M}^{EOF} \in \mathbb{R}^{T\times d}$ be the projection of the observations X_{obs} (e.g., temperature, precipitation) and climate responses from model simulations X_M (e.g., temperature, precipitation) onto the first d EOFs of internal (natural) variability, where T is the number of simulated years. X_M consists of k sets of forced simulations, typically $k = 2$ with

$X_M = \{X_{ANT}, X_{NAT}\}$, i.e., simulations with only anthropogenic and only natural forcing, respectively. The scaling parameters $\alpha \in \mathbb{R}^k$ corresponding to the set of forced simulations, e.g., α_{ANT} and α_{NAT}, are estimated from regression in the space of the EOFs:

$$\hat{X}_{obs}^{EOF} = \sum_{i=1}^{k} \alpha_i X_M^{i,EOF}. \quad (1)$$

The magnitude and confidence intervals of the scaling factors α_i indicate how much of the signal in the observations can be attributed to each particular forcing.

B. Data-driven Detection and Attribution

Let $Y \in \mathbb{R}^n$ be the simulated radiative forcing (e.g., anthropogenic, volcanic, solar, GHG, CO2), and $X \in \mathbb{R}^{n \times p}$ the matrix of specific climate variable measurements (e.g., temperature, precipitation, humidity) from climate model simulations in any given year, where n is the number of samples (i.e., the total number of simulated years across all simulations) and p the dimensionality of the data (the number of features or spatial grid cells). The radiative forcing is the net change in the energy balance of the Earth system due to some imposed perturbation [9]. Here, the matrix X is obtained by concatenating the different model simulation runs, and not through averaging as in the case of X_M from (1), therefore $n = m \times T$, where m is the number of model simulation runs and T the number of years simulated for each model.

The alternative data-driven D&A approach that we propose here predicts the external forcing $y \in Y$ directly from model simulations $x \in X$. Let $f_\beta(x)$ be the function that predicts y and is parameterized by β. The parameters β are estimated by minimizing a loss function $l(y, f_\beta(x))$ over the population drawn from some target population distribution $(x, y) \sim P$:

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}}\, \mathbb{E}_{(x,y)\sim P}[l(y, f_\beta(x))]. \quad (2)$$

The model used $f_\beta(\cdot)$ can be a linear or nonlinear (kernel) regression model, a random forest or a deep neural network [5]. The estimator in (2) only optimizes over one target population distribution $(x, y) \sim P$, however as we will see in the next sections, changes in the distribution of the data, e.g., stronger solar or volcanic forcing, can lead to poor prediction results. Our goal here is to protect ourselves against such distributional changes in the external forcings and optimize over a whole class of distributions in order to ensure robustness (for details see Sects. II-C and II-D).

In the statistical model from (2), the parameters β are learned from climate model simulations, and can be used to predict the external forcing from observations:

$$\hat{y}_{obs} = f_\beta(x_{obs}),$$

where x_{obs} are the full observational maps, and $\hat{y}_{obs}$ is the predicted observed forcing. We focus in this short paper on a perfect model scenario where we predict data from model simulations and leave the prediction of the observations for future work.

Traditional D&A tries to explain an *observed* climate pattern as a function of *modelled* climate patterns from simulations driven with different external forcings as in eq. (1). However, due to the high-dimensionality, this step cannot be performed directly on the original data. D&A therefore first extracts the fingerprints using an EOF analysis and the regression is performed in the EOF space. Our direct approach is an alternative to this step of fingerprint extraction. Instead of extracting the fingerprint using an (unsupervised, and therefore somewhat arbitrary) EOF analysis, we extract the fingerprint using directly the information contained in the radiative forcing in a supervised way. We note that our goal is not to predict the radiative forcing, but to use it to extract the fingerprint and, as explained in the following, to define a test statistic used for detection and attribution.

In the data-driven supervised approach that we propose, *detection* is done by testing against the null hypothesis that the predicted forcing does not differ from internal (natural) variability, i.e., the predicted forcing is not significantly different from zero. Practically, this is done by considering the predicted forcing $\hat{y}$ (either from model simulations in a perfect model scenario, or from observations) as a one-dimensional vector test statistic and computing the confidence intervals of the prediction. Detection occurs if the test statistic $\hat{y}$ is outside the pre-industrial range, i.e., the confidence intervals do not contain zero; and *attribution* is established if the true forcing lies within the confidence intervals of the predicted forcing.

C. Distributional robustness

The climate response x, e.g., temperature, is potentially influenced by multiple external forcings. Let us consider that we have three external forcings, e.g., solar, volcanic and anthropogenic forcings (see causal diagram in Fig. 1), and let's say we want to predict the anthropogenic forcing $y = F_3$. The diagram can be extended to include other forcings if necessary, or to split the anthropogenic forcing into its constituent parts, e.g., GHG, CO_2.

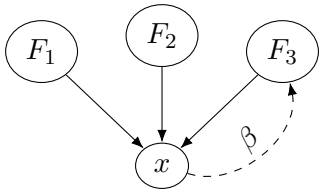


Fig. 1. Causal diagram for the effect (plain arrows) of external forcings F_1, F_2, and F_3, e.g., solar, volcanic and anthropogenic forcing, on the climate response x, e.g., temperature, and the regression problem (dashed arrow) for predicting $y = F_3$.

As discussed in the previous section, if we were to predict the anthropogenic forcing using the regression model in (2), we would only optimize over β for the observed distribution $(x, y) \sim P$. But, instead of seeking an estimator $f_\beta(\cdot)$ which is just a good predictor of the value of the forcing for the given distribution, we actually would like $f_\beta(\cdot)$ to capture the specific effect of the targeted forcing regardless of the strength of the other forcings. In other words, we would like to guarantee good prediction results even under distributional changes and we want the null distribution of the test statistic (in case of detection) to be valid even under changed solar/volcanic forcing.

The class of distributions $\mathcal{Q}$ over which we want to achieve robustness is generated both by interventions on the climate models (e.g., control runs, RCPs, anthropogenic runs, natural runs), and the class of shift interventions, i.e., interventions that shift the value of a variable in a given direction [10], on the external forcings. For example, in the graph from Fig. 1, the shift distributions $Q \in \mathcal{Q}$ are obtained by shifting the forcing F_1 or F_2, e.g., solar or volcanic forcing.

The distributionally robust form of the estimator in (2) is given by

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q}[l(y, f_\beta(x))], \tag{3}$$

that optimizes over a whole class of distributions $\mathcal{Q}$ instead of just a single target population distribution P [10], [11]. Distributional robustness is formulated here as a worst-case scenario, where solving for the most difficult case guarantees good prediction results for unseen future distributions.

D. Anchor regression

In the example from Fig. 1, we would like to protect ourselves against changes in the distribution of the solar forcing F_1, the volcanic forcing F_2, or both. We call these variables *anchors*, and in a linear setting where $f_\beta(x) = x^T\beta$ and the loss function is the least squares empirical risk $\sum_{i=1}^n l(y_i, f_\beta(x_i)) = \|Y - X\beta\|_2^2$, we use anchor regression [12] to achieve the distributional robustness from (3).

Let the anchor variables be $A \in \mathbb{R}^{n\times q}$, where n is the number of samples and q is the number of anchors. The robust estimator of anchor regression is given by

$$\hat{\beta}^\gamma = \underset{\beta}{\operatorname{argmin}} \|(I_n - \Pi_A)(Y - X\beta)\|_2^2 + \gamma\|\Pi_A(Y - X\beta)\|_2^2, \tag{4}$$

where $\Pi_A \in \mathbb{R}^{n\times n}$ is the matrix that projects on the column space of A, i.e., $\Pi_A = A(A^TA)^{-1}A^T$, $I_n \in \mathbb{R}^{n\times n}$ is the identity matrix, and γ is the "causal" regularization parameter that gives the strength of the shift intervention on the anchor variable. The causal regularization encourages orthogonality (or uncorrelatedness) of the residuals with the anchor variable. For the graph in Fig. 1, this ensures that the prediction accuracy remains good even if the strength of the solar or volcanic forcing changes. For $\gamma = 1$, the projection of the residuals on Π_A vanishes between the two terms and anchor regression coincides with ordinary least squares:

$$\hat{\beta}^1 = \underset{\beta}{\operatorname{argmin}} \|Y - X\beta\|_2^2, \tag{5}$$

where the parameters $\beta \in \mathbb{R}^p$ are the maps of regression coefficients (that can be interpreted in a more traditional sense in climate science as "fingerprints"). As the solution of both ordinary least squares from (5) and anchor regression from (4) can be prone to overfitting, we include a regularization term in the optimization problem. Because we want to ensure the smoothness of the maps β, we will use ridge (Tikhonov) regularization [13], and the estimator for the model in (4) can be written as ridge regression on a transformed data set:

$$\hat{\beta}^\gamma = \underset{\beta}{\operatorname{argmin}} \|\tilde{Y} - \tilde{X}\beta\|_2^2 + \lambda\|\beta\|_2^2, \tag{6}$$

where λ is the regularization parameter that controls the bias-variance tradeoff, and $\tilde{X} = (I_n - \Pi_A)X + \sqrt{\gamma}\Pi_A X$ and $\tilde{Y} = (I_n - \Pi_A)Y + \sqrt{\gamma}\Pi_A Y$ are the transformed data sets. The second term in (6) penalizes large regression coefficients, and handles the multicollinearity of the predictors.

Anchor regression finds the direction β that explains the component of the climate response to the forcing of interest, e.g., the anthropogenic forcing, that is orthogonal to other components that are (possibly) common in the response to other forcings. This allows us to do attribution of the detected change in the climate variable: if the projection on β (predicted forcing) is

similar enough to the forcing of interest (true forcing), the change can be attributed to the respective forcing.

III. DATA

We use data from climate model simulations from CMIP5 (Climate Model Intercomparison Project) [14] and consists of control runs and Representative Concentration Pathways (RCPs) [15] – RCP 2.6, RCP 4.5, RCP 6, RCP 8.5 – that outline plausible forcing trajectories throughout the 21st century that are used to drive climate model simulations. Here we use 42 control run simulations and 40 RCP 8.5 model simulations, so 82 model simulations from 21 climate models. Each model simulation has an annual resolution and runs for 231 years from 1870 to 2100. In total there are $n = 82 \times 231 = 18,942$ samples. The samples are two-dimensional spatial maps, and the spatial resolution is $p = 144 \times 72 = 10,368$ dimensions.

We first subtract the mean of the period 1870-1920 from each model individually in order to remove model biases in mean temperature, and then standardize the data prior to regression analysis. The regularization coefficient λ is chosen by cross validation with the folds built model-wise, i.e., we make sure that data from the same climate model falls in the same fold. We use here $k = 3$ folds. Likewise, the splitting into training and testing is also done model-wise. This ensures that we are testing only on full models that have not been seen during training. The data is split into 75% of models for training, and the remaining 25% of models for testing.

IV. EXPERIMENTS AND RESULTS

We report results for the prediction of the anthropogenic forcing using the volcanic forcing as anchor (Fig. 2). The first row shows the results of ridge regression (anchor regression with $\gamma = 1$), while the second and third row show the results for anchor regression with two different values of the "causal" parameter γ. The first column shows the raw coefficients β of the regression; the second column shows the prediction results together with the RMSE and R2 score for each case; and the last column plots the residuals against the anchor variable. We would like to obtain residuals that are uncorrelated with (or ideally independent from) the anchor to guarantee good prediction results even if the anchor changes. Constraining with the volcanic anchor slightly lowers the prediction accuracy (lower R2 and higher RMSE), however it also protects against a strong volcanic forcing. The correlation of the residuals with the anchor (last column) goes to zero as we increase the parameter γ from anchor regression. In the middle column we observe one testing model that behaves fairly different from the rest of the models (represented by the points that deviate the most from the black line). The raw coefficients with low causal regularization have mostly positive values indicating that all grid points contribute to explain the warming, but rely more on the tropical oceans because they have less variability than polar regions. Also land areas are chosen less than adjacent ocean regions, and the ENSO region is not chosen because it shows variability that is irrelevant w.r.t. the anthropogenic forcing. With the increase in the causal regularization parameter, the contrast in the maps also increases, as the coefficients give more weight to regions that play a role in explaining the anthropogenic forcing, but not the volcanic forcing. We note here that this approach is not intended to find estimates of the historical radiative forcing. Instead, we find through regression a spatial pattern that captures the (linear) relationship between the temperature and the existing radiative forcing estimates, and subsequently (in future work) we will use these spatial patterns to predict the observed radiative forcing for detection and attribution.

Fig. 3 shows how detection and attribution work using an RCP run with the anthropogenic forcing as target variable. The signal is detected starting around 1990, i.e., the confidence intervals after this time don't contain zero anymore, and we can attribute the signal to the anthropogenic forcing because the true forcing (black) lies within the confidence intervals ($\pm 2\sigma$) of the predicted forcing (red). We compute the confidence intervals for each model separately by scaling the standard deviation of the residuals of the prediction for each value of the forcing by the standard deviation of the residuals for the corresponding scenario, i.e., RCP vs control runs. The confidence intervals are defined here with respect to the residuals of the prediction, and therefore they allow us to use them for hypothesis testing in detection and attribution. However we note that this definition of confidence intervals should not be confused with the standard definition in the statistical learning literature where the confidence intervals are defined with respect to the mean value of the predictions.

V. CONCLUSION

We have introduced a novel supervised statistical learning approach for studying the detection and attribution of climate change that protects against distributional changes in the external forcings. The class of distributions that we would like to protect ourselves against is generated by both interventions on the climate

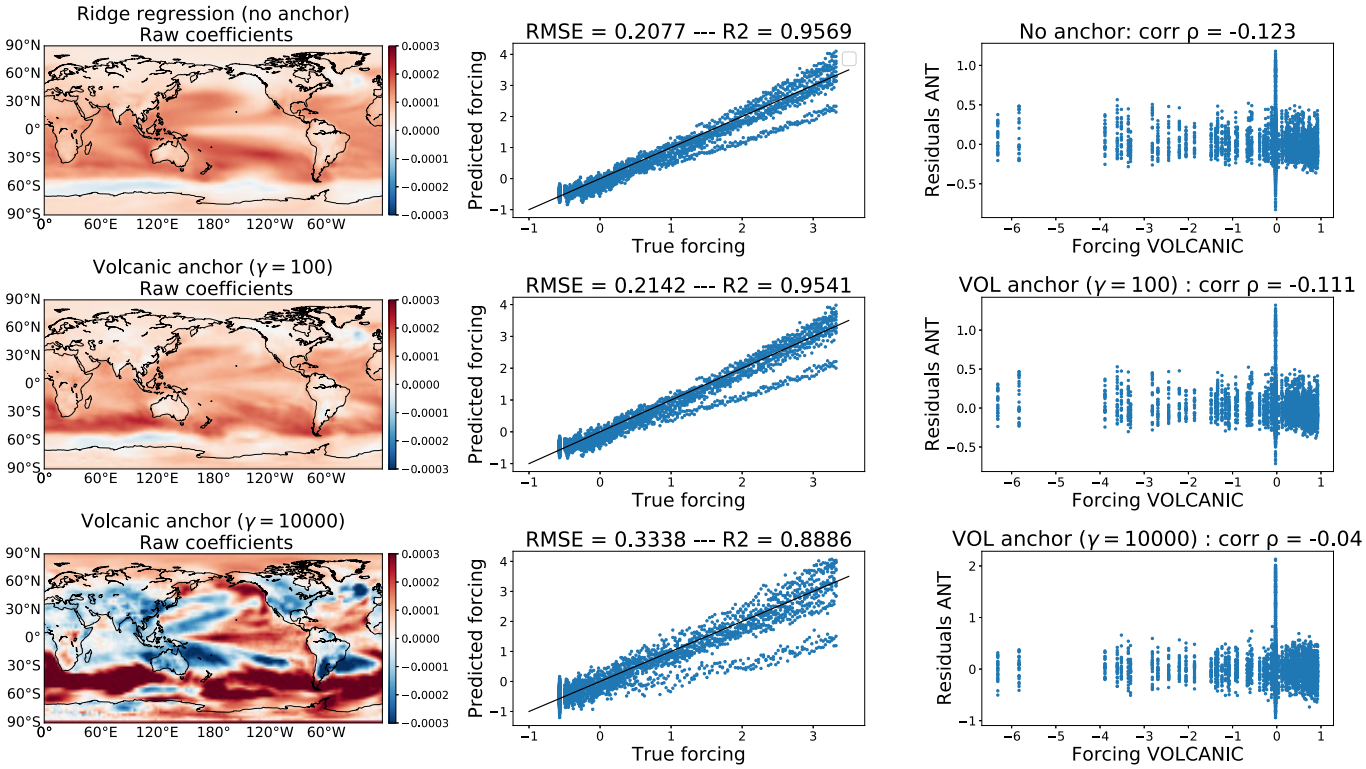


Fig. 2. Prediction of the anthropogenic forcing using anchor regression with the volcanic anchor. The first row shows the results for ridge regression, while the second and third row show results for anchor regression with two values of the causal regularization parameter γ.

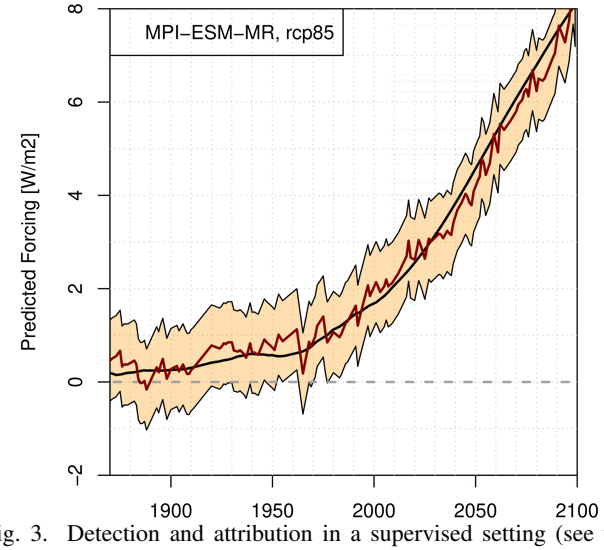


Fig. 3. Detection and attribution in a supervised setting (see text for details).

models and implicit shift interventions via the anchor method on the external forcings. In future work we plan to extend the framework to other forcings and go towards independence of the residuals with the anchor instead of just orthogonality. Another future direction is to incorporate temporal information into our framework, using for example Takens embedding (time-delay coordinates) [16], [17]. As the climate response to external forcings is not instantaneous, such information might help disentangle the different forcings which act on different timescales.

ACKNOWLEDGMENTS

This work was partly funded by the Swiss Data Science Center within the project "Data-science informed attribution of changes in the hydrological cycle" (DASH, C17-01). We thank Urs Beyerle for the preparation and maintenance of the CMIP5 data. We thank two anonymous reviewers for their valuable comments.

REFERENCES

[1] M. R. Allen and S. F. B. Tett, "Checking for model consistency in optimal fingerprinting," *Climate Dynamics*, vol. 15, no. 6, pp. 419–434, 1999.

[2] G. Hegerl and F. Zwiers, "Use of models in detection and attribution of climate change," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 2, pp. 570 – 591, 2011.

[3] N. L. Bindoff *et al.*, "Detection and attribution of climate change: from global to regional (IPCC - Ch. 10)," *Climate Change 2013: The Physical Science Basis*, 2013.

[4] A. Hannart, A. Carrassi, M. Bocquet, M. Ghil, P. Naveau, J. Ruiz, M. Pulido, and P. Tandeo, "Data assimilation for detection and attribution of weather and climate-related events," *Climatic Change*, vol. 136, pp. 155–174, 2016.

[5] E. Barnes, C. Anderson, and I. Ebert-Uphoff, "An AI approach to determining time of emergence of climate change," *Proceedings of the Eighth International Workshop on Climate Informatics (CI 2018)*, 2018.

[6] S. Sippel, N. Meinshausen, E. M. Fischer, E. Székely, and R. Knutti, "Climate change detected from today's global weather," *In review*, 2019.
[7] G. C. Hegerl, H. von Storch, K. Hasselmann, B. D. Santer, U. Cubasch, and P. D. Jones, "Detecting greenhouse-gas-induced climate change with an optimal fingerprint method," *Journal of Climate*, vol. 9, no. 10, pp. 2281–2306, 1996.
[8] B. D. Santer, S. Po-Chedley, M. D. Zelinka, I. Cvijanovic, C. Bonfils, P. J. Durack, Q. Fu, J. Kiehl, C. Mears, J. Painter, *et al.*, "Human influence on the seasonal cycle of tropospheric temperature," *Science*, vol. 361, no. 6399, p. eaas8806, 2018.
[9] G. D. Myhre *et al.*, "Anthropogenic and natural radiative forcing (IPCC - Ch. 8)," *Climate Change 2013: The Physical Science Basis*, 2013.
[10] N. Meinshausen, "Causality from a distributional robustness point of view," *IEEE Data Science Workshop*, pp. 6–10, 2018.
[11] P. Bühlmann, "Invariance, causality and robustness," *arXiv e-prints arXiv:1812.08233*, 2018.
[12] D. Rothenhäusler, N. Meinshausen, P. Bühlmann, and J. Peters, "Anchor regression: heterogeneous data meets causality," *arXiv e-print arXiv:1801.06229*, 2019.
[13] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*. Springer, 2 ed., 2009.
[14] K. E. Taylor, R. J. Stouffer, and G. A. Meehl, "An overview of cmip5 and the experiment design," *Bulletin of the American Meteorological Society*, vol. 93, no. 4, pp. 485–498, 2012.
[15] R. H. Moss, J. A. Edmonds, K. A. Hibbard, M. R. Manning, S. K. Rose, D. P. Van Vuuren, T. R. Carter, S. Emori, M. Kainuma, T. Kram, *et al.*, "The next generation of scenarios for climate change research and assessment," *Nature*, vol. 463, no. 7282, p. 747, 2010.
[16] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980*, vol. 898 of *Lecture Notes in Mathematics*, pp. 366–381, Berlin: Springer, 1981.
[17] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *J. Stat. Phys.*, vol. 65, no. 3–4, pp. 579–616, 1991.

UNSUPERVISED INPAINTING FOR OCCLUDED SEA SURFACE TEMPERATURE SEQUENCES

Yuan Yin[1*], Arthur Pajot[1], Emmanuel de Bézenac[1], Patrick Gallinari[1,2]

***Abstract*—Sea Surface Temperature (SST) data remotely measured from satellites are occluded by meteorological factors like clouds or rain. Classical SST reconstruction methods rely on interpolation or on model-based data assimilation requiring some knowledge about the underlying physical process. We propose a purely data-driven approach for the reconstruction of occluded SST zones. It is based on a spatio-temporal model of image sequence generation and it is implemented with generative adversarial networks (GANs). This approach is capable to complete the occluded areas without requiring any supervision or ground truth data. Our framework is evaluated on real SST data with simulated clouds.**

***Index Terms*—Sea surface temperature (SST), generative adversarial networks (GANs), image sequence inpainting, unsupervised learning.**

I. INTRODUCTION

Sea surface temperature (SST) is a key parameter for the modeling of weather and climate forecasts, ocean-atmosphere exchanges, tropical cyclones monitoring [1], etc. Most of the widely used satellite-derived SST sensing approaches are affected by meteorological perturbations. In particular, infrared (IR) radiation tracking cannot retrieve SST data from regions covered by clouds. The missing rate of IR-derived SST is up to 26-65% according to [2]. Reconstruction of SST missing values is important for geoscience applications and several approaches have been developed for filling the missing areas. They are often based on interpolation or data assimilation requiring explicit or implicit knowledge of the dynamical model of SST [3]–[5].

From a machine learning (ML) perspective, the problem shares common characteristics with image and video imputation. Generative models developed for learning data distributions, e.g. generative adversarial networks (GANs) [6], have motivated several developments for image or video reconstruction [7], [8]. However, most of these contributions are supervised and require the ground truth behind the missing part, which is unavailable for cloudy SST completion. Recently, unsupervised learning for data imputation using generative models has also motivated some work for still images [9]–[11]. We propose to extend these ideas to sequences and investigate the potential of generative models for completing cloud-occluded SST sequences without supervision. We thus develop a model-free unsupervised approach based on a spatio-temporal model of SST sequence generation. The completion or the reconstruction of dense image sequences is a new problem from an ML perspective. On the geoscience side, this is a new approach to this problem, and we will see that it compares well to existing methods.

After a brief state of the art in Section II, the model is introduced in Section III, experiments are detailed in Section IV.

Corresponding author: Y. Yin, yuan.yin@lip6.fr [1]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France [2]Criteo AI Lab, Paris, France [*]Work done while author was an intern at LIP6.

II. RELATED WORK

We briefly review different approaches developed in geoscience and ML for reconstructing SST images, and recent ML approaches for image and video imputation.

A. Cloud Removal for SST

Optimal Interpolation (OI), producing a linear estimate for the occluded area, is widely used in operational products [12]. Model-based assimilation methods, e.g. [3], rely on explicit physical dynamic priors with a significant computational cost. Data-driven methods based on empirical orthogonal functions (EOF) [13] use matrix factorization to achieve temporal interpolation. Recent advances in Analog Data Assimilation (AnDA) [5], [14] combine analog forecasting methods with data-driven assimilation using implicit knowledge of the dynamical prior. These methods rely either on interpolation or exploit some priors on the nature of the underlying process. Shibata et al. [15] propose to apply learning-based frame-level inpainting enhanced with optical flow using simple assumptions on pixel movement. In a later paper, [16], they recover the missing data using an adversarial approach.

B. Image Inpainting/Reconstruction

Since SST data can also be treated as image frames, pixel reconstruction can be considered as an inpainting problem. Early attempts are all supervised. [17] uses convolutional NNs for regressing observations to ground truth images. This typically produces blurry outputs. To overcome the issue, some authors introduce textures as in [18], while others make use of GANs [19], [20].

More recently, unsupervised approaches have been developed by considering only corrupted observations. [21], [22] show that it is possible to learn the underlying data distribution and reconstruct images from observations when the observation model is given or the noise is zero-mean. Bora et al. [9] introduce AmbientGAN to generate the missing part without supervision from corrupted data under the assumption that the stochastic observation process is known. Pajot et al. [10] propose to conditionally recover the corrupted image by solving a maximum *a posteriori* (MAP) estimation problem, which can be implemented with an adversarial framework without supervision.

C. Video Inpainting

Patch-based and object-based approaches for video inpainting were introduced before the deep learning era, but they generally rely on strong assumptions on videos, such as the reappearance of spatiotemporal patterns in occluded area or the prior segmentation of moving objects and background. More recently, flow-based methods have been used to model the spatial appearance and the local pixel movement between consecutive frames through neural optical flow estimation [8], [23]. There exist also several works extending image inpainting methods to video using 3D convolutional neural networks (NNs). They all require huge computational resources [24]. Wang et al. [7] propose frame-level generation decomposition by combining a video inpainter with a frame-wise refinement inpainter. All these methods are trained with supervision and have been developed for natural videos, with characteristics very different from dense SST image sequences.

III. METHOD

A. Problem Setting

We suppose that there exists an unknown original spatio-temporal sequence $\boldsymbol{x} \sim \mathrm{p}_{\boldsymbol{X}}, \boldsymbol{x} \in \mathbb{R}^{C\times T\times H\times W}$, where $\boldsymbol{x}$ is a fourth-order tensor denoting a C-channel sequence composed of T frames of $H \times W$ pixels. We denote $\boldsymbol{x}_t$ the t-th frame of the sequence and $\boldsymbol{x}_{t_1}^{t_2}$ the sub-sequence from the t_1-th to the t_2-th frame inclusive. With this notation, $\boldsymbol{x} \equiv \boldsymbol{x}_1^T$. We do not have access to the original signal $\boldsymbol{x}$ but only to corrupted observation sequences $\boldsymbol{y} \sim \mathrm{p}_{\boldsymbol{Y}}, \boldsymbol{y} \in \mathbb{R}^{C\times T\times H\times W}$. Our objective is to reconstruct $\boldsymbol{x}$ from the corresponding observation $\boldsymbol{y}$. For our application, $\boldsymbol{x}$ is the unknown sequence of SST images, and $\boldsymbol{y}$ is the cloud corrupted associated sequence of observations.

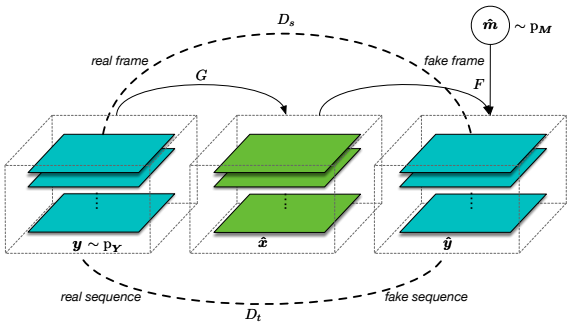


Fig. 1. Schema of our model. Generator G takes a sequence $\boldsymbol{y}$ and outputs an inpainted sequence $\hat{\boldsymbol{x}}$; measurement process F takes the inpainted sequence then outputs fake observations $\hat{\boldsymbol{y}}$.

We suppose that $\boldsymbol{y}$ is obtained from $\boldsymbol{x}$ via a measurement process modeled through a stochastic operator F as follows:

$$\boldsymbol{y} = F_{\boldsymbol{x}\sim \mathrm{p}_{\boldsymbol{X}}, \boldsymbol{m}\sim \mathrm{p}_{\boldsymbol{M}}}(\boldsymbol{x}, \boldsymbol{m}) = \boldsymbol{x} \odot \boldsymbol{m} + \mathbf{1} \cdot c \odot \bar{\boldsymbol{m}} \quad (1)$$

$\boldsymbol{m} \sim \mathrm{p}_{\boldsymbol{M}}$ is an occlusion mask, generated from a known distribution with the same size as $\boldsymbol{x}$ and with components in $\{0, 1\}$, where 0 holds for the masked pixel. $\bar{\boldsymbol{m}}$ denotes the complement of $\boldsymbol{m}$, $\odot$ is the element-wise multiplication, all the masked pixels are supposed to be reset to a constant c. In our experiments, we set $c = 1$. The random variables $\boldsymbol{X}, \boldsymbol{M}$ are assumed to be independent. Note that F is differentiable w.r.t. its first argument $\boldsymbol{x}$. We suppose that $\boldsymbol{m}$ can be retrieved from the observation $\boldsymbol{y}$, which is not very restrictive since the occlusion can be easily characterized in images.

Our objective is then to recover the sequence $\boldsymbol{x}$ from the observations $\boldsymbol{y}$ and the corresponding binary masks $\boldsymbol{m}$. Formally, we want to select a reconstruction $\boldsymbol{x}^*$ which is the most plausible under the posterior distribution $\mathrm{p}_{\boldsymbol{X}|\boldsymbol{Y}}(\cdot|\boldsymbol{y})$.

B. Model

We then formulate the problem as finding the most probable sequence conditioned on observations, i.e.:

$$\boldsymbol{x}^* = \arg\max_{\boldsymbol{x}} \log \mathrm{p}_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y}) \quad (2)$$

We introduce a parametric mapping $G : \boldsymbol{Y} \mapsto \boldsymbol{X}$ to infer $\boldsymbol{x}$ from $\boldsymbol{y}$, so that Equation 2 becomes:

$$G^* = \arg\max_{G} \mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\boldsymbol{Y}}}[\log \mathrm{p}_{\boldsymbol{X}}(G(\boldsymbol{y})) + \log \mathrm{p}_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|G(\boldsymbol{y}))] \quad (3)$$

We firstly handle the prior term $\mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\boldsymbol{Y}}}[\mathrm{p}_{\boldsymbol{X}}(G(\boldsymbol{y}))]$ in Equation 3. In order to make this term close to $\mathrm{p}_{\boldsymbol{X}}$, we will use an adversarial approach as in [9], [10]. Generator G will act as the inpainter to product a completed $\hat{\boldsymbol{x}} \equiv G(\boldsymbol{y})$ from true observations, then fake observations are generated through $\hat{\boldsymbol{y}} \equiv F(\hat{\boldsymbol{x}}, \hat{\boldsymbol{m}})$ with $\hat{\boldsymbol{m}} \sim \mathrm{p}_{\boldsymbol{M}}$. We will use two discriminators D_s and D_t respectively associated with the spatial features and the temporal dependence to optimize G. We found out that using two separated discriminators worked better than a unique spatio-temporal one. Spatial discriminator D_s will distinguish real frames $\boldsymbol{y}_t$ from fake ones $\hat{\boldsymbol{y}}_t$. Temporal discriminator D_t separates sequences $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$. The loss function used for training G, D_s, and D_t is:

$$\begin{aligned}\min_G \mathcal{L}(G) = \max_{D_s, D_t} \mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\boldsymbol{Y}}, \hat{\boldsymbol{y}}\sim \mathrm{p}^G_{\boldsymbol{Y}}} \\ [\log D_t(\boldsymbol{y}) + \log(1 - D_t(\hat{\boldsymbol{y}})) + \\ \frac{1}{T}\sum_{t=1}^{T} \log D_s(\boldsymbol{y}_t) + \log(1 - D_s(\hat{\boldsymbol{y}}_t))]\end{aligned} \tag{4}$$

with $\mathrm{p}^G_{\boldsymbol{Y}}(\boldsymbol{y}) \equiv \mathbb{E}_{\boldsymbol{m}\sim \mathrm{p}_{\boldsymbol{M}}, \boldsymbol{x}\sim \mathrm{p}^G_{\boldsymbol{X}}}[\mathrm{p}_{\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{M}}(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{m})]$, corresponding to the measurement operator F, and $\mathrm{p}^G_{\boldsymbol{X}}(\boldsymbol{x}) \equiv \mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\boldsymbol{Y}}}[\mathrm{p}^G_{\boldsymbol{X}|\boldsymbol{Y}}(\boldsymbol{x}|\boldsymbol{y})] \equiv \mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\boldsymbol{Y}}}[\delta(\|\boldsymbol{x} - G(\boldsymbol{y})\|)]$. Note that $\delta(\cdot)$ is the Dirac delta function, for point-to-point distribution matching. G is optimized by descending the loss and D_s, D_t by ascending it. In practice, discriminator D_s operates in the usual way while we force D_t to focus on temporal features by classifying real/generated difference pair ($\boldsymbol{\delta}_1^{T-1} \equiv \boldsymbol{y}_1^{T-1} - \boldsymbol{y}_2^{T}, \hat{\boldsymbol{\delta}}_1^{T-1} \equiv \hat{\boldsymbol{y}}_1^{T-1} - \hat{\boldsymbol{y}}_2^{T}$). The schema of our model is illustrated in Fig. 1.

For the conditional likelihood part $\mathbb{E}_{\boldsymbol{y}\sim \mathrm{p}_{\boldsymbol{Y}}}[\log \mathrm{p}_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|G(\boldsymbol{y}))]$, we make generator G to complete the occluded area and reuse the known values for the non-occluded area:

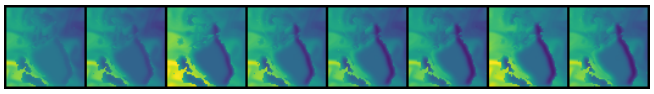
$$G(\boldsymbol{y}, \boldsymbol{m}) \equiv G_\phi(\boldsymbol{y}) \odot \bar{\boldsymbol{m}} + \boldsymbol{y} \odot \boldsymbol{m} \tag{5}$$

where G_ϕ maps an observation sequence to a reconstructed one.

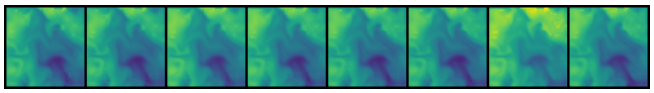
IV. EXPERIMENTS

A. SST Dataset

Our full SST dataset $\{\boldsymbol{x}^i\}_{i=1\ldots N}$ includes 2 subsets of GLOBAL Sea Physical Analysis and Forecasting Product from E.U. Copernicus Marine Service Information[1]. Training and validation set comes from the hourly SST in 2018 on a marine region of 64×64 pixels (20°-25.25°N, 34.75°-40°W, North Atlantic Ocean). Test set corresponds to the first 60 days of 2019 in another region of the same size (14.75°-20°S, 14.75°-20°W, South Atlantic Ocean). The frame rate is fixed to 6 hours per frame to make the movement of SST images more visible. Each sequence is composed of 24 frames.



(a) Generated sequence taken at the beginning of the training. MAE: 0.3081°C, RMSE: 0.3930°C, FID: 69.93, FVD: 279.79



(b) Generated sequence taken after 50 epochs. MAE: 0.0768°C, RMSE: 0.1127°C, FID: 8.84, FVD: 23.74

Fig. 2. Comparison of quality and scores of inpainted SST sequences. Scores are shown respectively under the examples. The lower are FID and FVD, the better is the quality of generated sequence.

B. Measurement Process: Cloud Simulation

The SST observation dataset $\{(\boldsymbol{y}^i, \boldsymbol{m}^i)\}_{i=1\ldots N}$ has been obtained by simulating clouds on the full SST data. This is obtained by applying a stochastic measurement operator F to the full dataset with mask sequences $\{\boldsymbol{m}^i\}_{i=1\ldots N}$ randomly chosen from the cloud dataset. The mask distribution is simulated using Liquid Water Path data (LWP) representing the total amount of liquid water present in the atmosphere between two points and measured in g/m^2. The LWP dataset is generated by PyCLES[2] [25], an open-source large eddy simulation (LES) system. It simulates clouds in 3D based on a variant of anelastic equations of atmospheric motion. The LWP dataset is the satellite view observation of the clouds. The binary masks are then obtained by marking pixels with 0 when their LWP value is above a threshold. We selected threshold values from 55 to 80 g/m^2 to simulate clouds at different occlusion rates. Some statistics about the occluded area at different thresholds are presented in Table I.

C. Network Architecture and Training

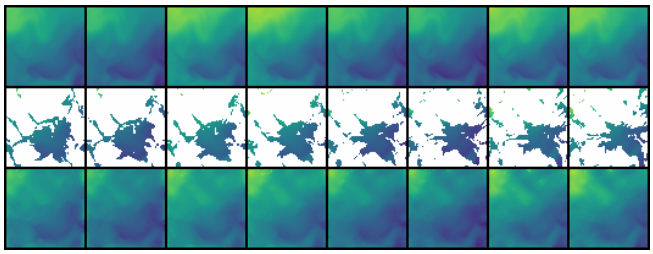
We briefly describe our model architecture and training setting. For full implementation details, please refer to the repository at https://github.com/yuan-yin/UNISST.

Our model utilizes a ResNet-type self-attention network [26] for the generator G, composed of 3D-ResNet blocks and spatial self-attention layers. Spatial discriminator D_s is a 2D convolutional NN for binary classification. Temporal discriminator D_t uses the same structure as D_s but with 3D convolutions.

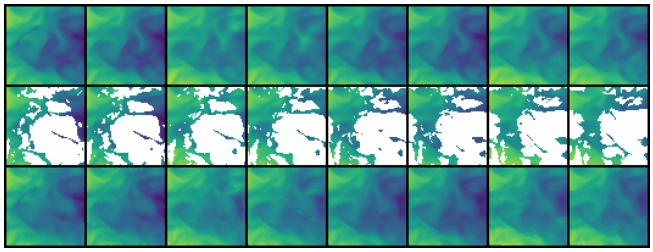
The model is trained on 300 sequences, validated on 66 sequences, and tested on 60 sequences. Each

[1]Provided online at http://marine.copernicus.eu.

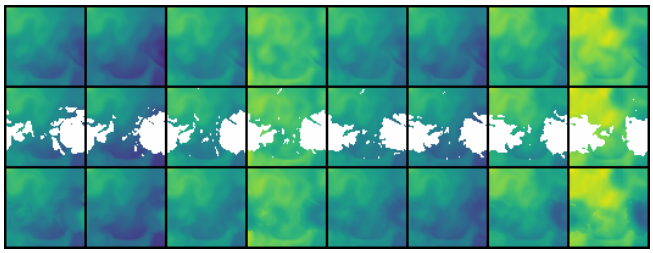
[2]https://github.com/pressel/pycles



(a) Cloud generated with LWP threshold 60 g/m^2



(b) Cloud generated with LWP threshold 70 g/m^2



(c) Cloud generated with LWP threshold 80 g/m^2

Fig. 3. Examples from test sets with cloud masks generated with different LWP thresholds. Each example from top to bottom: unknown ground truth, observation, and recovered sequence. Data are normalized to the same range of value for better visualization.

sequence is composed of 24 frames of 64×64 image. One can process sequences of any length. The number of frames is only limited GPU memory. We use SST data degraded by cloud masks at LWP threshold 70 g/m^2 for training since they include sufficient information for both SST and cloud dynamics. We apply hinge loss for Equation 4 as in [26]. All three networks are trained using Adam optimizer with a learning rate of 1×10^{-4} and $(\beta_1, \beta_2) = (0, 0.999)$ with batch size 1. All networks are initialized with normal distributions with a gain of 0.02. The experiments were made on one Nvidia GeForce GTX TITAN X GPU.

D. Evaluation Metrics

For the evaluation of reconstruction, we choose the Mean Average Error (MAE) and the Root Mean Square Error (RMSE) as metrics, which indicate the absolute and the standard deviation from the real data. They are calculated uniquely within the occluded area.

We use Fréchet Inception Distance (FID) [27] and Fréchet Video Distance (FVD) [28] to evaluate the quality of generated sequences. They compare the activation distribution of the generated samples from $p_{\boldsymbol{X}}^G$ to the real one sampled from $p_{\boldsymbol{X}}$. These distributions are extracted from activation layers of NNs, which are pre-trained on natural image tasks for FID and video tasks for FVD. The two distances are calculated for the whole sequence including occluded and non-occluded regions.

TABLE I
TEST RESULTS WITH CLOUDS GENERATED WITH DIFFERENT LWP THRESHOLDS

LWP (g/m^2)	Occluded Area (%)	MAE (°C)	RMSE (°C)	FID	FVD
55	79.9± 9.6	.1273±.0443	.1899±.0647	32.49	134.40
60	69.6±12.8	.1047±.0396	.1609±.0619	22.95	79.13
65	55.9±15.1	.0988±.0378	.1493±.0591	17.75	75.07
70	39.5±14.6	.0739±.0324	.1166±.0553	8.01	40.76
75	24.5±11.5	.0698±.0305	.1029±.0517	5.58	30.07
80	13.4± 7.8	.0497±.0237	.0763±.0420	1.77	9.89
All	47.1±11.9	.0874±.0347	.1327±.0558	14.76	61.55

TABLE II
COMPARISON WITH DINEOF ON TEST SET WITH LWP THRESHOLD 70 G/M^2

Method	MAE (°C)	RMSE (°C)	FID	FVD
DINEOF [13]	.1214±.0248	.1614±.0387	27.99	323.61
Ours	**.0739±.0324**	**.1166±.0553**	**8.01**	**40.76**

E. Results

Table I shows the results with SST data and clouds at different occlusion rates. For most of the occlusion rates, the generated sequences have an MAE under 0.1°C. They also have good FID and FVD values, which means that they are spatially and temporally realistic (See Fig. 3 for examples). For heavily occluded area recovery, our model can recover data around the border, but the performance in the core of clouds is impacted. Table II shows a comparison with a strong model-free baseline, DINEOF [13]. The error reduction w.r.t. DINEOF is about 40% for MAE and 30% for RMSE.

V. CONCLUSION

We proposed a GAN-based framework for occluded SST sequence inpainting. Our model utilizes an inpainter to complete the occluded sequences with the help of two discriminators classifying real and generated observation sequences. We show that it is possible to reconstruct SST sequences without ground truth supervision when we have a stochastic model of the occlusion process. The results show that even with heavy occlusion, the sequences completed by our approach are realistic. Our model could be further tested using non-simulated

stochastic operator with cloud masks extracted from satellite images as in [29] or cloud data service [30].

Acknowledgments

This work was partially funded by ANR project LOCUST - ANR-15-CE23-0027 and by CLEAR - Center for LEArning & data Retrieval - joint lab. With Thales (www.thalesgroup.com).

REFERENCES

[1] C. J. Donlon, P. J. Minnett, C. Gentemann, T. J. Nightingale, I. J. Barton, B. Ward, and M. J. Murray, "Toward improved validation of satellite sea surface skin temperature measurements for climate research," *Journal of Climate*, vol. 15, no. 4, pp. 353–369, 2002.

[2] L. Guan and H. Kawamura, "Sst availabilities of satellite infrared and microwave measurements," *Journal of Oceanography*, vol. 59, pp. 201–209, Apr 2003.

[3] C. Ubelmann, P. Klein, and L.-L. Fu, "Dynamic interpolation of sea surface height and potential applications for future high-resolution altimetry mapping," *Journal of Atmospheric and Oceanic Technology*, vol. 32, no. 1, pp. 177–184, 2015.

[4] D. Sirjacobs, A. Alvera-Azcárate, A. Barth, G. Lacroix, Y. Park, B. Nechad, K. Ruddick, and J.-M. Beckers, "Cloud filling of ocean colour and sea surface temperature remote sensing products over the southern north sea by the data interpolating empirical orthogonal functions methodology," *Journal of Sea Research*, vol. 65, no. 1, pp. 114 – 130, 2011.

[5] R. Lguensat, P. Tandeo, P. Ailliot, M. PULIDO, and R. Fablet, "The Analog Data Assimilation," *Monthly Weather Review*, vol. 145, pp. 4093 – 4107, Oct. 2017.

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27* (Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, eds.), pp. 2672–2680, Curran Associates, Inc., 2014.

[7] C. Wang, H. Huang, X. Han, and J. Wang, "Video inpainting by jointly learning temporal structure and spatial details," *CoRR*, vol. abs/1806.08482, 2018.

[8] R. Xu, X. Li, B. Zhou, and C. C. Loy, "Deep flow-guided video inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] A. Bora, E. Price, and A. G. Dimakis, "AmbientGAN: Generative models from lossy measurements," in *International Conference on Learning Representations*, 2018.

[10] A. Pajot, E. de Bézenac, and P. Gallinari, "Unsupervised adversarial image reconstruction," in *International Conference on Learning Representations*, 2019.

[11] S. C. Li, B. Jiang, and B. M. Marlin, "MisGAN: Learning from incomplete data with generative adversarial networks," *CoRR*, vol. abs/1902.09599, 2019.

[12] C. J. Donlon, M. Martin, J. Stark, J. Roberts-Jones, E. Fiedler, and W. Wimmer, "The operational sea surface temperature and sea ice analysis (OSTIA) system," *Remote Sensing of Environment*, vol. 116, pp. 140 – 158, 2012. Advanced Along Track Scanning Radiometer (AATSR) Special Issue.

[13] J. M. Beckers and M. Rixen, "Eof calculations and data filling from incomplete oceanographic datasets," *Journal of Atmospheric and Oceanic Technology*, vol. 20, no. 12, pp. 1839–1856, 2003.

[14] R. Fablet, P. Huynh Viet, R. Lguensat, P.-H. Horrein, and B. Chapron, "Spatio-temporal interpolation of cloudy SST fields using conditional analog data assimilation," *Remote Sensing*, vol. 10, no. 2, 2018.

[15] S. Shibata, M. Iiyama, A. Hashimoto, and M. Minoh, "Restoration of sea surface temperature images by learning-based and optical-flow-based inpainting," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 193–198, July 2017.

[16] S. Shibata, M. Iiyama, A. Hashimoto, and M. Minoh, "Restoration of sea surface temperature satellite images using a partially occluded training set," in *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 2771–2776, Aug 2018.

[17] J. Xie, L. Xu, and E. Chen, "Image denoising and inpainting with deep neural networks," in *Advances in Neural Information Processing Systems 25* (F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, eds.), pp. 341–349, Curran Associates, Inc., 2012.

[18] C. Yang, X. Lu, Z. Lin, E. Shechtman, O. Wang, and H. Li, "High-resolution image inpainting using multi-scale neural patch synthesis," *CoRR*, vol. abs/1611.09969, 2016.

[19] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," *CoRR*, vol. abs/1604.07379, 2016.

[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," *CoRR*, vol. abs/1801.07892, 2018.

[21] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Deep image prior," *CoRR*, vol. abs/1711.10925, 2017.

[22] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila, "Noise2Noise: Learning image restoration without clean data," *CoRR*, vol. abs/1803.04189, 2018.

[23] D. Kim, S. Woo, J.-Y. Lee, and I. S. Kweon, "Deep video inpainting," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[24] Y. Chang, Z. Y. Liu, K. Lee, and W. Hsu, "Free-form video inpainting with 3D gated convolution and temporal PatchGAN," *CoRR*, vol. abs/1904.10247, 2019.

[25] K. G. Pressel, C. M. Kaul, T. Schneider, Z. Tan, and S. Mishra, "Large-eddy simulation in an anelastic framework with closed water and entropy balances," *Journal of Advances in Modeling Earth Systems*, vol. 7, no. 3, pp. 1425–1456, 2015.

[26] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, (Long Beach, California, USA), pp. 7354–7363, PMLR, 09–15 Jun 2019.

[27] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, G. Klambauer, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a nash equilibrium," *CoRR*, vol. abs/1706.08500, 2017.

[28] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "Towards accurate generative models of video: A new metric & challenges," *CoRR*, vol. abs/1812.01717, 2018.

[29] P. Singh and N. Komodakis, "Cloud-GAN: Cloud removal for Sentinel-2 imagery using a cyclic consistent generative adversarial networks," in *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1772–1775, July 2018.

[30] A. H. Young, K. R. Knapp, A. Inamdar, W. Hankins, and W. B. Rossow, "The international satellite cloud climatology project

H-series climate data record product," *Earth System Science Data*, vol. 10, no. 1, pp. 583–593, 2018.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

IMPROVING WEATHER AND CLIMATE PREDICTIONS BY TRAINING OF SUPERMODELS

Francine Schevenhoven[1,2], Frank Selten[3], Alberto Carrassi[4,1,2], Noel Keenlyside[1,2]

***Abstract*—Recent studies demonstrate that weather and climate predictions potentially improve by dynamically combining different models into a so called "supermodel". In this study we use a weighted supermodel - the supermodel's time derivative is a weighted superposition of the time-derivatives of the imperfect models. The weights of the supermodel are trained on the basis of past observations. Here we apply two different learning methods to a supermodel of up to four different versions of the global atmosphere-ocean-land model SPEEDO. The standard version is regarded as truth. The first training method is based on an idea called Cross Pollination in Time (CPT), where models exchange states during the training. The second method is a synchronization based learning rule, originally developed for parameter estimation. We demonstrate that both training methods produce climate simulations and weather predictions of superior quality than the individual model versions. Supermodel predictions also outperform predictions based on the commonly used Multi-Model Ensemble (MME) mean. In principle the proposed training schemes are applicable to state-of-the-art models and historical observations.**

I. INTRODUCTION

A. Multi-model-ensemble

Although weather and climate models continue to improve, they will inevitably remain imperfect [1]. Nevertheless, with the best possible models in hands, more accurate predictions can be obtained by making good use of all of them thus exploiting multi-model information. In order to reduce the impact of model errors on predictions, it is common practice to combine the predictions of a collection of different models in a statistical fashion. This is referred to as the Multi-Model Ensemble (MME) approach: the MME mean prediction is often more skillful as model errors tend to average out [2], whereas the spread between the model predictions is naturally interpreted as a measure of the uncertainty about the mean [3]. Although MME generally improves predictions of climate statistics (i.e. mean and variance), a major drawback is that it is not designed to produce an improved trajectory that can be seen as a specific climate forecast, given that averaging uncorrelated climate trajectories from different models leads to variance reduction and smoothing.

Corresponding author: F Schevenhoven, francine.schevenhoven@uib.no [1]Geophysical Institute, University of Bergen, Bergen, Norway [2]Bjerknes Centre for Climate Research, Bergen, Norway [3]Royal Netherlands Meteorological Institute, De Bilt, The Netherlands [4]Nansen Environmental and Remote Sensing Center, Bergen, Norway

B. Supermodeling

Improving the trajectory of a model is precisely what supermodeling attempts to achieve [4]. In a supermodel, different models exchange information during the simulation at every time step and form a consensus on a single best prediction, hence the models are synchronized. Because of the synchronization, a supermodel will not suffer from variance reduction and smoothing as with the MME approach. Nevertheless, before supermodeling becomes suitable for the class of large dimensional state-of-the-art weather and climate models, we need to rely upon training schemes that are computationally suitable for that context. In this paper we apply and compare CPT and the synch rule to train a weighted supermodel based on the intermediate complex global coupled atmosphere-ocean-land model SPEEDO [5].

C. Weighted supermodeling

A weighted supermodel based on $N = 2$ imperfect models with parametric error is given by

$$\dot{\mathbf{x}}_1 = \mathbf{f}(\mathbf{x}_s, \mathbf{p}_1) \tag{1a}$$

$$\dot{\mathbf{x}}_2 = \mathbf{f}(\mathbf{x}_s, \mathbf{p}_2) \tag{1b}$$

$$\dot{\mathbf{x}}_s = \mathbf{W}_1\dot{\mathbf{x}}_1 + \mathbf{W}_2\dot{\mathbf{x}}_2, \tag{1c}$$

where $\mathbf{x}_s \in \mathbb{R}^n$ represents the supermodel state vector whereas diagonal matrices $\mathbf{W}_i = \text{diag}(\mathbf{w}_i)$ with $\mathbf{w}_i \in \mathbb{R}^n$, $i \in \{1, 2\}$ denote the weights. In a weighted supermodel the states are imposed to be perfectly synchronized. Training a weighted supermodel implies training the weights $\mathbf{w}_i$.

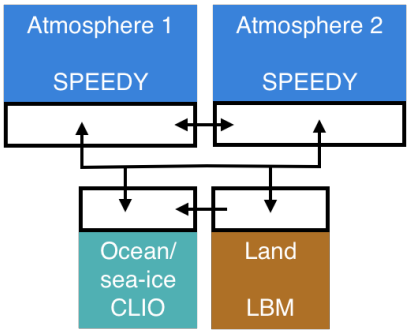


Fig. 1. Adapted from [6]. Schematic representation of the SPEEDO climate supermodel based on two imperfect atmosphere models. The two atmosphere models exchange water, heat and momentum with the perfect ocean and land model. The ocean and land model send their state information to both atmosphere models. The atmosphere models exchange state information in order to combine their time-derivatives.

II. SPEEDO CLIMATE MODEL

The SPEEDO global climate model consists of an atmospheric component (SPEEDY) that exchanges information with a land (LBM) and an ocean-sea-ice component (CLIO) using coupling routines. The atmospheric model SPEEDY describes the evolution of the two horizontal wind components U (east-west) and V (north-south), temperature T, and specific humidity q at eight vertical levels and the surface pressure p_s. Relatively simple calculations of heating and cooling rates due to radiation, convective transports, cloud amounts, precipitation and turbulent heat, water and momentum exchange at the surface are performed at a computational grid of approximately 3.75 degree horizontal spacing (48x96 grid cells). A detailed description of SPEEDO can be found in [5] and [8].

A. SPEEDO supermodel

The training experiments of this study are evaluated in a noisy free observation framework, with perfect observations generated by sampling a reference model trajectory. This "perfect model" provides a set of time-ordered observations, called the "truth". We consider the SPEEDO climate model with standard parameter values as perfect model (specified later) and create imperfect models by perturbing parameter values in the atmospheric component. A supermodel is formed by combining the imperfect atmosphere models through a weighted superposition of the time derivatives of the imperfect models (Eq. 1) which are each coupled to the same ocean and land model (Fig. 1).

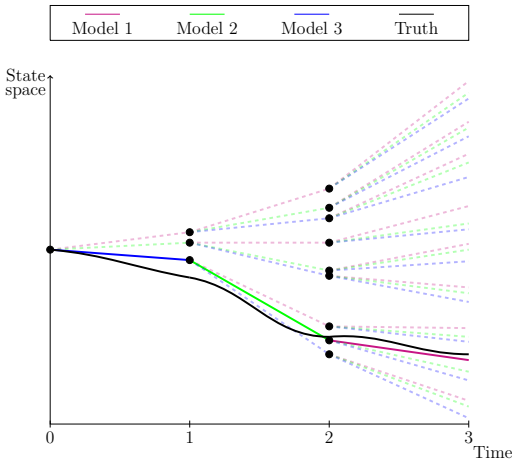


Fig. 2. Adapted from [6]. A one dimensional schematic of a pruned CPT ensemble for 3 models. Note that the *truth* has been drawn here as a continuous line for illustrative purpose. In practice the *truth* is only known at discrete times (the observation times) and the distance with respect to model trajectories is computed at those times only.

III. LEARNING METHODS

Two different learning strategies are evaluated in this study in order to train the SPEEDO weighted supermodel: (i) learning based on CPT as developed and applied to low-order dynamical systems in [7], and (ii) learning based on synchronization as applied to a connected SPEEDO supermodel in [8].

A. Cross pollination in time

The Cross Pollination in Time (CPT) learning approach is based on an idea proposed by [9]. CPT "crosses" trajectories of different models in order to create a larger solution space. The aim is to generate trajectories that follow the truth more closely. The training phase of CPT starts from an observed initial condition in state space. For simplicity assume the model is one dimensional. From the same initial state, the imperfect models compute one time step each ending in a different state. Next, all models compute another time step from each of these new states. Continuing this process leads to a rapid increase in the number of trajectories with time that will ultimately cover a larger area of the state space. Among the full set of mixed trajectories, the one which is closest to the truth (i.e. to the data) is continued, the others are discarded resulting in a pruned ensemble, as is depicted in Fig. 2.

The training period is terminated when the CPT trajectory starts to deviate from the truth beyond a given pre-specified threshold. After training an optimal

trajectory is obtained that is produced by a combination of different imperfect models. Next we count how often during training each model has produced the best prediction of a particular component of the state vector. This frequency of occurrences is used to compute weights **W** for the corresponding time-derivative of the state vector. The superposition of weighted imperfect models forms a weighted supermodel, as expressed in the example of Eq. 1.

B. Synchronization based learning

For the training of a supermodel based on synchronization a learning rule (the synch rule) is used that updates the weights such that synchronization errors between truth and supermodel are minimized. In contrast to CPT learning, initial values for the weights need to be chosen and the weights are updated during training. Under certain conditions the supermodel will fall into synchronized motion with the truth as the weights are updated and the supermodel is nudged to the truth.

The synch rule for the weights is an application of the general synchronization based parameter estimation approach suggested in [10]:

$$\dot{\mathbf{w}}_{kj} \quad = \quad -\delta_j \mathbf{e}_j \mathbf{f}_{kj} \tag{2}$$

Here $\mathbf{w}_{kj}$ denotes the weight for model k corresponding to variable j, $\mathbf{e}_j$ is the difference between the supermodel and the truth, $\mathbf{f}_{kj}$ the time derivative of imperfect model k and δ_j an adjustable rate of learning scaling factor.

C. Construction of imperfect models

The perturbed parameters in this study are the convection relaxation timescale, the relative humidity threshold and the momentum diffusion timescale. The reason to perturb these parameters is because the uncertainty in climate models mostly lies in the parameterization of clouds and convection, and perturbing these parameters in the SPEEDO model results in a spread in the simulated climate that characterizes this uncertainty. The parameters are listed in Table I.

The first supermodel will consist of a weighted superposition of models 1 and 2. The second supermodel will consist of a weighted superposition of models 1,3,4 and 5. The parameter values of these models are chosen such that they form a so-called convex hull around the true parameter values (see [7] for a discussion on the convex hull principle). A convex hull implies that, provided the model functional dependence on the parameter is linear, the true parameter values can be obtained as a linear combination with positive weights of the parameter values of the imperfect models. Note that we use only two perturbed values for each parameter, the imperfect models differ only in the combination of these values.

TABLE I
PARAMETER VALUES OF PERFECT AND IMPERFECT MODELS.

model	convection relaxation timescale	relative humidity threshold	momentum diffusion timescale
perfect	6 hours	0.9	24 hours
model 1	4 hours	0.85	18 hours
model 2	8 hours	0.95	30 hours
model 3	4 hours	0.95	30 hours
model 4	8 hours	0.95	18 hours
model 5	8 hours	0.85	30 hours

For both CPT and the synch rule we choose to work with global weights, which means that for each meteorological variable we use the same weight at every grid point. Best results were obtained by exchanging the state information of temperature, vorticity and divergence between the models.

IV. RESULTS

For the synch rule it takes the weights approximately one year to converge, hence we choose to have a training period of one year. For CPT the weights converge already within one week, but in order to check the difference between the CPT weights during a year, the CPT method is applied for each week during one year. After each week, the values for all prognostic variables are reset to the truth, and the procedure is repeated.

A. Experiment 1

After training the imperfect models and the supermodel are integrated for 40 years in time, starting from January 1st of model year 2001. The climatology is defined as the average over years 11-40. Global mean time-series for surface air temperature, precipitation, wind, surface solar radiation and cloud cover for the different models demonstrate that both weighted supermodels behave very similar and remain close to the perfect model (not shown). We computed an optimal weighted average of the climatology of both imperfect models (optimal in the sense that the RMSE in the climatology is minimized) as in [8]. This MME mean climatology has errors in the same order of magnitude as both trained weighted supermodels due to fact that the imperfect model errors are near-mirror images of each other.

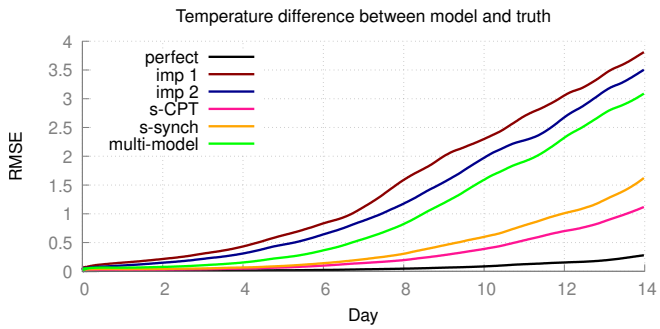


Fig. 3. Adapted from [6]. Forecast quality as measured by the root-mean-squared error (RMSE) of the truth and a model with a perturbed initial condition. The control is the difference between the perfect model and the perfect model with a perturbed initial condition.

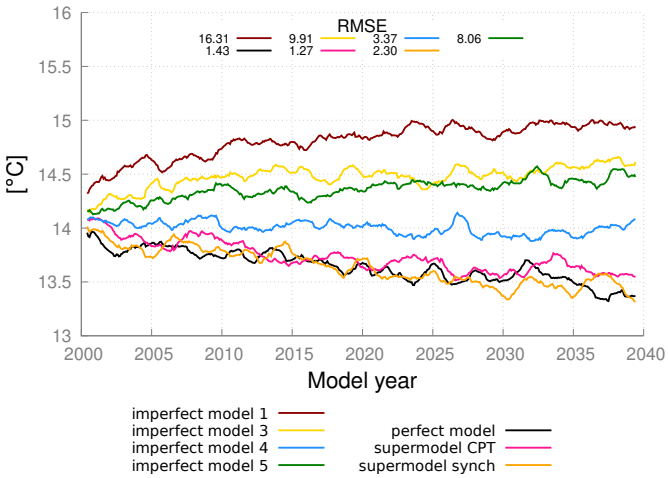


Fig. 4. Adapted from [6]. Global mean temperature. The normalized RMSE in the climatology of model years 2011-2040 with respect to the climatology of the truth.

In order to assess the quality of short-term forecasts, we initialized the various models from slightly perturbed states of the truth and integrated the models for two weeks. For comparison we computed the forecast error of the weighted mean forecast of both imperfect models using the same weights as in the calculation of the optimal climatology. This MME mean forecast has smaller forecast errors than the imperfect models, yet both supermodels are clearly superior. Figure 3 shows the RMSE for surface air temperature.

B. Experiment 2: convex hull principle

This experiment (Fig. 4) demonstrates the potential of supermodels to mitigate common errors, and thereby clearly outperform the standard MME-approach. Since all imperfect models overestimate the global average temperature, a standard weighted MME-approach results in a climatological forecast worse than the best imperfect model. Even though the parameters do not appear linearly in the equations, it seems the convex hull approach works well. Furthermore this experiment shows that by combining imperfect models dynamically we can not only in the short-term error, but also for the climatology surpass the multi-model mean.

V. SUMMARY AND CONCLUSION

We have demonstrated the potential of weighted supermodeling to improve weather and climate predictions using the global coupled atmosphere-ocean-land model SPEEDO in the presence of parametric error. The weights are trained using data from the perfect model using two different training schemes having low computational cost. The first method is based on Cross Pollination in Time (CPT), where different model trajectories are "crossed" in order to create a larger ensemble of possible trajectories. The second method is a synchronization based learning rule (synch rule), which adapts the weights of the different imperfect models during training such that the supermodel synchronizes with the perfect model.

Both training methods yield supermodels that outperform the individual imperfect models, in short-term forecasts as well as in long-term climate simulations. In addition, supermodels can outperform results from the standard MME-approach. CPT training required shorter training periods (one week as opposed to a year for the synch rule). An advantage of the synch rule is that it could allow for negative weights that can potentially improve the weighted supermodel in case model errors do not compensate for positive weights (not shown, see [6]).

The ultimate goal of our research is to apply supermodeling to realistic climate models. But, will it work? The real world is not simply a perturbed parameter version of these complex models. It is essential to extend the approach to other sources of model error towards the application with real climate models. Secondly, in this study observations were assumed to be perfect, i.e. to be complete and noisy free. To use real data it will thus be necessary to extend the supermodel training methods to account for the observational noise. These latter problems are the subjects of ongoing research of scientists which are making use of ideas and techniques from data assimilation. Data assimilation based supermodeling is also envisionable to account for generic source of model error in the construction of the supermodel, and it will be the subject of future research.

This publication is partially reprinted from the EGU Journal of Earth System Dynamics: Schevenhoven, F., Selten, F., Carrassi, A., and Keenlyside, N.: Improving weather and climate predictions by training of supermodels, Earth Syst. Dynam. Discuss., https://doi.org/10.5194/esd-2019-32, in review, 2019.

REFERENCES

[1] P. Bauer, A. Thorpe, and G. Brunet, "The quiet revolution of numerical weather prediction," *Nature*, vol. 525, pp. 47 EP –, 09 2015.

[2] A. P. Weigel, M. A. Liniger, and C. Appenzeller, "Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts?," *Quarterly Journal of the Royal Meteorological Society*, vol. 134, pp. 241–260, Jan. 2008.

[3] IPCC, *Climate Change 2013: The Physical Science Basis*. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.) Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp], 2013.

[4] L. A. van den Berge, F. M. Selten, W. Wiegerinck, and G. S. Duane, "A multi-model ensemble method that combines imperfect models through learning," *Earth System Dynamics*, vol. 2, no. 1, pp. 161–177, 2011.

[5] C. A. Severijns and W. Hazeleger, "The efficient global primitive equation climate model speedo v2.0," *Geoscientific Model Development*, vol. 3, no. 1, pp. 105–122, 2010.

[6] F. Schevenhoven, F. Selten, A. Carrassi, and N. Keenlyside, "Improving weather and climate predictions by training of supermodels," *Earth System Dynamics Discussions*, vol. 2019, pp. 1–32, 2019.

[7] F. J. Schevenhoven and F. M. Selten, "An efficient training scheme for supermodels," *Earth System Dynamics*, vol. 8, no. 2, pp. 429–438, 2017.

[8] F. M. Selten, F. J. Schevenhoven, and G. S. Duane, "Simulating climate with a synchronization-based supermodel," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 12, p. 126903, 2017.

[9] L. A. Smith, *Nonlinear Dynamics and Statistics*, ch. Disentangling Uncertainty and Error: On the Predictability of Nonlinear Systems, pp. 31–64. Boston, MA: Mees, Alistair I., Birkhäuser Boston, 2001.

[10] G. S. Duane, D. Yu, and L. Kocarev, "Identical synchronization, with translation invariance, implies parameter estimation," *Physics Letters A*, vol. 371, no. 5–6, pp. 416 – 420, 2007.

CAUSAL LINK DETECTION AND THE PREDICTION OF THE INDIAN SUMMER MONSOON

Moumita Saha, Dhanendra Soni, Brandon Finley, Claire Monteleoni

***Abstract*—Indian summer monsoons are complex phenomena influenced by different climatic variables at geographical distances. A reliable prediction of monsoon in advance is vital for the economic development of a country and combating extreme conditions. A technique, called causal discovery, is used to identify the regions over the Pacific Ocean influencing the Indian summer monsoons. The PC Algorithm aids in learning the links between the sea surface temperature of the Pacific Ocean and Indian monsoon at lead months. Identified causes are considered to predict both the aggregate and regional Indian monsoons at a lead of three months with a random forest model. The model predicts the aggregate Indian monsoon with a mean absolute error of 5.2% as well as four regional monsoons with mean absolute errors of 5.8%, 6.7%, 6.9%, and 5.1%. The predictions are found to be superior to those of the Niño 3.4 index and comparable to the forecasts by the India Meteorological Department.**

I. MOTIVATION

Indian summer monsoons are critical climatic phenomena. The prediction of monsoons is challenging for their chaotic nature and high inter-annual variability. A variation of 10% on either upper or lower values from the long period average (LPA) rainfall can breed extreme events like floods or droughts. The precipitation occurring during the June-September time frame in India comprises more than 75% of annual rainfall and is called the Indian summer monsoon. The summer monsoon is vital for fresh-water renewal, hydro-electricity generation, agricultural productivity, and nourishment of the flora-fauna in the country. The prediction of monsoons in advance is critical for designing proper agricultural policies, combating the extremes, and thus, overall economic development of the country.

The term, causality, refers to the phenomenon where a previous state contributes to the occurrence of a future state. The former state is called the 'cause,' whereas the latter is called the 'effect.' A cause can be responsible for a multitude of effects, and similarly, an effect can be the result of numerous causes. The study of these cause-effect relationships is significant for interpreting the underlying complexities of different climatic phenomena. Granger introduced one of the most preliminary and essential concepts in causality [1]. Granger causality is defined as follows: a time series, C, Granger-caused another series, E, if the past values of C assists in predicting the present values of E more accurately than only using the previous values of E. The introduction to causal calculus during 1980 strengthed the concept of causality [2]. Graphical models were introduced by Pearl [3] to represent probabilistic-independence relationships between different variables. Sprites and Glymour [4] assisted in detecting the hidden-common causes, which further helped in causal interpretation of the graph structures. Ebert-Uphoff and Deng [5] elaborated on different causal discovery methods for climatic events with a focus on constraint-based structure learning. Strenberg [6] investigated the presumed causal links between drought and severe winter conditions in Mongolia and found a contradiction in the linkage between the extreme conditions. Causal relationships between El Niño Southern Oscillation (ENSO) and the Indian Ocean dipole were studied using a future simulation of the CMIP5 model [7]. Jebli and Hadhri [8] performed a study on examining the dynamic, causality relationships between tourism and carbon-dioxide emission via transportation, gross domestic productivity, and energy usage. Thus, we observed that causality exists between various climatic variables and phenomena, which makes causal discovery a critical tool to uncover the underlying association and understand the complex climatic phenomenon.

Indian summer monsoons are widely studied [9], [10], [11] for understanding and unraveling their composite mechanism. The influence of different climatic indexes like ENSO or the Indian Ocean dipole on Indian monsoons is also investigated [12], [13]. Various machine learning methods are used to predict the monsoon rainfall at a lead time for aiding in policy

Corresponding author: M. Saha, moumita.saha@colorado.edu, Department of Computer Science, University of Colorado, Boulder

design and maximizing agricultural productivity [12], [14], [15]. The ENSO event over the equatorial Pacific Ocean influences several climatic phenomena across the globe (including Indian summer monsoons) [12].

In this study, we aim to identify the causal links between the sea surface temperature of the Pacific Ocean and the Indian monsoon. Causal link detection helps to consider the temporal lag and also adds interpretability to the predictor variables affecting the monsoon phenomenon. The motivation of the work is two-fold–(i) causal discovery between the sea surface temperature of the Pacific Ocean and Indian monsoons phenomena (occurring during June-September), and (ii) predictions of the rainfall at a lead time with the identified causes using a random forest prediction model. We studied the monsoons for both aggregate India as well as its regional parts. The study is performed months in advance, and predictions of monsoon rainfall by the proposed approach are compared with that by the Niño 3.4 index and India Meteorological Department models.

II. METHOD

The proposed approach to the identification of causal links between the Pacific Ocean and Indian monsoons, and thereby, predictions of the Indian summer monsoons, is shown in Figure 1.

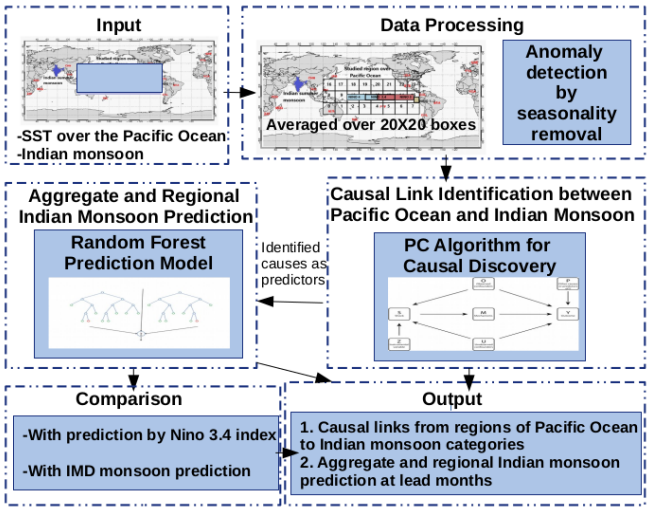


Fig. 1. Identification of causal links between the Pacific Ocean and Indian monsoons, and the prediction of Indian summer monsoons

A. Data Processing

The studied region over the Pacific Ocean (160°E to 80°W and 30°N to 30°S) is modularized into boxes of dimension 20° latitude × 20° longitude. This region encompasses a much broader region of the ocean including all of the Niño indexes (the Niño 1+2, the Niño 3, the Niño 4, and the Niño 3.4). Hence, we call them together as the Niño 1-4 indexes. The choice of this region assists in learning the influence of both the close and far away regions from the Niño 1-4 indexes on the Indian monsoons. The studied ocean region is comprised of 24 boxes (each of dimension 20° × 20°), as shown in Figure 2.

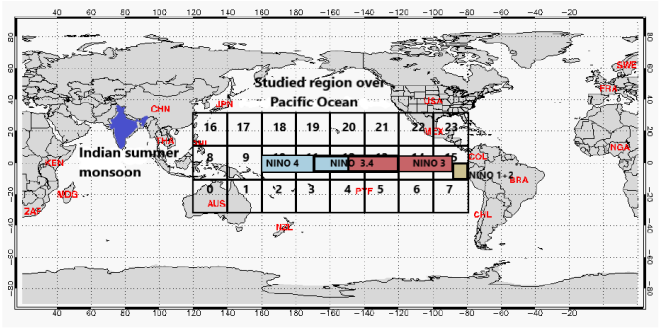


Fig. 2. Studied regions as distinct boxes over the Pacific Ocean (the color highlights the regions of the Niño 1-4 indexes)

We considered the sea surface temperature (SST) of the ocean for detecting the causal links. SST monthly data is collected from NOAA NCDC ERSST version3b [16], available at 2°× 2° resolution. The data is collected for 1951-2018 and processed by removing the seasonality, as shown in Equation 1.

$$\text{Process\_Temp}_{\text{m}}^{\text{y}} = \text{Temp}_{\text{m}}^{\text{y}} - \text{mean}(\text{Temp}_{\text{m}}), \qquad (1)$$

where, $\text{Temp}_{\text{m}}^{\text{y}}$ is the SST for the m^{th} month of the y^{th} year, and $\text{mean}(\text{Temp}_{\text{m}})$ denotes the average of the m^{th} month's SST over thirty years (1951–1980).

We represented each box (with a dimension of 20° × 20°) with the average SST of all the data points lying inside the box. These boxes are considered to be nodes in the causal graph for link discoveries. We divided the data into two sets–(i) training (1951-2007), and (ii) test (2008-2018). We used the training set for the identification of causal links between the Pacific Ocean and the Indian monsoons. We used the same set for the training of the random forest model with the detected causes acting as predictors. Finally, we used the test set for evaluating the accuracy of the monsoon predictions with the identified causes.

We collected the rainfall data from the India Meteorological Department for 1951-2018. We considered the rainfall for the aggregate Indian region (AI) and its four homogeneous regions: central (CI), north-east (NEI), north-west (NWI), and south-peninsular India (SPI) [17]. The standard deviation of rainfall over the regional

parts are higher than aggregate India, and thus, they are more challenging to forecast.

B. Causal Link Detection

The boxes over the Pacific Ocean (24 boxes) and the five rainfall categories (AI, CI, NEI, NWI, and SPI) are considered to be nodes in the causal graph structure. The primary aim of this study is to identify the causal links between the Pacific Ocean regions and the Indian monsoons. The motivation of using causal discovery includes considering the temporal lag to identify the Pacific Ocean regions influencing the Indian monsoons, and it also adds some interpretability to the findings. The monsoons correspond to the June-September total rainfall, i.e., a single value for a year. Similarly, we have selected the SST of the Pacific Ocean for March of each year to evaluate the causal links affecting the monsoons (June-September) at a lead of three months. We normalized the SST and rainfall data to the same scale for discovering the causal dependencies between them.

We used Tigramite (a causal analysis tool in Python) for the identification of the causal links between the regions of the Pacific Ocean and the Indian monsoons [18]. The tool assists in constructing the causal graph using the nodes with significant relationships. The causal discovery method is based on the linear and non-parametric conditional independence test and performed as a two-step process. Firstly, PC algorithm [4] is used for condition selection. In PC algorithm, a superset of parents $P(X_t^v)$ is estimated for each variable v using an iterative process, which helps in avoiding the overhead of conditioning on irrelevant variables. The PC method is followed by the momentary conditional independence test, which is performed to check the sustainability of the obtained graph. The algorithm yields the causal graph as an output, where each edge is associated with a time-lag, p-value (statistical significance), and edge weight. We have considered the statistically significant edges, and have ranked them in order of edge weights. Though the aggregate and regional rainfall may share some dependencies, we mostly focus on the region of the Pacific Ocean, which bears causal dependence link to the different monsoon categories.

C. Monsoon Predictions

We predicted the aggregate and regional Indian summer monsoons with the causes identified from the Pacific Ocean region. For every rainfall categories, we ranked the source nodes of causal links to the rainfall nodes in order of their edge weights. Finally, we built the predictor sets with the top three, five, eight, and ten ranked predictors (D1-D4). Predictions of rainfall (for June-September) are provided at a lead of three months in March for monsoons during June-September.

A random forest model [19] is used for the monsoon prediction. The model merges a set of weak learners, that learns using regression tree and the training data. The ensemble model can predict response for new data by summing the forecasts from its subsequent weak learner models. We used the bagging technique for training the underlying learners based on regression trees. The number of weak learner models or trees in an ensemble is selected empirically with a five-cross validation method. It assists in maintaining a balance between the speed of the algorithm and its accuracy performance. In our case, five weak learner models are the ensemble to forecast the monsoon rainfalls. The model takes the set of predictors as input, trains the weak learners using bagging algorithms based on tree learners and finally predicts the Indian summer monsoon rainfalls.

III. EXPERIMENTAL RESULTS AND ANALYSIS

The causal links between the SST of regions of the Pacific Ocean and the Indian summer monsoons are identified with the proposed approach and presented in this section. Additionally, the prediction skill of the identified causes acting as predictors is also assessed both for the aggregate and regional Indian summer monsoons. Finally, we compared the prediction accuracy by our proposed approach with the prediction by Niño 3.4 index and India meteorological department models.

A. Identified Causal Links

We evaluated the causal links to all the five monsoon categories (AI, NEI, NWI, and SPI) at a lead of three months. The lead of three months signifies that the SST of the Pacific Ocean in March affects Indian monsoons (during June-September). Table I highlights the regions (boxes) of the Pacific Ocean that have causal links to the respective monsoon categories. We ranked the areas by their weights (i.e., following their importance). We observed that boxes 12 and 13 (encompassing the Niño 3.4 index), and box 22 (region above the Niño 3.4 index) have strong causal links to the aggregate and central Indian monsoons. However, the areas above the Niño 4 index (boxes 19 and 20) are

evaluated as potent causes for the north-eastern monsoon rainfall. Finally, the region adjacent to the Niño 4 index (box 9) and areas below the Niño 3 index (boxes 6 and 7) are assessed as active regions influencing the north-western and the southern-peninsular monsoons. Thus, we noticed that the rainfalls are affected by many areas of the Pacific Ocean, which are different from the Niño 1-4 indexes.

TABLE I
CAUSAL LINKS BETWEEN REGIONS OF THE PACIFIC OCEAN AND DIFFERENT CATEGORIES OF INDIAN SUMMER MONSOONS

Rainfall categories	**Source nodes of causal links**
AI	12, 22, 13, 9, 6, 14, 11, 20, 15, 7
CI	12, 22, 13, 11, 4, 21, 9, 6, 14, 7
NEI	20, 19, 10, 5, 17
NWI	6, 7, 9, 10, 22
SPI	6, 3, 9, 13, 12, 15, 5, 22, 14, 1

B. Prediction of Indian Summer Monsoons

The Indian summer monsoons (June-September) are predicted in March (at three months lead) using a random forest model. The mean absolute errors of the predictions of all five categories of monsoons during the test-period 2008-2018 are presented in Table II.

TABLE II
MEAN ABSOLUTE ERRORS OF THE PREDICTIONS OF INDIAN SUMMER MONSOONS DURING THE TEST PERIOD 2008-2018

Regions	**D1**	**D2**	**D3**	**D4**
AI	8.1	5.9	5.2	5.9
CI	7.8	5.8	8.3	8.8
NEI	8.5	6.7	-	-
NWI	8.7	6.9	-	-
SPI	5.6	7.3	6.4	5.1

The predictor sets (D1-D4), built with the top 3, 5, 8, and 10 ranked predictors provide predictions. NEI and NWI have predictions by D1 and D2 only, as the number of significant causal links is evaluated to be five for each. We predicted the aggregate Indian monsoon in March with a mean absolute error of 5.2%. Similarly, the rainfall of the central region of India shows an error of 5.8%. We predicted the NEI and NWI rainfall with errors of 6.7% and 6.9%, respectively. Finally, we predicted the monsoon of the south-peninsular region with a 5.1% error. We showed the variation of the observed and predicted rainfall for aggregate India in Figure 3. We noticed the predicted rainfall follows the same pattern as the observed. The extreme monsoon values are also well captured- except for the year 2009 (which was a severe drought for India).

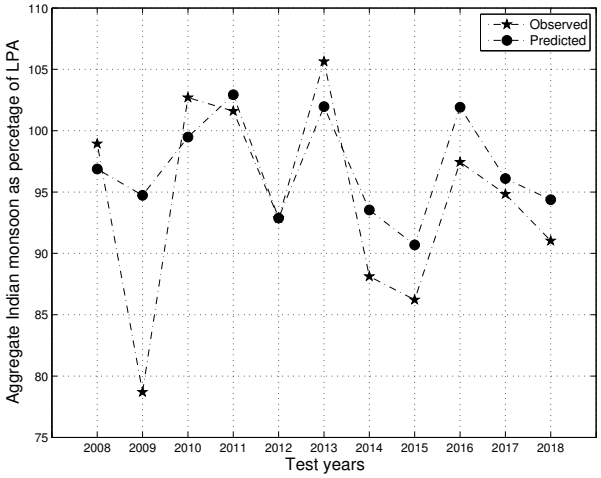


Fig. 3. Observed and predicted rainfalls for the aggregate Indian region during the test period 2008-2018

C. Comparison with the Niño Index and India Meteorological Department Model

We compared the monsoon predictions by the proposed approach of causal discovery with the forecasts by the Niño 3.4 index. Both of the predictions are provided at a lead of three months. In addition to comparing the predictions by the Niño 3.4 as a single index, we have also divided the region of the Niño 3.4 index into 3, 5, 8, and 10 different parts (same as our predictor sets) to evaluate an apples-to-apples comparison. As then, the regression model of the random forest will have the same number of predictors for both the cases and thus result in a more fair comparison. The monsoon predictions for aggregate and regional India by the Niño 3.4 index and the proposed method are presented in Table III. The last row of the table provides the prediction by the Niño 3.4 as a single index. We observed that the predictions by the discovered predictors are better than the Niño 3.4 index for all the categories of Indian monsoons.

TABLE III
MEAN ABSOLUTE ERRORS OF THE PREDICTIONS OF INDIAN SUMMER MONSOONS BY THE PROPOSED MODEL (MO.) AND THE NIÑO 3.4 INDEX (NI.) DURING THE TEST PERIOD 2008-2018

Num. of Pred.	**AI**		**CI**		**NEI**		**NWI**		**SPI**	
	Mo.	Ni.	Mo.	Ni.	Mo.	Ni.	Mo.	Ni.	Mo.	Ni.
P-3	8.1	8.9	7.8	13.0	8.5	8.9	8.7	18.5	5.6	13.7
P-5	5.9	8.5	5.8	11.9	6.7	8.1	6.9	14.6	7.3	12.1
P-8	5.2	9.1	8.3	10.4	-	-	-	-	6.4	10.6
P-10	5.9	7.7	8.8	9.9	-	-	-	-	5.1	12.0
Niño	10.6		11.7		8.9		19.2		13.4	

We also compared the monsoon predictions by the

proposed approach to the forecasts by the India Meteorological Department (IMD) [20]. IMD provides the rainfall forecast for aggregate India (AI) twice in a year: at a lead of two months (in April) and at no lead (in June). However, IMD provides the predictions for the regional monsoons (CI, NEI, NWI, and SPI) only once in June (at no lead). We compare the predictions by our proposed method at three months lead with the IMD predictions for the test-period, 2008-2018. We showed the mean absolute errors for the aggregate and regional monsoon predictions by our method and IMD models in Figure 4. We observed that the prediction by

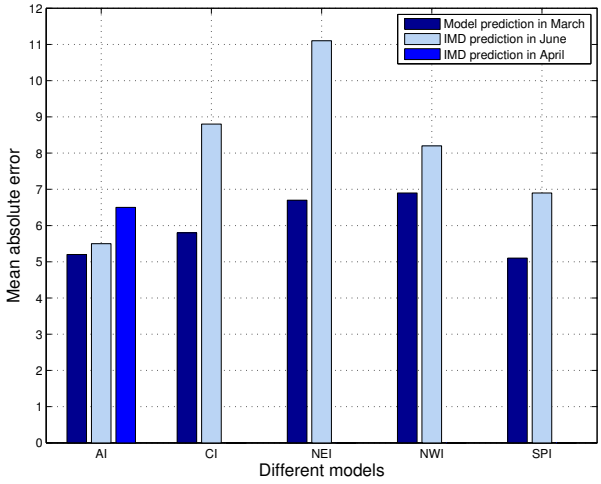


Fig. 4. Predictions of aggregate and regional Indian monsoons by the causal discovery approach and IMD models during the test period 2008-2018

the proposed approach is comparable to the predictions provided by the IMD models for aggregate Indian monsoon. However, the proposed method shows superior performance for prediction of the regional Indian monsoons. Thus, this causal discovery-based approach can identify the predictors over the Pacific Ocean more precisely, which can engross the high variability of regional monsoon rainfall and predict them with higher accuracy.

IV. CONCLUSION

Causal link detection is used to identify the significant edges between regions of the Pacific Ocean and different Indian summer monsoon categories. The PC algorithm is used to discover the causal links with statistical significance. The source nodes of the links are considered as predictors to predict both the aggregate and regional monsoons at three months lead. The monsoon predictions by the identified causes are observed to be superior to the forecasts by the Niño 3.4 index, and they are also comparable to the forecasts by the India Meteorological Department.

In the future, these causal discoveries can be further explored and analyzed for the physical interpretation of the identified causes over the Pacific Ocean influencing the Indian monsoon phenomenon. A detailed study of their influence and importance to the Indian summer monsoons can also be focused.

REFERENCES

[1] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, 1969.

[2] G. Rebane and J. Pearl, "The recovery of causal poly-trees from statistical data," *Int. J. Approx. Reasoning*, vol. 2, p. 341, 1987.

[3] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1988.

[4] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social Science Computer Review*, vol. 9, no. 1, pp. 62–72, 1991.

[5] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[6] T. Strenberg, "Investigating the presumed causal links between drought and dzud in Mongolia," *Natural Hazards*, vol. 92, no. 1, p. 27, 2018.

[7] T. Le and D. Bae, "Causal links on interannual timescale between ENSO and the IOD in CMIP5 future simulations," *Geophysical Research Letters*, vol. 46, pp. 2820–2828, 2019.

[8] M. B. Jebli and W. Hadhri, "The dynamic causal links between CO2 emissions from transport, real GDP, energy use and international tourism," *International Journal of Sustainable Development & World Ecology*, vol. 25, pp. 568–577, 2018.

[9] M. Rajeevan, "Prediction of Indian summer monsoon: Status, problems and prospects," *Current Science*, vol. 81, no. 11, pp. 1451–1458, 2001.

[10] S. Gadgil, M. Rajeevan, and R. Nanjundiah, "Monsoon prediction-Why yet another failure?," *Current Science*, vol. 88, no. 9, pp. 1389–1400, 2005.

[11] F. Kucharski, A. Bracco, J. Yahoo, and F. Molten, "Atlantic forced component of the Indian monsoon interannual variability," *Geophysical Research Letters*, vol. 35, p. L04706, 2008.

[12] A. Chakraborty and T. N. Krishnamurti, "A coupled model study on ENSO, MJO and Indian summer monsoon rainfall relationships," *Meteorology and Atmospheric Physics*, vol. 84, no. 3-4, pp. 243–254, 2003.

[13] A. Cherchi, S. Gualdi, S. Behera, J. J. Luo, S. Masson, T. Yamagata, and A. Navarra, "The influence of tropical Indian ocean SST on the Indian summer monsoon.," *Journal of climate*, vol. 20, no. 13, pp. 3083–3105, 2007.

[14] M. Saha, P. Mitra, and A. Chakraborty, "Fuzzy clustering-based ensemble approach to predicting Indian monsoon," *Advances in Meteorology*, vol. 2015, no. 32835, pp. 1–12, 2015.

[15] M. Saha, P. Mitra, and R. S. Nanjundiah, "Autoencoder-based identification of predictors of Indian monsoon," *Meteorology and Atmospheric Physics*, vol. 128, no. 5, pp. 613–628, 2016.

[16] R. W. Reynolds, N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, "An improved in situ and satellite SST analysis for climate," *Journal of Climate*, vol. 15, pp. 1609–1625, 2002.

[17] B. B. Parthasarathy, A. Munot, and D. Kothawale, "Monthly and seasonal rainfall series for All-India homogeneous regions and meteorological subdivisions: 1871-1994," *Indian Institute of Tropical Meteorology, Research Report RR-065*, 1995.

[18] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting causal associations in large nonlinear time series datasets," *arXiv:1702.07007v2*, 2018.

[19] W. Y. Loh, "Classification and regression tree methods," *Encyclopedia of statistics in quality and reliability*, pp. 315–323, 2008.

[20] M. Rajeevan, D. S. Pai, R. A. Kumar, and B. Lal, "New statistical models for long-range forecasting of southwest monsoon rainfall over India," *Climate Dynamics*, vol. 28, no. 7-8, pp. 813–828, 2007.

CHANGES IN INFORMATION HUBS OVER THE PACIFIC ENSO REGION

Moumita Saha, Dhanendra Soni, Brandon Finley, Claire Monteleoni

***Abstract*—The Pacific Ocean is an important region involving the El-Niño Southern Oscillation (ENSO) event. ENSO is known to influence several climatic phenomena around the world. In this paper, we focus on identifying the movement of information hubs in the Pacific Ocean region. We aim to characterize the ocean and assess its variations over two different periods. Causal discovery is used to identify relationships between different areas in the ocean at a lead. We obtain the information hubs from the identified causal graph. We observe the significant hubs in areas of the Niño 3.4 index during the past, which are found to get weaker during the present period. In the recent period, the regions surrounding the Niño 4 index emerge as stronger hubs in the Pacific Ocean. The information hubs are observed to be moving from the eastern to the western regions of the Pacific Ocean with decreasing strengths.**

I. MOTIVATION

The El-Niño Southern Oscillation (ENSO), occurring in the central and eastern tropical Pacific Ocean, is a recurring climatic pattern involving changes in the sea surface temperature. The phenomenon of the warming of the oceanic surface above the mean temperature is called El-Niño, and the counter cooling phase is called La-Niña. ENSO is known to be a driving force for many climatic phenomena around the world: rainfalls in East Asia [1], hurricane intensity [2], and the onset of Indian monsoons [3]. Thus, the study of the Pacific Ocean region (encompassing ENSO) is significant for characterizing and assessing the changes over time. The study also assists in understanding the effects of climate change on the global environment, such as oceans.

We focus on causal discovery for identifying the changes in information hubs over the Pacific Ocean region. Causal discovery attempts to recover the cause-effect relationships from data with the use of graphical models. The term, causality, refers to a phenomenon, where a state is responsible for the occurrence of another state. In such a case, the former state is called the 'cause' of the latter state (known as the 'effect'). A single cause can contribute to several effects, and likewise, an effect can also be the consequences of numerous causes. Causality assists in learning the cause-effect relationships and identifying the information hubs in the Pacific ENSO region.

Corresponding author: M. Saha, moumita.saha@colorado.edu, Department of Computer Science, University of Colorado, Boulder

Causality methods are widely used to explore different problems in climate domain. MacGraw and Barnis [4] highlighted the advantage of causality over lagged linear regression for climate variability studies. Lozano et al. [5] proposed causal modeling for the attribution of climate change phenomenon. Mosedale et al. [6] used Granger causality to diagnose the effects of ocean surface temperature on the values of the North Atlantic Oscillation. Granger causality was also used to assess the relationship between the global land surface temperature and carbon dioxide in the atmosphere [7]. Jajcay et al. [8] showed synchronization and causal relations across different time scales of ENSO. Causal links of interannual variability between ENSO and the Indian Ocean dipole were also investigated using a future simulation of the CMIP5 model [9]. Ebert-Uphoff and Deng [10] introduced a probabilistic-graphical climate network for comparing the features of boreal winter and summer conditions. The characteristics of atmospheric information that flew around the globe were studied to identify pathways of information flow using a climate network based on causal discovery and a graphical model [11]. Ebert-Uphoff and Deng [12] used constraint-based structure learning for exploring the causal relationships among four prominent atmospheric oscillations (the Western Pacific Oscillation, Eastern Pacific Oscillation, PacificNorth America pattern, and North Atlantic Oscillation). Song et al. [13] established the causal links between ENSO and different abnormal climate events in remote regions using a vector autoregressive model estimation method. Additionally, Nowack and Runge [14] analyzed the robustness of several causal network discovery methods using long stationary time-series data from different control runs of CMIP5 model. Thus, causal discovery methods are

seen to be used explicitly and elaborately due to their significance and efficiency.

Causal discovery is an unsupervised method that considers the temporal dimension (time) into account. Learning temporal causal graphs reveals the dependency relationships between present time and past, and thus, helps in better understanding of the complex phenomenon. The lacking skills of clustering methods in including different lags at a point make them inadequate to conclude about relationships and changes over time. We are inclined to use causal techniques and included three lags, which allowed us to uncover the movements of information hubs across two periods of 34 years. The causal discovery method assists in leaning the information centers or hubs by discovering the pathways of information flow over the Pacific ENSO region.

II. METHOD

The proposed approach to identifying the changes in information hubs over the Pacific Ocean (including the ENSO region) is shown in Figure 1.

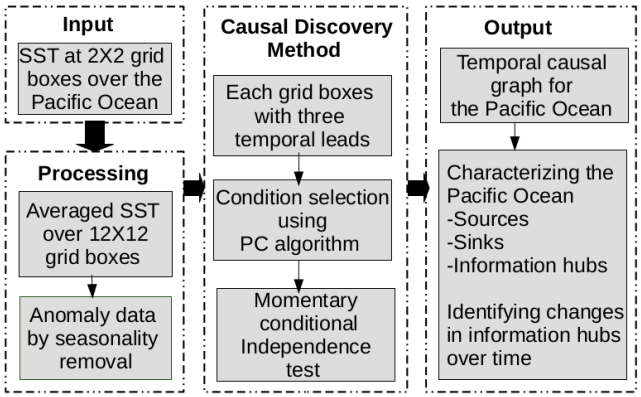


Fig. 1: Causal discovery approach to identifying the changes in information hubs over the Pacific Ocean

A. Data Processing

We used the sea surface temperature (SST) of the Pacific Ocean in this study. The investigation area includes a broader district of the Pacific Ocean with all the regions of the Niño indexes: the Niño 3.4, the Niño 3, the Niño 4, and the Niño 1+2 (we call them the Niño 1-4 indexes, henceforth). The data was collected from NOAA ERSST [15] at 2° × 2° resolution for 1951-2018. The geographical region expands over 18°N to 30°S and 160°E to 80°W. We removed the seasonality of the SST time-series for every geographical grid using the relation shown in Equation 1.

$$\text{Anomaly\_SST}_{\text{m}}^{\text{y}} = \text{SST}_{\text{m}}^{\text{y}} - \text{mean}(\text{SST}_{\text{m}}), \qquad (1)$$

where, $\text{SST}_{\text{m}}^{\text{y}}$ is the SST for the m^{th} month of the y^{th} year, and $\text{mean}(\text{SST}_{\text{m}})$ denotes the average of the m^{th} month of SST over thirty years (1951-1980).
Finally, we divided the studied period into two periods: (i) the past time-period: TP1 (1951-1984), and (ii) the present time-period: TP2 (1985-2018). These two periods are investigated independently to track the information hubs and their changes over time.

We partitioned the studied area into square grids of dimension 12° ×12°. The variation of individual SST series within a particular grid is low, so we have considered the average of all SST series present inside that grid as its representative. We obtain forty grid boxes over the studied area. These grid boxes are considered to be the nodes for building the causal graph structure. The graph will be populated by learning the relationships between these grid boxes with three different leads. The grid divisions are shown in Figure 2. The figure also highlights the regions of the Niño 1-4 indexes with different color boxes.

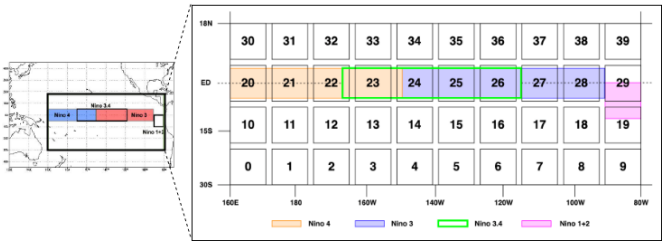


Fig. 2: Grid boxes over the studied region of the Pacific Ocean (with marked regions for the Niño 1-4 indexes)

B. Causal Discovery for Identifying the Changes in Information Hubs Over the Pacific Ocean

The primary goal of the study is to discover the information hubs over the Pacific Ocean. Learning the interactions between different regions of the ocean helps in identifying the information hubs. The temporal dimension is added to enhance the understanding of these relationships at lead months. We considered leads of one, two, and three months to evaluate the causal graph structure of the Pacific ENSO region. We restrict with a maximum lead of three months as we are concentrating on a localized area of the Pacific Ocean, where a grid region of the ocean is most likely to influence another at lesser lead only.

The forty grid-boxes are the input nodes to the causal discovery method. We have used Tigramite (a causal discovery library in Python) [16] for learning the cause-effect relationships with lead months. We have used the default setting of Tigramite for our study. The causal dependence graph is built with the addition of appropriate links between the nodes; in our case, it is the SST time-series of the nodes. The relationships can exist at various lead times (we considered a lead of one to three months). For example, if a region R1 has a causal link to another region R2 with lead three, it signifies that the SST variation in R1 at three months in the past is likely to be the cause of SST observed in R2 at present.

Causal discovery consists of two major steps: (i) condition selection, and (ii) conditional independence test. The PC algorithm developed by Spirtes and Glymour [17] is used for condition selection, which helps in overcoming the overhead of conditioning on irrelevant variables. The algorithm tries to learn a directed acyclic graph from the time-series nodes, which allows one to assess the impact of change in one variable onto other variables. The condition selection step estimates a super-set of parent regions $\mathrm{Par}(X_t^R)$ for each region R using the iterative PC algorithm. The second step is a momentary conditional independence test, which verifies the sustainability of the generated causal links (Equation 2). We calculated the partial correlation for the independence test (as the default setting of the Tigramite), which ascertain the linear relationships between the causally linked nodes of the ENSO Pacific Ocean region.

$$X_{t-\tau}^S \perp X_t^R \quad | \quad \mathrm{Par}(X_t^R), \quad \mathrm{Par}(X_{t-\tau}^S), \tag{2}$$

where, S and R are two regions, Par() represents the parent set, and τ is the lead month.

We applied the causal discovery method for both periods, TP1 and TP2 (each with 34 years). The casual graph structure for TP1 and TP2 assists in assessing the changes in information hubs over the Pacific Ocean. The method outputs a causal graph, where each causal link is associated with a time lead, p-value (statistical significance), and edge weight. We only considered the statistically significant edges with a p-value of 0.01 or less (confidence: 99%) for further evaluation of the hubs and their strengths. The same causal discovery method is also applied to an example with known relationships to validate the outcomes of the proposed approach.

III. Experimental Results and Analysis

We characterized the Pacific Ocean in terms of information hubs for the two time-periods (TP1: 1951-1984 and TP2: 1985-2018). Additionally, we also tracked the changes in the information hubs over the ocean.

A. *Information Hubs in the Pacific Ocean*

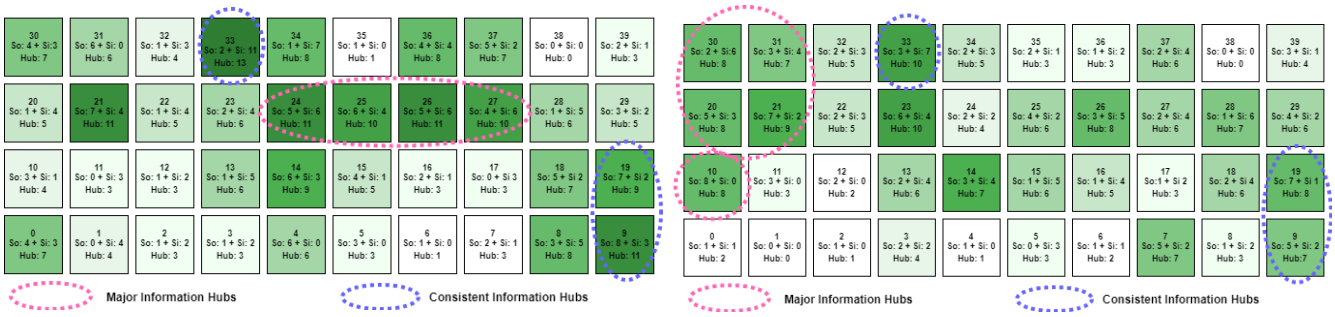
Causal discovery leads to the identification of the causal links between different regions of the Pacific Ocean at lead months. We have considered only statistically significant edges with a p-value of 0.01 or less to identify the hub regions over the ocean. The information hubs for the two time periods, TP1 and TP2, are shown in Figures 3a and 3b, respectively.

We call a region a 'source' if it has out-going edges, and similarly, we call a region a 'sink' if it has in-coming edges. In layman's terms, the source can be seen as the 'cause' while the sink as the 'effect.' We determined the strength of a source or sink by the count of out-going or in-coming edges, respectively. Finally, a region is known as a 'hub' if it behaves as both source and sink. In essence, a region is an information hub, when there is a continuous flow of information; the information flow may be out-going (when it acts as source), or it may be in-coming (when it is a sink). All the regions in the figures (Figures 3a and 3b) are marked with the source, sink, and hub strengths. The color gradient reflects the strength of the hubs. A darker shade represents a strong hub, while a lighter shade represents a weaker one.

For the first period, the primary sources are observed to be similar to the regions of the Niño 3.4 and 1+2 indexes. However, in the second period, the areas of the Niño 3.4 index seem to no longer be strong sources. The regions corresponding to the Niño 4 index and its surrounding are observed to emerge as strong sources at present. The Niño 1+2 index continues to be a strong source from the prior period.

We observed the sinks along the regions of the Niño 1-4 indexes in TP1; however, the strongest sink corresponds to region 33 with a strength of 11. In the present, most of the areas of the Niño 1-4 indexes lost their strengths as sinks. Even the power of the strongest sink of TP1 also dropped from 11 to 7 in TP2. The regions below the Niño 3.4 index are observed to evolve as sinks in the present.

Finally, the information hubs in TP1 mainly include the regions of the Niño 3.4 and 1+2 indexes. The strongest hub of this period is region 33. However, all the information hubs are observed to lose their strengths



(a) (b)

Fig. 3: Information hubs over the Pacific Ocean during (a) TP1: 1951-1984 and (b) TP2: 1985-2018

in TP2. New hubs are seen to be evolving around Niño 4 index in the current period.

B. *Changes in Information Hubs over the Pacific Ocean*

It is interesting to assess the changes in information hubs over the Pacific ENSO region for the two periods. It helps in analyzing the overall shifting of hubs from the past to the present period.

We observed the important hubs in the regions of the Niño 3.4 index during 1951-1984 (highlighted by a pink oval area in Figure 3a). An area of the Niño 4 index (region 21) and the Niño 1+2 index are also active hubs. These results certify the efficiency of the proposed technique for identifying information hubs, as the Niño 1-4 indexes are well-known for their significance [18]. We also identified the neighboring areas above the Niño 3.4 index as hubs (their strengths may be due to their proximity with other strong hubs of the period).

The first observation during 1985-2018 is an apparent trend in the decrease in the strength of the hubs. However, it is exciting to find that the region of the Niño 4 index and its surrounding are emerging as stronger information hubs in the current period (highlighted by a pink area in Figure 3b). The Niño 1+2 index with its below region and the region 33 are consistent information hubs in both TP1 and TP2 (marked by the blue boundary in Figures 3a and 3b).

It is important to note that the information hubs are losing their strengths in the east and gaining them in the west. In other words, the information hubs are observed to be shifting from the eastern to the western Pacific Ocean at the present period.

C. *Validation of the Causal Discovery Method*

We also applied the proposed causal discovery approach to an example data set, where the causes and effects are known. The above method is a validation step to emphasize the fact about the meaningfulness of the causal links discovered with the algorithm. The study consists of three different variables: daily mean temperature (TMP), cloud cover (CLD), and wet day frequency (WET). The variables are considered on a daily scale from November 2017 to December 2018. We collected the variables from the CRU dataset [19]. The reason for selecting the daily data for 14 months (which is equivalent to time-series length of 426) is to keep the length of time-series for the validation set as similar to our original experiment (which is monthly data for 34 years: sums to time-series length of 408). Liu et al. [20] used the same example to understand the cause-effect relationships between temperature and several more variables (the possible causes for temperature change). We considered the data of the Green state (latitude: 45.25° and longitude: -107.25°) of the Liu et al. [20] study for this experimentation. Their result suggests that both the parameters CLD and WET are affecting the temperature (TMP) variable. We will consider their result as ground truth to validate the graph obtained from our causal discovery method. We showed the discovered causal graph structure with our algorithm in Figure 4.

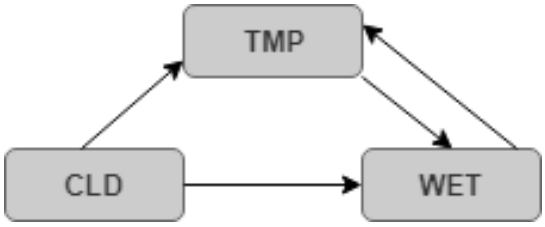


Fig. 4: Causal discovery for the example problem relating temperature (TMP), cloud cover (CLD), and frequency of wet days (WET) of the Green State [20]

It is observed that incoming edges to the TMP variable are present from both the CLD and WET variables,

supporting the results of Liu et al. [20]. In their study, they tried to identify the effect of other variables on temperature only, so they have only incident edges to the variable temperature. However, in our study, we aim to find the relationships between all the variables, and thus, we observe two other links: (i) CLD to WET: signifies that cloudiness has an effect on the number of wet days, and (ii) TMP to WET: signifies that temperature change in the atmosphere can lead to wet days (rainfalls). Therefore, this example helps us to validate the results of information hubs obtained for the Pacific Ocean region and their changes over time.

IV. CONCLUSION

The changes in the information hubs over the Pacific Ocean are evident. We observed the regions of the Niño 1-4 indexes and surrounding to vary in their hub strengths from the past to present period. The information hubs are noticed to be moving from the eastern to the western Pacific Ocean. The regions corresponding to the Niño 3.4 index were evaluated as significant hubs in the past time; the same areas are observed to lose their strengths in the present. The areas of the Niño 4 index and its surrounding are emerging as significant information hubs in the present period. These results leave us with an important question that arises from the observations: 'Why do we confine ourselves to the rectangular boxes of the Niño 1-4 indexes to consider them as the predictors?'- We believe that there are changes in these Niño 1-4 indexes as well.

In the future, we need to explore the physical interpretations of these identified hubs. We may also try to investigate the oceanic or atmospheric process that has influenced the change of these information hubs over time. We believe that the identified hubs may serve as good predictors for forecasting diverse climatic phenomena. It will be also interesting to assess the performance of these hubs in predicting different climatic events as compared to the Niño 1-4 indexes.

REFERENCES

[1] L. Chongyin, "Interaction between anomalous winter monsoon in East Asia and El Nino events," *Nature*, vol. 7, no. 1, pp. 36–46, 1990.

[2] J. Donnelly and J. Woodruff, "Intense hurricane activity over the past 5,000 years controlled by El Niño and the West African monsoon," *Advances in Atmospheric Sciences*, vol. 447, pp. 465–468, 2007.

[3] P. V. Joseph, J. K. Eischeid, and R. J. Pyle, "Interannual variability of the onset of the Indian summer monsoon and its association with atmospheric features, El Nio, and sea surface temperature anomalies," *Journal of Climate*, vol. 7, no. 1, pp. 81–105, 1994.

[4] M. C. McGraw and E. A. Barnes, "Memory matters: A case for Granger causality in climate variability studies," *Journal of Climate*, vol. 31, no. 8, pp. 3289–3300, 2018.

[5] A. Lozano, H. Li, A. Niculescu-Mizil, Y. Liu, P. C., H. J., and N. Abe, "Spatial-temporal causal modeling for climate change attribution," in *International conference on Knowledge discovery and data mining*, SIGKDD'09, pp. 587–596, 2009.

[6] T. J. Mosedale, D. B. Stephenson, M. Collins, and T. C. Mills, "Granger causality of coupled climate processes: Ocean feedback on the North Atlantic Oscillation," *Journal of Climate*, vol. 19, no. 7, pp. 1182–1194, 2006.

[7] E. Kodra, S. Chatterjee, and A. R. Ganguly, "Exploring Granger causality between global average observed time series of carbon dioxide and temperature," *Theoretical and Applied Climatology*, vol. 104, no. 3-4, pp. 325–335, 2011.

[8] N. Jajcay, S. Kravtsov, G. Sugihara, A. A. Tsonis, and M. Palus, "Synchronization and causality across time scales in El Niño Southern Oscillation," *Climate and Atmospheric Science*, vol. 1, no. 1, pp. 2397–3722, 2018.

[9] T. Le and D. Bae, "Causal links on interannual timescale between ENSO and the IOD in CMIP5 future simulations," *Geophysical Research Letters*, vol. 46, pp. 2820–2828, 2019.

[10] I. Ebert-Uphoff and Y. Deng, "A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer," *Geophysical Research Letters*, vol. 39, p. L19701, 2012.

[11] Y. Deng and I. Ebert-Uphoff, "Weakening of atmospheric information flow in a warming climate in the community climate system model," *Geophysical Research Letters*, vol. 41, no. 1, pp. 193–200, 2014.

[12] I. Ebert-Uphoff and Y. Deng, "Causal discovery for climate research using graphical models," *Journal of Climate*, vol. 25, no. 17, pp. 5648–5665, 2012.

[13] H. Song, J. Wang, J. Tian, J. Huang, and Z. Zhang, "Spatio-temporal climate data causality analytics - an analysis of ENSOs global impacts," in *International Workshop on Climate Informatics*, CI'2018, pp. 45–48, 2018.

[14] P. J. Nowack and J. Runge, "Large-scale causal network discovery in CMIP5 models: robustness and intercomparison," *American Geophysical Union, Fall Meeting 2018*, 2018.

[15] R. W. Reynolds, N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, "An improved in situ and satellite SST analysis for climate," *Journal of Climate*, vol. 15, pp. 1609–1625, 2002.

[16] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting causal associations in large nonlinear time series datasets," *arXiv:1702.07007v2*, 2018.

[17] P. Spirtes and C. Glymour, "An algorithm for fast recovery of sparse causal graphs," *Social Science Computer Review*, vol. 9, no. 1, pp. 62–72, 1991.

[18] K. Yamasaki, A. Gozolchiani, and S. Havlin, "Climate networks around the globe are significantly affected by El Niño," *Physical Review Letters*, vol. 100, p. 228501, 2008.

[19] I. Harris, P. Jones, T. Osborn, and D. Lister, "Updated high-resolution grids of monthly climatic observations the cru ts3.10 dataset," *International Journal of Climatology*, vol. 34, no. 3, pp. 623–642, 2014.

[20] Y. Liu, A. Niculescu-Mizil, A. Lozano, and Y. Lu, "Learning temporal causal graphs for relational time-series analysis," in *International Conference on International Conference on Machine Learning*, ICML'10, pp. 687–694, 2010.

CAN AVALANCHE DEPOSITS BE EFFECTIVELY DETECTED BY DEEP LEARNING ON SENTINEL-1 SATELLITE SAR IMAGES?

Saumya Sinha*[1], Sophie Giffard-Roisin*[1], Fatima Karbou[2], Michael Deschatres[3], Anna Karas[2], Nicolas Eckert[3], Cécile Coléou[2], Claire Monteleoni[1]

***Abstract*—Achieving reliable observations of avalanche debris is crucial for many applications including avalanche forecasting. The ability to continuously monitor the avalanche activity, in space and time, would provide indicators on the potential instability of the snowpack and would allow a better characterization of avalanche risk periods and zones. In this work, we use Sentinel-1 SAR (synthetic aperture radar) data and an independent in-situ avalanche inventory (ground truth) to automatically detect avalanche debris in the French Alps during the remarkable winter season 2017-18. Convolutional neural networks are applied on SAR image patches to locate avalanche debris signatures. We are able to successfully locate new avalanche deposits with as much as 77% confidence on the most susceptible mountain zone (compared to 53% with a baseline method). One of the challenges of this study is to make an efficient use of remote sensing measurements on a complex terrain. It explores the following questions: to what extent can deep learning methods improve the detection of avalanche deposits and help us to derive relevant avalanche activity statistics at different scales (in time and space) that could be useful for a large number of users (researchers, forecasters, government operators)?**

I. INTRODUCTION

Remote sensing of avalanche debris in mountain areas offers new opportunities to improve our understanding of avalanche activity and to evaluate the physical models of avalanche hazard forecasts. The location of avalanche debris and the estimation of their sizes are of great interest for studies dealing with the stability of the snowpack and also the variability of natural avalanche activity, which could be related to climate change. In addition, time series of avalanche events, with relevant time and space resolutions, would be highly relevant to better identify avalanche risk zones and periods. Such time series would be a great addition to the existing databases, mostly based on visual observations. Despite their great value, these in-situ data are scarce and are limited by the terrain accessibility, the weather conditions and the danger of avalanches themselves. In this study we use backscatter coefficients at C-band from the SAR onboard Sentinel-1A and -1B satellites launched between 2014 and 2016. The French Alps are observed every 6 days.

The study period covers the winter of 2017-18, which was marked by particularly high avalanche activity recorded in the French Alps. Microwave backscattering over snow surfaces is complex because it combines several phenomena including reflection on the snow surface, scattering within the snowpack (which depends on its layers properties) and reflection at the snow-soil boundary. To detect avalanche debris, change detection methods are typically used to isolate avalanche debris-like features based on the backscatter contrast between avalanche debris and the surrounding undisturbed snowpack [1]. Debris detection is based on major changes in the backscatter coefficients due to changes in snow properties following the avalanche event (height, density, roughness), Figure 1 shows an example of an RGB composition map using 3 Sentinel-1 images at VH polarization. The large avalanche event near "Les Houches" can be seen in green.

Recent work [1], [2] demonstrated the potential of Sentinel-1 SAR data for avalanche mapping on specific examples. Karbou et al. [3] applied a change detection method and combined Sentinel-1 ascending/descending orbits to automatically detect avalanches at the scale of a mountain chain. However, the complexity of the interaction of the radar signal and the snow medium necessitate the development of more advanced algorithms that are also able to better manage the consistent data

Corresponding author: S. Giffard-Roisin, sophie.giffard-roisin@mines-saint-etienne.org [1]University of Colorado Boulder, USA [2]CNRM-GAME, Météo-France, and CNRS, Centre d'Etudes de la Neige, Grenoble, France [3]Irstea, Université Grenoble Alpes, France. * authors contributed equally.

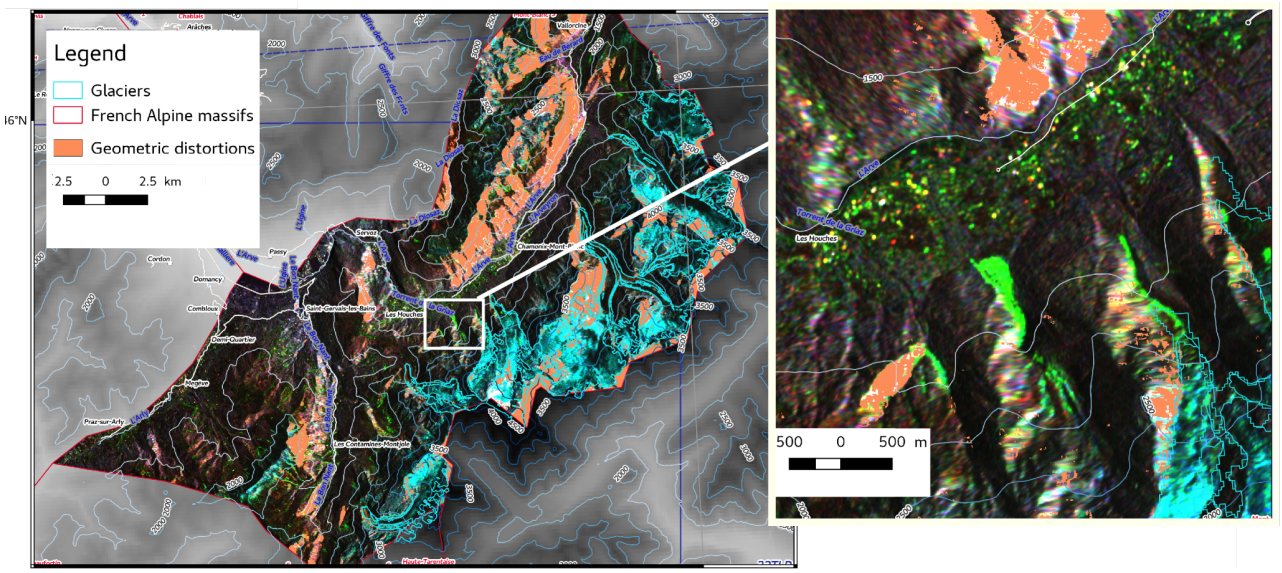


Fig. 1: An RGB composition SAR image over the Mont Blanc chain (one of the 23 French alpine massifs of the database) using 3 sentinel-1 VH images (R: 2017/08/24, G: 2018/01/15, B: 2018/01/09) highlighting avalanche debris signatures in light green for events between the 09th and the 15th of January 2018, such as the avalanche event occurred near les Houches (see zoom).

flow.

With the advances in machine learning, recent works proposed classification methods for this task, using a random forest classifier [4] or convolutional neural networks [2]. The results are very promising; however they both rely on expert labelling from the same SAR imagery. This has two major limitations: i) no study has been made on the accuracy of expert labelling from SAR signals; ii) it is not possible to differentiate between a new avalanche and an old one that is still visible.

We propose in this paper to couple the SAR data with an independent ground truth database which would answer both issues. Specifically, we used an avalanche inventory covering more than 3000 avalanche corridors in the French Alps, which are collected by forest rangers from ONF (Office National des Forêts) and stored by Irstea research institute. From the partial information available in the inventory (the specific delineation of the avalanche is not accessible), we automatically cropped some image patches containing the zone of deposition for every avalanche in the database. We then constructed and trained a convolutional neural network model able to classify the satellite image patches as *avalanche* or *no avalanche*. By using the SAR acquisitions of both current time and previous acquisition (6 days earlier), we trained our network for detecting only the new avalanches. From a database of more than 1300 samples from 16 mountain chains (out of 23 for the whole alps), we were able to detect new avalanches. We compared our results with a baseline method. Moreover, we performed an analysis to generate insights on the types of avalanches that can be identified.

II. METHOD

A. Data Processing

a) Avalanche inventory: The EPA (Enquête Permanente sur les Avalanches) database includes field observations on more than 3000 paths (mountain corridors where avalanches occur). Avalanche occurrences are recorded, along with quantitative and qualitative data (runout altitudes, release cause, damages, etc.). We used more than 4000 avalanche events annotated from the 2017-18 season and attributed to the different EPA paths. With SAR data, we can observe a relative increase of the backscatter due to snow deposit. Consquently, we automatically extracted the lowest elevation parts of the EPA corridors, where most of the avalanches would have their zone of deposition.

b) Sentinel-1 SAR imagery: We extracted the Sentinel-1 synthetic aperture radar (SAR) polarizations VV and VH on the whole region from the descending relative orbit 139, with a 20m resolution[1]. SAR

[1]We used the Level-1 Ground Range Detected (GRD) products made available through the Copernicus web site https://scihub.copernicus.eu/dhus/.

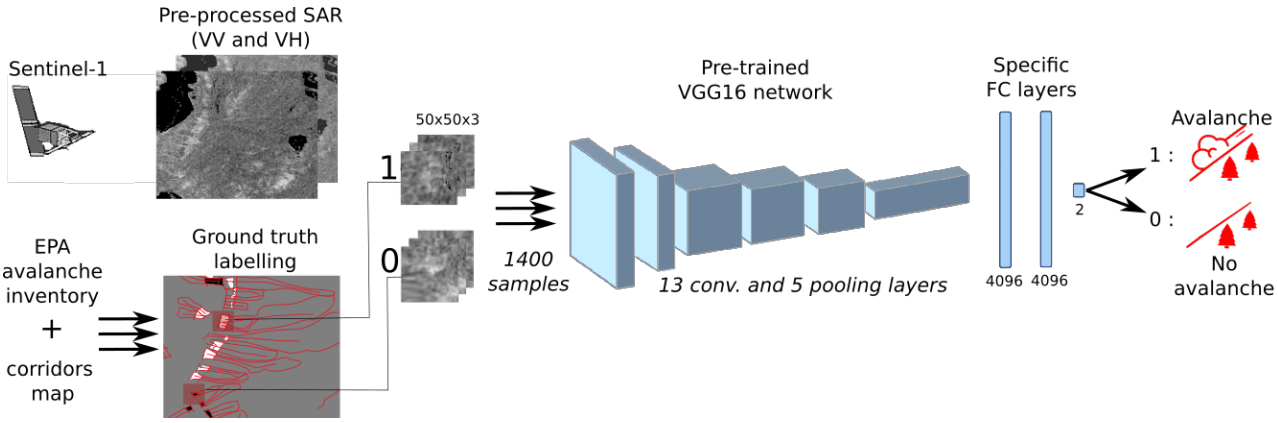


Fig. 2: Avalanche detection pipeline. From an independent ground truth labelling, 50% positive and 50% negative SAR satellite image patches are feeding a convolutional neural network composed of convolutional layers (Conv) and fully connected layers (FC). The 3 input channels are VV^*, VV^*_{old} from the previous satellite acquisition (6 days earlier), and $VH_{diff} = VH^* - VH^*_{old}$.

data have been processed using the ESA Sentinel-1 Toolbox (speckle filtering, radiometric calibration, terrain correction, etc.). With one acquisition every 6 days, we collected a total of 32 dates in the season. Sentinel-1 has a side-looking imaging geometry which causes geometric distortion occurrences in mountains including shadow, layover and foreshortening effects. These areas are screened out. We calculated the images ratio (with respect to snow-free summer images): $VV^* = 10\log 10(VV/VV_{summer})$ and $VH^* = 10\log 10(VH/VH_{summer})$ as well as the difference $VH_{diff} = VH^* - VH^*_{old}$ with the previous satellite acquisition (6 days earlier).

c) Label maps: For every SAR acquisition date, we calculated a label map where a zone is positive if an avalanche was monitored between the last acquisition and this one (6-day window); negative if not. The zones outside of the EPA corridors (thus not monitored) are considered as unknown and not labelled. Moreover, if the uncertainty on the date at which the avalanche occurred was larger than the 6 days between the acquisitions, the zone was also not labelled.

B. Learning Framework

a) Satellite image patches: Because the zone of deposition of every event is roughly localized (from the corridors of the EPA map), a segmentation task is not possible. That is why we used satellite image patches (of 50x50 pixels, so $1km^2$) centered on the lowest elevation part of the corridors, assuming that any snow deposit would be included. We stored 3 feature image channels: VV^*, VV^*_{old} (from the previous acquisition 6 days earlier) and the difference VH_{diff}. 658 positive patches (containing an avalanche) were available, and we randomly extracted the same number of negative patches from non-active corridors. The 1316 samples were randomly separated into 3 sets as follows: train (60%) / valid (20%) / test (20%), where different acquisitions (dates) of a same corridor were kept in a unique set.

b) Convolutional neural networks (CNN) model: Because of the limited number of samples, we developed a transfer learning method that uses a pre-trained CNN network which is then fine-tuned for our specific problem. Following recent studies [2], [5], we used the VGG 16 network [6] (composed of 13 convolutional layers) trained on the ImageNet database, and optimized only the 3 fully connected (FC) layers (see Figure 2). We used a cross-entropy loss as criterion for the classification task. To reduce over-fitting, each sample was subject to random augmentation: flipping of axes and 50x50 cropping from a 64x64 initial patch. Moreover, we also used a 50%-rate dropout on the FC layers. The evaluation was repeated three times and an average score was computed in order to assess the robustness of the random weights initialization. We used an early-stopping model selection on a maximum of 250 epochs.

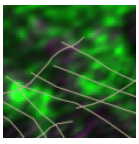
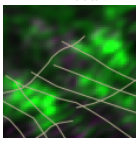
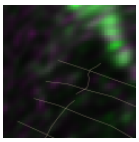
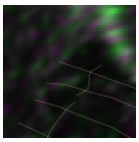
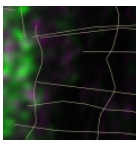
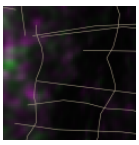
III. EVALUATION

a) Quantitative results: Once the best model using the validation set selected, we present in Table I the results of the test set (kept hidden). We compared our results with the thresholding method [3]. In order

TABLE I:
Comparison between our method and the baseline (thresholding method). Test set: 211 samples from which 1/4 are from the Haute Maurienne chain.

	Haute Maurienne		All Alps	
	CNN	**Baseline**	**CNN**	**Baseline**
Accuracy	0.77	0.53	0.69	0.58
Precision	0.81	0.51	0.69	0.57
Recall	0.74	0.72	0.69	0.59
F1-score	0.78	0.6	0.69	0.58

TABLE II:
Examples of classification results. RGB composition of the 2 VV images (as R: VV_{summer}, G: VV, B: VV_{summer}) given as input. The light green should reflect avalanche deposits.

VV	VV_{old}	**Label**	**Prediction**
		**1**	**1**
		**0**	**0**
		**1**	**0**

to automate the threshold for image classification, we calculated the number of positive threshold pixels per image that gave the best result on the validation set (above which the whole patch is considered as *positive*), and used this parameter on the test set. We can see that our method outperforms the baseline on all of the metrics (accuracy, recall, precision and F1-score). We can see that the results in the Haute Maurienne chain (77% accuracy), containing a quarter of the total number of samples, are clearly better than the result on the whole 16 mountain chains (69% accuracy). This seems to indicate that it is easier to detect avalanches in zones where we have a good amount of data, even if the corridors in the test set were unseen. Table II shows three examples of classification in an RGB composition where the green should reflect avalanche deposits.

b) Analysis: The results were then further analyzed, in order to understand what types of avalanches can be detected. We observe no significant difference in terms of season (month) and local slope between true positives (TP, avalanches correctly detected) and false negatives (FN, avalanches missed). Yet, we noticed that the small avalanches (area< $70m^2$ according to the EPA database) were more missed than others (40% of them were not detected). We also noticed a difference in the orientation of the mountain patches. The proportion of FN patches facing East is 69%, while it is only 44% for TP. This might be due to the angle at which the satellite is facing the mountain (seeing better the slopes facing West for the descending orbit). Lastly, as we have seen with Haute Maurienne, the mountain chains with the larger number of samples had the best results, probably because different conditions (orientation, preprocessing of the signal,ect.) are dependent on the mountain zone. This is currently a limitation of the method, however it should be resolved by (i) increasing the size of the database with more seasons (since 2015), (ii) increasing the confidence on the result by combining several satellite orbits (4 relevant ascending/descending orbits in our test zone), for a better coverage of the mountains.

CONCLUSION

This is the first quantitative study combining SAR imaging data with an independent in-situ avalanche inventory. The complexity of the SAR signal and the uncertainty on the labels make this problem particularly challenging. We showed that by selecting some patches centered on the lower part of inventoried avalanche corridors, a convolutional neural network can detect the presence of avalanche debris with an accuracy of up to 77% in the most susceptible mountain zone. Moreover, we identified two causes of misclassification: the size of the avalanche debris, and the orientation of mountains. These new insights can help to make an efficient use of remote sensing measurements on this complex terrain. This is an encouraging first step towards a efficient use of remote sensing for avalanche forecasters and local policies.

ACKNOWLEDGMENTS

This work was supported by the Programme National de Télédétection Spatiale (www.insu.cnrs.fr/pntsgrant N. PNTS-2019-10). Sentinel-1 data are freely available from the European Space Agency (ESA) as well as the SNAP toolbox to process raw SAR observations.

We acknowledge the support provided by ESA for the processing resources and data access provided through the RSS CloudToolbox service, the support of the Geohazard Exploitation Platefom (GEP) initiative and the support of Copernicus Research and User Support (RUS) service.

REFERENCES

[1] F. Karbou, M. Lefort, M. Dumont, N. Eckert, M. Deschrtres, and R. Martin, "Multi-temporal avalanche debris mapping in the french mountains using synthetic aperture radar observations from sentinel-1," in *EGU General Assembly Conference Abstracts*, vol. 20, p. 18024, 2018.

[2] A. U. Waldeland, J. H. Reksten, and A.-B. Salberg, "Avalanche detection in sar images using deep learning," in *IGARSS 2018-2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2386–2389, IEEE, 2018.

[3] F. Karbou, C. Coléou, M. Lefort, M. Deschatres, N. Eckert, R. Martin, G. Charvet, and A. Dufour, "Monitoring avalanche debris in the french mountains using sar observations from sentinel-1 satellites," in *International snow science workshop (ISSW)*, 2018.

[4] J. B. Hamar, A.-B. Salberg, and F. Ardelean, "Automatic detection and mapping of avalanches in sar images," in *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 689–692, IEEE, 2016.

[5] P. Egil Kummervold, E. Malnes, M. Eckerstorfer, I. Arntzen, and F. M. Bianchi, "Avalanche detection in sentinel-1 radar images using convolutional neural networks," 10 2018.

[6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

USE OF IMAGE ANALYSIS TOOLS TO EXPLORE THE SPATIAL PATTERNS OF EXTREME RAINFALLS IN ASIA: COMPARING REMOTE SENSING AND MODEL-BASED RAINFALL DATA

Carlos H. R. Lima[1] Hyun-Han Kwon [2]

***Abstract*—The spatial patterns of large-scale, extreme rainfall storms are complex to understand and hard to simulate in climate models. Here we seek to advance this field by using image analysis tools to investigate some spatial features of extreme rainfalls as provided by remote sensing and model-based data. We focus on rainfall datasets provided by satellites (GPM and PERSIANN datasets), general circulation models (HadGEM2) and reanalysis data (NCEP/NCAR). Daily rainfall data for a period of about 5 years are obtained for the Asia region, including part of the continent and part of the Pacific ocean, and three key variables are extracted and compared across the different datasets: centroids, approximated areas of the extreme rainfall fields and distances among the rainfall storm clusters. The results reveal significant discrepancies among the datasets. In particular, the centroids of extreme rainfalls tend to be located in the Pacific ocean for the satellite-based data, while the model-based rainfall data have centers dispersed over the domain, but predominantly on the continental region. The satellite rainfall data have also a tendency of having smaller storm areas (average $\sim$ 1.000 km^2) and shorter distances between storm clusters (average of minimum distance $\sim$ 200 km), when compared with model-based rainfall data (average area $\sim$ 3.000.000 km^2 and minimum distances between $\sim$ 800 km and $\sim$ 1.300 km). These results stress some deficiencies of atmospheric models to simulate the correct spatial scale and location of extreme rainfalls, pointing to relevant aspects that need a better representation in the models.**

I. MOTIVATION

The recent availability of large amounts of remote sensing rainfall data provides an opportunity to investigate new spatial patterns and features which otherwise would not be possible with ground station data only. For instance, the spatial patterns of rainfall storms over large areas, particularly over the oceans, are not well understood given the lack of ground observations. Likewise, we do not perceive the ability of atmospheric general circulation models (GCMs) and reanalysis data to reproduce the spatial features of rainfall fields as estimated by satellite data, which could be seen as a proxy of the actual rainfall field in areas where no ground observations exist.

We noticed in the literature some studies that investigated the spatial dynamics of rainfall data as provided by ground and radar sources ([1], [2], [3]), which have smaller spatial scales and higher frequencies than rainfall data as estimated by satellites and simulated by most atmospheric models. We also observed some efforts to properly simulate the rainfall field using dynamical models ([4], [5]) and, for large spatial domains, we noted that the focus has been on the analysis of storm tracks ([6], [7]), which has some advantages but misses features related to the organization of storms over the entire domain. Therefore, we believe that the spatial analysis of rainfall over large domains deserves further investigation and, in this framework, we apply here image analysis tools to readily extract key features (centroids, areas and distance across clusters) of rainfall storms as provided by different remote sensing and model-based sources.

II. METHOD

Table 1 provides the different remote sensing and model-based sources for the rainfall data used in this study. We have two datasets from satellite sources, two purely based on models (HadGEM2) and one that combines observations with model-based simulations (reanalysis data). The time step for all datasets is daily and the spatial domain is delimited by the coordinates

Corresponding author: C. H. R. Lima, chrlima@unb.br
[1]Department of Civil and Environmental Engineering, University of Brasilia, Brazil [2]Sejong University, Seoul, South Korea

90°E – 150°E longitude and 0° – 45°N latitude, which includes a large portion of Asian and the western Pacific Ocean (Fig. 1). The relatively short period results from the limited data availability of the rainfall sources, especially for the GPM data. Note also that we include in the comparison future scenarios of rainfall as simulated by the HadGEM2 model under the Representative Concentration Pathway (RCP) scenario 8.5.

TABLE I
RAINFALL DATA SOURCES

Dataset	Source	Period	Resolution
GPM [8]	Satellite	2014-2018	0.20°x0.20°
PERSIANN [9]	Satellite	2014-2018	0.25°x0.25°
NCEP/NCAR [10]	Reanalysis	2014-2018	1.87°x1.90°
HadGEM2-H [11]	GCM	2014-2018	1.87°x1.25°
HadGEM2-F [11]	GCM	2015-2025	1.87°x1.25°

The methodology to extract the spatial features for each dataset at each day of the record period, as indicated in Table 1, can be summarized as follows:

- The 99.5th rainfall quantile over the entire spatio-temporal domain (for all pixels, including the non-precipitating ones) is obtained. Each pixel value below this value is masked with zero, so that only the rainfall extremes are displayed (Fig. 2 top). The 99.5th quantile is somewhat arbitrary to define extreme rainfall events and we noticed no meaningful differences in the results as we change to values between the 95th and the 99.9th quantiles.
- For each day of the record, the Canny edge detection algorithm [12] is applied to identify the boundaries of the rainfall storms, which are marked with value 1, while the remaining pixels are assigned with zero values (Fig. 2 bottom);
- For each edge/rainfall storm identified, we attempt to fit circles that enclose the edge of the images (Fig. 3);
- For each day of the record, the following features are extracted from the circles and stored: centroids (black asterisks in Fig. 3), circle areas and distances among circles. It is worth mentioning that the centroids coincide with the center of the circles, and not necessarily with the actual center of the rainfall field. These features are then used in all subsequent analysis;
- The entire procedure is repeated for each rainfall data source.

Most of the routines used here were developed with the support of the Image Processing Toolbox in MAT-

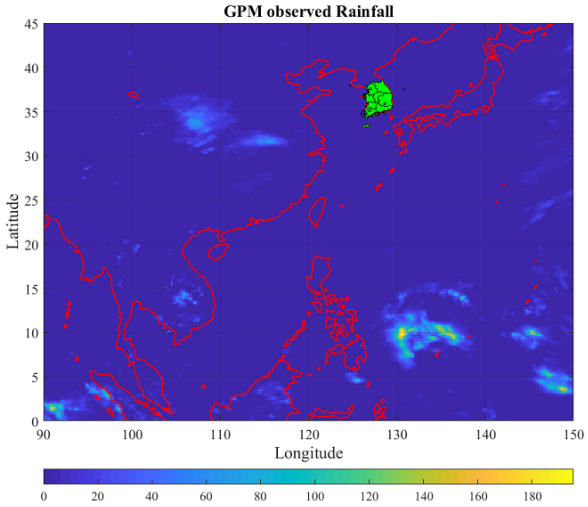


Fig. 1. GPM observed rainfall (in mm) for a given random day. The Korean Peninsula is featured in green, while the red lines show the coastline.

LAB.

III. EVALUATION

The centroids and the respective density for the GPM set are shown in Figure 4. Clearly, the centers of extreme storms are mostly located in the southwest part of the domain, which encloses the Pacific ocean. A very few points appear in the Asian continent. The Persiann rainfall dataset, which is also derived from satellites, shows also a similar pattern, but with a higher density in the southeast part of the domain (top left panel of Fig. 5). On the other hand, the model based rainfall storms have centroids mostly concentrated in the Asia continent (Fig. 5).

The density distribution of the areas of the rainfall storms, as estimated by enclosed circles (Fig. 3), is shown in Figure 6 (top panel). Again, the sizes of the satellite sources (average ~ 1.000 km^2) are similar and much smaller than those obtained for the model-based rainfall data (average between $\sim$ 500.000 km^2 and $\sim$ 3.000.000 km^2). It is worth mentioning that these numbers should be depicted in a relative context, as the circles are just roughly approximations of the actual size of the extreme rainfall storms (See Fig. 3). Due to the small sample size, they should neither be seen as the climatology of rainfall storm areas.

For each day of the record, we also estimate all the distances across the rainfall storm circles and then store the minimum distance as a surrogate measure of the clustering degree. Figure 6 (bottom panel) displays the

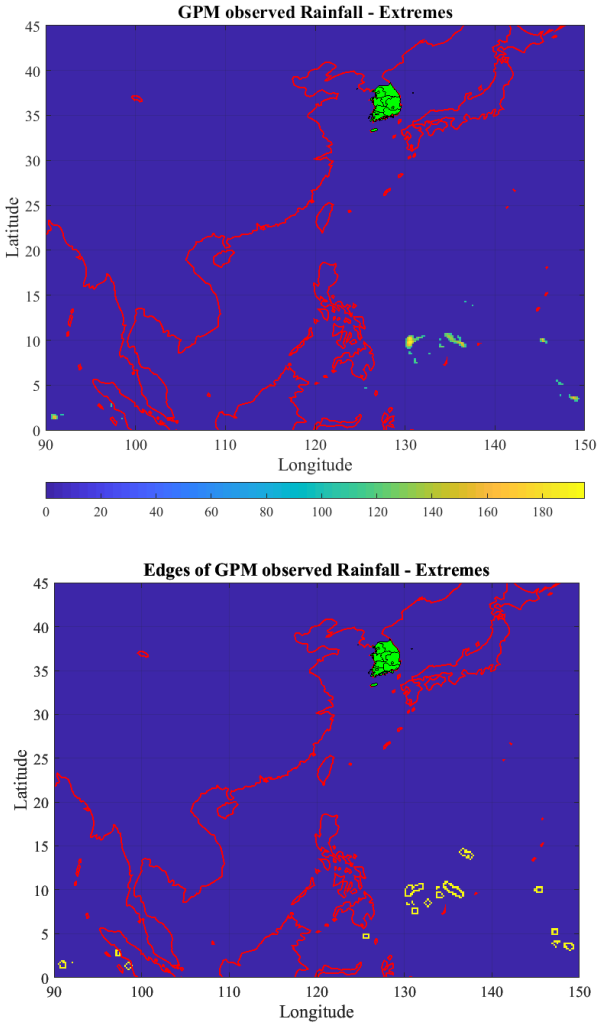


Fig. 2. As in Figure 1, but highlighting the extreme rainfalls (top panel) and the respective edges detected (bottom panel).

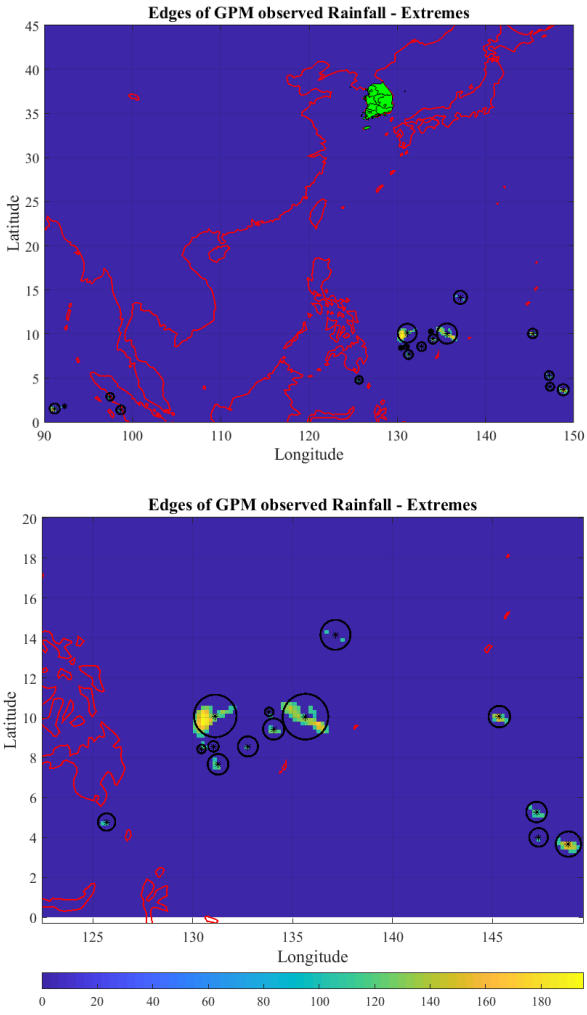


Fig. 3. Top panel: circles (black lines) and respective centroids (black asterisks) enclosing the extreme rainfalls identified in Figure 2. Bottom panel: zooming a key area in the top panel.

distribution of the minimum distances across days and for all datasets. The satellite based rainfall estimates show smaller and more concentrated distances (average $\sim$ 200 km), while the model-based estimates have larger and more spread distances (average between $\sim$ 800 km and $\sim$ 1.300 km). Part of this conclusion is coherent with the previous one, as storms enclosed by larger circles tend to be more apart from those enclosed by smaller circles.

Although the differences in the grid resolution (Table 1) across the datasets might lead to differences in the size and distances of rainfall storms (Fig. 6), we believe that atmospheric models are in fact overestimating the sizes of the extreme storms, as they are not able to capture the actual size when using large-resolution grid cells. Likewise, while the storm size identified by the satellite data resembles the scale of the grid cell ($\sim$ 700 km), the size estimated by the model-based data is much larger than the respective spatial scale of the grid ($\sim$ 28.000 km$-$43.000 km). On the other hand, in case extreme storms had the sizes of those determined by the GCM models and reanalysis data, then the satellites would be able to estimate those sizes as they have finer resolution grids. Whether atmospheric models with the same physics but with finer-resolution grids cells would correctly simulate such storm sizes is still a question.

The results obtained in this work highlight the deficiencies of atmospheric models to simulate the correct spatial scale of extreme rainfalls, assuming that the satellite estimates are a trusty proxy of the actual rainfall field. The scale reproduced by the atmospheric models is 10x-100x larger than their own grid resolution

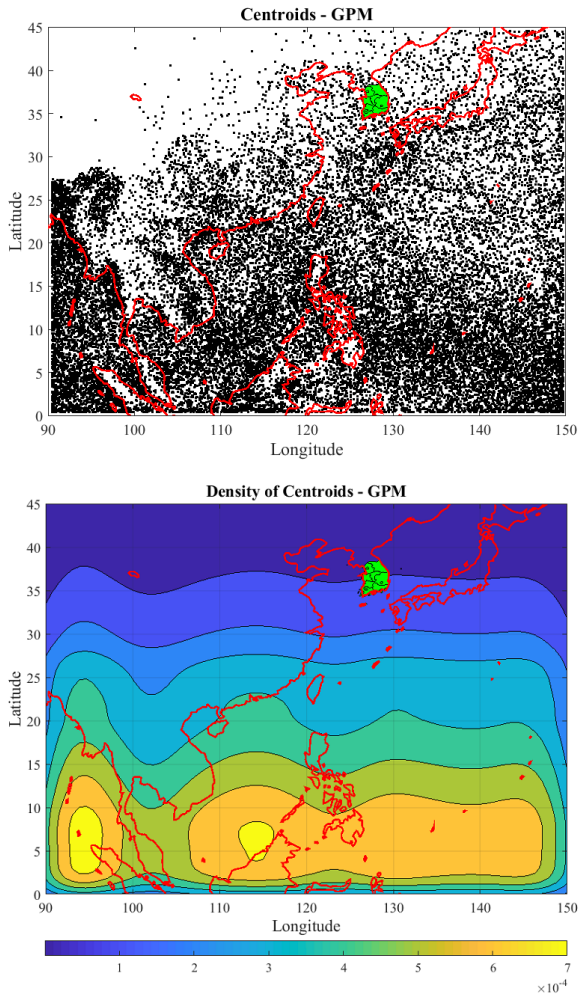


Fig. 4. Location (top panel) and density (bottom panel) of centroids for the GPM dataset. The red lines show the coastline, while the Peninsula Korea is featured in green.

and 500x-3.000x than the satellite estimates. Moreover, atmospheric models tend to estimate the centroids of extreme storms over the Asian continent, while the satellite data suggest that, for this specific area, the most extreme rainfalls tend to occur over the Pacific ocean. These discrepancies should be carefully recognized when using these data for studies where the spatial dynamics of extreme rainfalls is an important element to consider. Future work will focus on better exploring the effects of grid resolution as well as improving the estimates of rainfall sizes and extend the analysis to other datasets.

ACKNOWLEDGMENTS

The first author acknowledges the financial support of FAPDF to attend The 9th International Workshop on Climate Informatics. This work was funded by

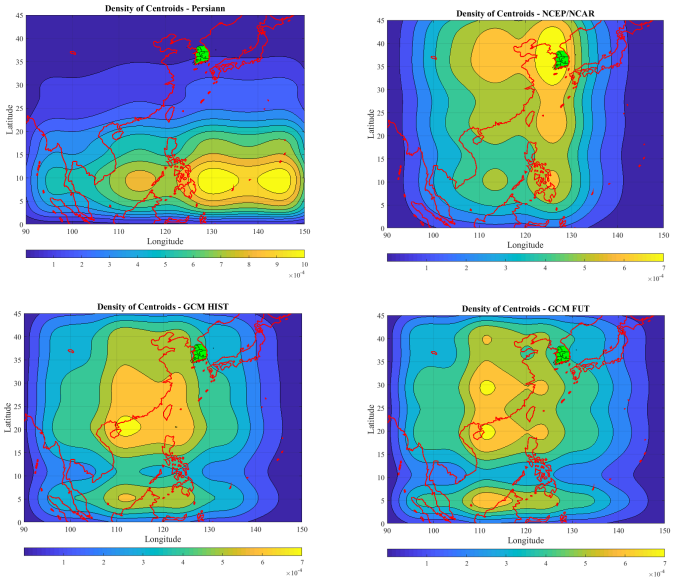


Fig. 5. Density of centroids for different datasets.

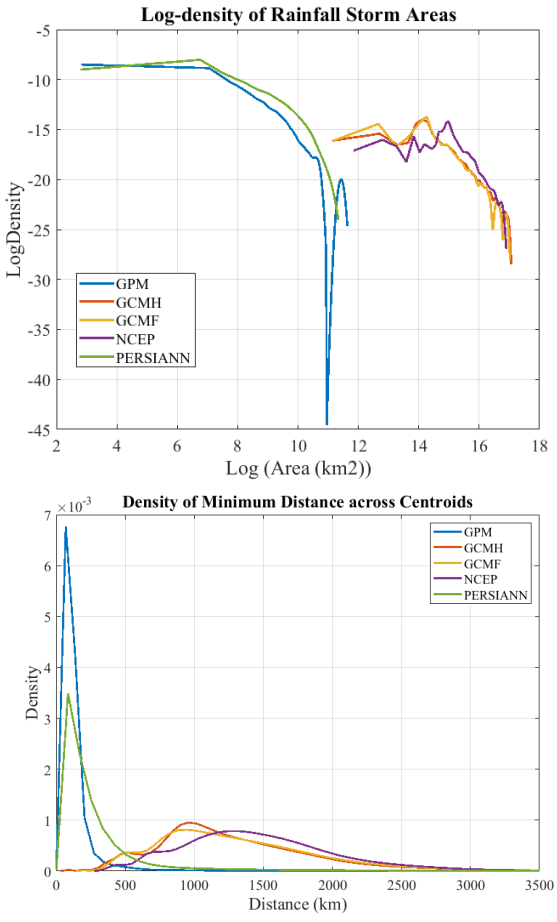


Fig. 6. Top panel: distribution of the logarithm of circular areas (in km^2) of storms. Bottom panel: distribution of the logarithm of the minimum distance (in km) among the rainfall storms.

the Korea Meteorological Administration Research and Development Program under Grant KMI2018-07010. We also thank all agencies for providing the rainfall datasets.

REFERENCES

[1] Y. Tramblay, C. Bouvier, P.-A. Ayral, and A. Marchandise, "Impact of rainfall spatial distribution on rainfall-runoff modelling efficiency and initial soil moisture conditions estimation," *Natural Hazards and Earth System Sciences*, vol. 11, no. 1, pp. 157–170, 2011.

[2] E. Cristiano, M.-C. ten Veldhuis, and N. van de Giesen, "Spatial and temporal variability of rainfall and their effects on hydrological response in urban areas – a review," *Hydrology and Earth System Sciences*, vol. 21, no. 7, pp. 3859–3878, 2017.

[3] Q. Hu, Z. Li, L. Wang, Y. Huang, Y. Wang, and L. Li, "Rainfall Spatial Estimations: A Review from Spatial Interpolation to Multi-Source Data Merging," *Water*, vol. 11, no. 3, 2019.

[4] A. F. Prein, A. Gobiet, M. Suklitsch, H. Truhetz, N. K. Awan, K. Keuler, and G. Georgievski, "Added value of convection permitting seasonal simulations," *Climate Dynamics*, vol. 41, pp. 2655–2677, Nov 2013.

[5] W. Chang, J. Wang, J. Marohnic, V. R. Kotamarthi, and E. J. Moyer, "Diagnosing added value of convection-permitting regional models using precipitation event identification and tracking," *Climate Dynamics*, Jul 2018.

[6] C. J. Matyas, "Comparing the Spatial Patterns of Rainfall and Atmospheric Moisture among Tropical Cyclones Having a Track Similar to Hurricane Irene (2011)," *Atmosphere*, vol. 8, no. 9, 2017.

[7] J. F. Booth, Y.-O. Kwon, S. Ko, R. J. Small, and R. Msadek, "Spatial Patterns and Intensity of the Surface Storm Tracks in CMIP5 Models," *Journal of Climate*, vol. 30, no. 13, pp. 4965–4981, 2017.

[8] A. Y. Hou, R. K. Kakar, S. Neeck, A. A. Azarbarzin, C. D. Kummerow, M. Kojima, R. Oki, K. Nakamura, and T. Iguchi, "The Global Precipitation Measurement Mission," *Bulletin of the American Meteorological Society*, vol. 95, no. 5, pp. 701–722, 2014.

[9] P. Nguyen, E. Shearer, H. Tran, M. Ombadi, N. Hayatbini, T. Palacios, P. Huynh, G. Updegraff, K. Hsu, B. Kuligowski, W. Logan, and S. Sorooshian, "The CHRS Data Portal, an easily accessible public repository for PERSIANN global satellite precipitation data," *Nature Scientific Data*, vol. 6, no. 180296, 2019.

[10] E. Kalnay, M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, "The NCEP/NCAR 40-Year Reanalysis Project," *Bulletin of the American Meteorological Society*, vol. 77, no. 3, pp. 437–472, 1996.

[11] T. H. D. T. G. M. Martin, N. Bellouin, W. J. Collins, I. D. Culverwell, P. R. Halloran, S. C. Hardiman, T. J. Hinton, C. D. Jones, R. E. McDonald, A. J. McLaren, F. M. O'Connor, M. J. Roberts, J. M. Rodriguez, S. Woodward, M. J. Best, M. E. Brooks, A. R. Brown, N. Butchart, C. Dearden, S. H. Derbyshire, I. Dharssi, M. Doutriaux-Boucher, J. M. Edwards, P. D. Falloon, N. Gedney, L. J. Gray, H. T. Hewitt, M. Hobson, M. R. Huddleston, J. Hughes, S. Ineson, W. J. Ingram, P. M. James, T. C. Johns, C. E. Johnson, A. Jones, C. P. Jones, M. M. Joshi, A. B. Keen, S. Liddicoat, A. P. Lock, A. V. Maidens, J. C. Manners, S. F. Milton, J. G. L. Rae, J. K. Ridley, A. Sellar, C. A. Senior, I. J. Totterdell, A. Verhoef, P. L. Vidale, and A. Wiltshire, "The HadGEM2 family of Met Office Unified Model climate configurations," *Geoscientific Model Development*, vol. 4, no. 3, pp. 723–757, 2011.

[12] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

VARIABILITY OF AIR POLLUTION (PM1) EXPLAINED USING A MACHINE LEARNING APPROACH

Roland Stirnberg[1,2], Jan Cermak[1,2], Simone Kotthaus[3], Martial Haeffelin[3], Julia Fuchs[1,2], Hendrik Andersen[1,2], Miae Kim[1,2]

***Abstract*—Air pollution and in particular high concentrations of particulate matter (PM) are known to be harmful to human health. However, the quantification of factors leading to high levels of PM remains challenging, as both anthropogenic and meteorological factors contribute to high pollution events. Here, a novel approach using a machine learning algorithm is proposed to identify and quantify drivers of concentrations of speciated particles with a diameter below 1μm (PM1) using meteorological data from the Site Instrumental de Recherche par Télédétection Atmosphérique (SIRTA) observatory located southwest of Paris and PM1 observation from an Aerosol Chemical Speciation Monitor instrument (ACSM). PM1 concentrations were modelled and effects of meteorological conditions on modelled PM1 concentrations were analyzed. Mixing layer, wind direction and temperatures showed to have high explanatory power to the model. Mixing layer is positively related to total PM1 predictions up to $\sim$800m. Winds from northeastern direction substantially increase PM1 predictions, as do low ($<\sim$5°C) temperatures. This is shown in one exemplary episode of high winter PM1 concentrations.**

I. MOTIVATION

Implications of high ambient particle concentrations on human health are well documented [1], [2] and discussions on measures to improve air quality are ongoing. Proposals include e.g. partial traffic bans [3] or the expansion of urban vegetation [4]. High pollution situations are, however, not exclusively driven by anthropogenic activities, but also due to changes in meteorological conditions. For the analysis presented here, speciated PM1 based on ACSM [5], [6] and Aethalometer measurements [7] between the years 2012-2018 and meteorological data from the SIRTA supersite (located 25km southwest of Paris [8]) are used. Measured peak pollution concentrations at the SIRTA observatory are often characterized by stagnant conditions and a mixture of high local contribution from wood burning [7] and regionally transported nitrates from traffic sources, particularly originating from Paris city [9], [10]. Air quality is furthermore influenced by meteorological parameters such as the mixing layer height (MLH), which determines the vertical distribution of aerosols in the lower troposphere, [11], [12] and precipitation, which leads to a substantial reduction in PM concentrations by wet scavenging [13].

The magnitude of effects of meteorological conditions on PM1 concentrations are difficult to constrain as meteorological parameters strongly interact with each other, hampering the isolation of the influence of individual factors. Thus, the main motivation behind this study was to detect and quantify meteorological influences on high pollution situations, eventually being able to identify conditions that favor high concentrations of particulate matter. While machine learning has been previously applied to address these points using PM10 and PM2.5 [14], this study uses speciated PM1 measurements. The differentiation in PM species allows for a detailed source attribution and helps to understand the driving mechanisms of high pollution situations. A novel statistical approach is presented, which allows to retrace the statistical model's decisions and attributes importance values to input features with respect to the model outcome. This way, meteorological influences on daily variations of air quality are identified and discussed.

II. METHOD

As the analysis focuses on day-to-day variations of PM1 concentrations, mean PM1 values are calculated for the period midday-3p.m., providing one value for

Corresponding author: R Stirnberg, Roland.Stirnberg@kit.edu
[1]Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology (KIT), [2]Institute of Photogrammetry and Remote Sensing, Karlsruhe Institute of Technology (KIT), [3]Institut Pierre Simon Laplace, École Polytechnique

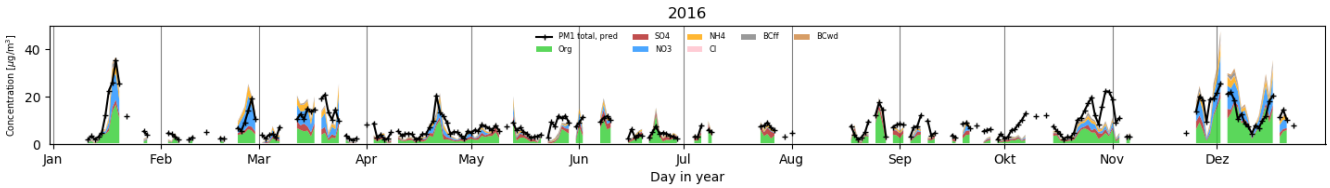


Fig. 1. Course of measured and predicted PM1 concentrations for the year 2016. Individual species are highlighted in color. Data gaps are due to measurement interruptions

each day at afternoon conditions. Species measured by the ACSM include organics (Org), ammonium (NH4), sulfate (SO4), nitrate (NO3) and chloride (Cl). The Aethalometer measures black carbon, which is differentiated in black carbon from fossil fuels (BCff) and from wood burning (BCwd) sources. Meteorological parameters used as model inputs include MLH based on Ceilometer data [15], [16], temperature in 2m height, relative humidity, air pressure, precipitation, wind speed, wind direction and total surface solar radiation measured at the SIRTA observatory. To capture the accumulation effect of missing precipitation, the hours since the last precipitation event were counted and used as input. Wind information is given as u and v components. These were normalized in order to distinguish between speed and direction information. Wind speed was then used as additional model input. Gradient Boosted Regression Trees (GBRT, [17], [18]) are used to predict daily PM1 concentrations. GBRT produce results more effectively than standard Random Forests (RF) [19], as trees are built systematically and less iterations are required [20]. GBRT hyperparameters, which determine the architecture of the statistical model, are similar to RF. One peculiar parameter that can be set is the magnitude of contribution of each tree to the model outcome, which is basically the step size of the gradient descent. Hyperparameters are tuned by executing a grid search, which tests several hyperparameter combinations and determines the best performing combination via a four-fold cross validation.

A tree regressor ”conveys” characteristics of the meteorological dataset to a statistical model, which then generalizes its learnings on an unseen data set and produces estimates of PM1. The tree regressor sets up decision trees based on a training data set. These trees split the data set along decision nodes, creating subsamples of the data while minimizing the variance of each subsample. For each subsample, regression trees fit the mean response of the observations that go into the model [20]. To increase confidence in the prediction, several decision trees are sequentially added to the ensemble [20], [21]. Each new tree that is added to the ensemble is fitted to the predecessors previous residual error using gradient descent [17], [20]. For each species and each year, one statistical model is set up. A total PM1 model is set up using the sum of all measured PM1 species. Yearly models are validated on one year subsequently to being trained on all complement years. For example, the 2016 model is trained on PM1 and meteorological data from the years 2012-2015 and 2017-2018 and validated on data from 2016. The models are well able to capture the yearly course of PM1, although peak PM1 values are not always fully reproduced. The coefficient of determination (R^2) was used to determine how much of the variation is explained by the model. Overall, R^2 ranges from 0.44 (SO_4) - 0.66 (BCff), depending on the predicted species. Predictions and observations of PM1 are largely consistent throughout the year (see figure 1).

The model is tested for its sensitivity to changes in meteorological conditions. Given that the model accurately reproduces daily variations of PM1, these sensitivity analyses allow to infer on physical processes driving variations of PM1. Therefore, the absolute contribution of each feature to PM1 predictions is derived using SHAP (SHapleyAdditive exPlanation) values [22], [23], [24]. Originally a method from gaming theory, it has been expanded to a feature attribution method suitable for tree ensembles. For each data instance, the contribution of each feature to the deviation from the average PM1 prediction is calculated. The feature contribution is calculated by training a model with that feature present, and another model with the feature withheld. Due to interactive effects between the feature and the complementary features, the effects of withholding the feature are calculated for all possible subsets [22]. This enables a quantitative analysis of effects of input features on modelled PM1 outcomes. The calculations are done using the SHAP python package (see https://github.com/slundberg/shap)

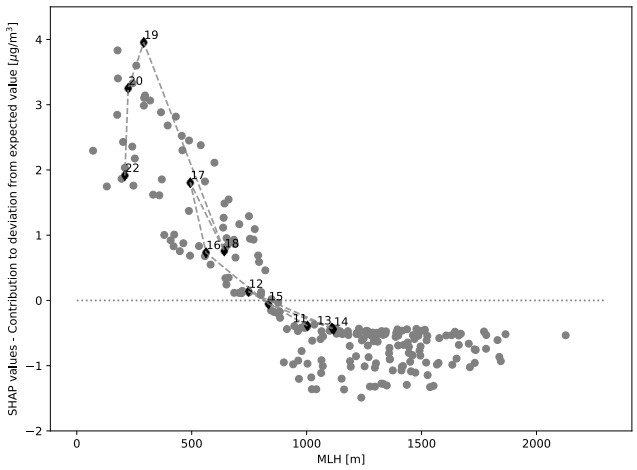


Fig. 2. SHAP values for MLH, i.e. the contribution of MLH to the deviation of modelled PM1 concentrations from the expected concentrations versus absolute MLH values. The dotted line indicates the evolution of MLH during a high pollution episode also shown in figure 3 (numbers represent the days in year)

III. EVALUATION

An analysis of the contribution of MLH to modelled PM1 concentrations in the year 2016 is shown in figure 2. It shows the SHAP values versus absolute measured MLH value, allowing to approximate the effect of variations in mixing layer height. For the modelled total PM1 values, low MLHs have the strongest positive contribution. This reflects a well-known mechanism in the atmosphere; during shallow mixing layer conditions, particles accumulate near ground as they cannot disperse to higher levels of the atmosphere and are confined to a smaller volume [11], [25]. At MLH values above ∼800m, the contribution becomes negative, thus reducing PM1 predictions as particles are transported from the near-surface and mixed within the atmosphere. The magnitude of the negative contribution above 1km remains constant. In other words, for the prediction of PM1, MLH is of importance only up to values of approximately 1km. Above that, PM1 predictions generally do not decrease with increasing MLH. Figure 2 also indicates the evolution of MLH during an exemplary episode of high PM1 (the numbers stand for the days in year), which is further analyzed in figure 3.

The exemplary episode of high PM1 concentrations (peak of 37 μg/m^3) in 2016 is shown in figure 3. Data are available for January 8th-20th and January 22nd. Modelled PM1 concentrations closely follow the measured concentrations.

In the upper panel, the total cumulative PM1 and the corresponding PM1 prediction are shown. Both agree

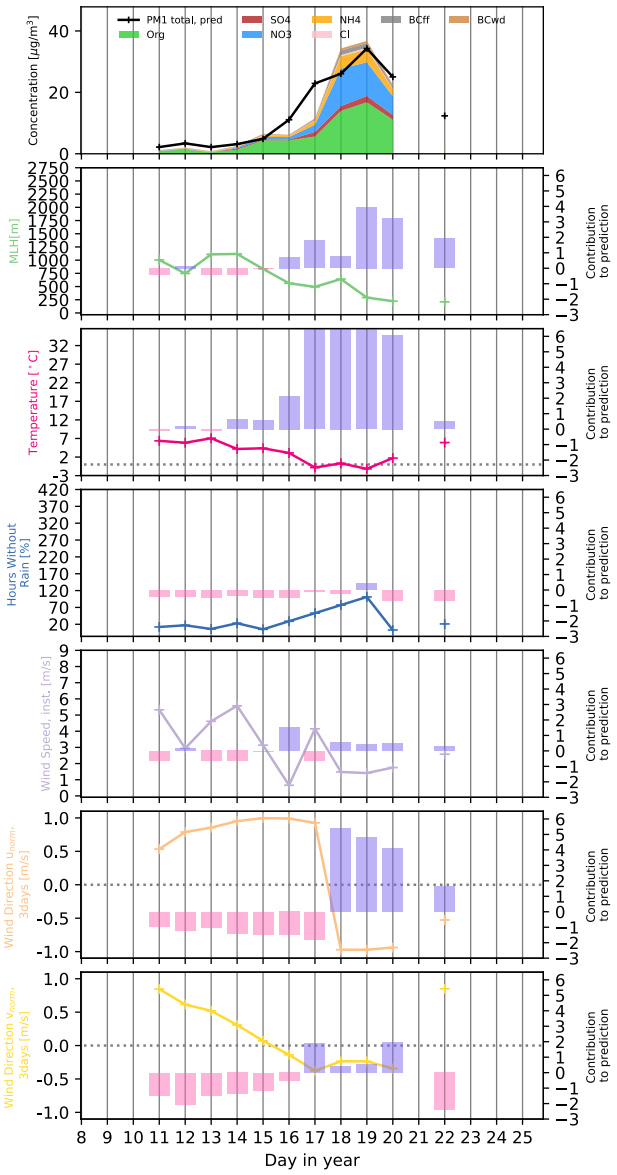


Fig. 3. Cumulative PM1 species and the corresponding PM1 prediction are shown in the upper panel. The underlying panels show absolute values of MLH [m], temperature [°C], hours without rain [h], wind speed [m/s] and normalized u and v wind component [m/s] (left y-axis). In addition, contributions of features to the final PM1 prediction are shown as columns (right y-axis).

well. In the underlying panels, absolute values of the most important input features such as MLH, temperature, hours without rain, wind speed and normalized u (v) wind component are shown (left y-axis). Positive u wind components indicate western winds, positive v components indicate southern winds. In addition, each panel shows the contribution to the PM1 prediction at each data instance as columns (right y-axis). The features' contributions reveal different processes behind the evolution of high PM1 concentrations. The episode

shown here is first characterized by a medium MLH (∼1000m), relatively mild temperatures (∼7°C), frequent precipitation and wind from the southwest. A regime change starting on day 16 causes temperatures to drop below freezing point. Low temperatures coinciding with decreasing values for MLH and wind speed indicate overall stagnant conditions. The strong contribution of very low temperatures to the overall modelled PM1 outcome could reflect increased local emissions from residential heating, in particular in the organic matter fraction [7]. In addition, low temperatures could stimulate new particle formation by shifting the gas-particle partitioning of ammonium and nitrate [9]. On day 17, the wind direction changes from west to east, eventually leading to the peak concentration of PM1. The positive contribution of northern and eastern winds (positive u and v) coincides with an increase in NO_3, pointing to a transport of particles from the Paris region [7]. The peak concentration is thus a result of both local sources and advected particles from the Paris region, mainly driven by increases in the organic fraction and NO_3. The strongest contribution of input features to the PM1 peak concentration comes from temperature and wind direction. The pollution episode then ends with rising temperatures and a change to a more southern wind direction. While MLH remains shallow, the effect of very low temperatures is missing, which prevents a higher PM1 prediction. The influence of precipitation is only marginal during this episode.

The analysis presented here will be expanded to further high pollution episodes, potentially identifying specific mechanisms leading to peak pollution concentrations.

ACKNOWLEDGMENTS

Roland Stirnberg was financially supported by the KIT Graduate School for Climate and Environment (GRACE).

REFERENCES

[1] A. Pope, R. Burnett, M. Thun, C. EE, K. D, K. I, and T. GD, "Long-term Exposure to Fine Particulate Air Pollution," *Jama*, vol. 287, no. 9, p. 1192, 2002.

[2] J. Lelieveld, K. Klingmüller, A. Pozzer, U. Pöschl, M. Fnais, A. Daiber, and T. Münzel, "Cardiovascular disease burden from ambient air pollution in Europe reassessed using novel hazard ratio functions," *Eur. Heart J.*, pp. 1–7, 2019.

[3] R. B. Ellison, S. P. Greaves, and D. A. Hensher, "Five years of London's low emission zone: Effects on vehicle fleet composition and air quality," *Transp. Res. Part D Transp. Environ.*, vol. 23, pp. 25–33, 2013.

[4] B. Bonn, E. Von Schneidemesser, D. Andrich, J. Quedenau, H. Gerwig, A. Lüdecke, J. Kura, A. Pietsch, C. Ehlers, D. Klemp, C. Kofahl, R. Nothard, A. Kerschbaumer, W. Junkermann, R. Grote, T. Pohl, K. Weber, B. Lode, P. Schönberger, G. Churkina, T. M. Butler, and M. G. Lawrence, "BAERLIN2014 -The influence of land surface types on and the horizontal heterogeneity of air pollutant levels in Berlin," *Atmos. Chem. Phys.*, vol. 16, no. 12, pp. 7785–7811, 2016.

[5] N. L. Ng, S. C. Herndon, A. Trimborn, M. R. Canagaratna, P. L. Croteau, T. B. Onasch, D. Sueper, D. R. Worsnop, Q. Zhang, Y. L. Sun, and J. T. Jayne, "An Aerosol Chemical Speciation Monitor (ACSM) for routine monitoring of the composition and mass concentrations of ambient aerosol," *Aerosol Sci. Technol.*, vol. 45, no. 7, pp. 770–784, 2011.

[6] J.-C. Dupont, M. Haeffelin, J. Badosa, T. Elias, O. Favez, J. Petit, F. Meleux, J. Sciare, V. Crenn, and J. Bonne, "Role of the boundary layer dynamics effects on an extreme air pollution event in Paris," *Atmos. Environ.*, 2016.

[7] J. E. Petit, O. Favez, J. Sciare, V. Crenn, R. Sarda-Estève, N. Bonnaire, G. Močnik, J. C. Dupont, M. Haeffelin, and E. Leoz-Garziandia, "Two years of near real-time chemical composition of submicron aerosols in the region of Paris using an Aerosol Chemical Speciation Monitor (ACSM) and a multi-wavelength Aethalometer," *Atmos. Chem. Phys.*, vol. 15, no. 6, pp. 2985–3005, 2015.

[8] M. Haeffelin, O. Bock, C. Boitel, S. Bony, D. Bouniol, H. Chepfer, M. Chiriaco, J. Cuesta, P. Drobinski, C. Flamant, M. Grall, A. Hodzic, F. Hourdin, F. Lapouge, A. Mathieu, Y. Morille, C. Naud, J. Pelon, C. Pietras, A. Protat, B. Romand, G. Scialom, and R. Vautard, "SIRTA, a ground-based atmospheric observatory for cloud and aerosol research," *Ann. Geophys.*, vol. 23, pp. 253–275, 2005.

[9] J. E. Petit, O. Favez, J. Sciare, F. Canonaco, P. Croteau, G. Močnik, J. Jayne, D. Worsnop, and E. Leoz-Garziandia, "Submicron aerosol source apportionment of wintertime pollution in Paris, France by double positive matrix factorization (PMF2) using an aerosol chemical speciation monitor (ACSM) and a multi-wavelength Aethalometer," *Atmos. Chem. Phys.*, vol. 14, no. 24, pp. 13773–13787, 2014.

[10] H. Petetin, M. Beekmann, J. Sciare, M. Bressi, A. Rosso, O. Sanchez, and V. Ghersi, "A novel model evaluation approach focusing on local and advected contributions to urban PM2.5 levels - Application to Paris, France," *Geosci. Model Dev.*, vol. 7, no. 4, pp. 1483–1505, 2014.

[11] P. Gupta and S. A. Christopher, "Particulate matter air quality assessment using integrated surface, satellite, and meteorological products: Multiple regression approach," *J. Geophys. Res.*, vol. 114, p. D14205, jul 2009.

[12] M. Bressi, J. Sciare, V. Ghersi, N. Bonnaire, J. B. Nicolas, J. E. Petit, S. Moukhtar, A. Rosso, N. Mihalopoulos, and A. Féron, "A one-year comprehensive chemical characterisation of fine aerosol (PM2.5) at urban, suburban and rural background sites in the region of Paris (France)," *Atmos. Chem. Phys.*, vol. 13, no. 15, pp. 7825–7844, 2013.

[13] J. Rost, T. Holst, E. Sahn, M. Klingner, K. Anke, D. Ahrens, and H. Mayer, "Variability of PM10 concentrations dependent on meteorological conditions," *Int. J. Environ. Pollut.*, vol. 36, no. March 2014, pp. 3–18, 2009.

[14] S. K. Grange, D. C. Carslaw, A. C. Lewis, E. Boleti, and C. Hueglin, "Random forest meteorological normalisation models for Swiss PM10 trend analysis," *Atmos. Chem. Phys.*, vol. 18, no. 9, pp. 6223–6239, 2018.

[15] S. Kotthaus and C. S. B. Grimmond, "Atmospheric boundary-layer characteristics from ceilometer measurements. Part 1: A

new method to track mixed layer height and classify clouds," *Q. J. R. Meteorol. Soc.*, vol. 144, no. 714, pp. 1525–1538, 2018.

[16] S. Kotthaus and C. S. B. Grimmond, "Atmospheric boundary-layer characteristics from ceilometer measurements. Part 2: Application to London's urban boundary layer," *Q. J. R. Meteorol. Soc.*, vol. 144, no. 714, pp. 1511–1524, 2018.

[17] J. H. Friedman, "Stochastic gradient boosting," *Comput. Stat. Data Anal.*, vol. 38, pp. 367–378, feb 2002.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine Learning in Python," jan 2012.

[19] A. Just, M. De Carli, A. Shtein, M. Dorman, A. Lyapustin, and I. Kloog, "Correcting Measurement Error in Satellite Aerosol Optical Depth with Machine Learning for Modeling PM2.5 in the Northeastern USA," *Remote Sens.*, vol. 10, no. 5, p. 803, 2018.

[20] J. Elith, J. R. Leathwick, and T. Hastie, "A working guide to boosted regression trees," no. Ml, pp. 802–813, 2008.

[21] Y. Rybarczyk, "applied sciences Machine Learning Approaches for Outdoor Air Quality Modelling : A Systematic Review," 2018.

[22] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," no. Section 2, pp. 1–10, 2017.

[23] S. M. Lundberg, G. G. Erion, and S.-I. Lee, "Consistent Individualized Feature Attribution for Tree Ensembles," no. 2, 2018.

[24] L. Shapley, "A Value for n-Person Games," in *Contrib. to Theory Games (AM-28), Vol. II* (H. W. Kuhn and A. W. Tucker, eds.), vol. 28, pp. 307–318, Princeton: Princeton University Press, dec 1953.

[25] P. Wagner and K. Schäfer, "Influence of mixing layer height on air pollutant concentrations in an urban street canyon," *Urban Clim.*, vol. 22, no. May 2018, pp. 64–79, 2017.

LEARNING CONSTRAINED DYNAMICAL EMBEDDINGS FOR GEOPHYSICAL DYNAMICS

Said Ouala[1], Steven L. Brunton[2], Duong Nguyen[1], Lucas Drumetz[1] and Ronan Fablet[1]

***Abstract*—In this work, we investigate the implementation of physical constraints for the regularization of linear quadratic dynamical representations of partially observed systems. We focus on energy preserving quadratic terms and propose to enforce this constraint within the learning criterion of the models. We further demonstrate on the Lorenz 63 system that the generalization performance is significantly improved to states beyond the attractor spanned by the observation data when this constraint is satisfied.**

I. INTRODUCTION

Recent advances in data driven modeling, especially in optimization techniques, machine learning and neural networks address the learning of data-driven representations of dynamical systems as relevant alternatives to model driven strategies for applications ranging from system identification [1], forecasting [2], reconstruction [3] and control [4]. When considering observation data issued from an a priori complex field as encountered in ocean, atmosphere and climate science, these powerful tools should be considered with care to account for the proper specifications of the underlying dynamics. For instance, when considering the data-driven identification of an Ordinary Differential Equation (ODE) from a set of observations $\mathbf{x}_t \in \mathbb{R}^n$, where $t \in \{t_0, ..., T\}$ is the temporal sampling and n the dimension of our observation space, the first question to answer is the existence (or not) of an appropriate ODE mapping in the observation space. For fully-observed systems, *i.e.* when observed variables $\mathbf{x}_t$ are governed by an ODE or are related to some underlying states $\mathbf{z}_t$ that are governed by an ODE according to a diffeomorphic mapping, recent advances [1], [5], [6] have shown that one can identify the governing equations of the dynamics of $\mathbf{z}$ from a representative dataset of observations $\{\mathbf{x}_{t_i}\}_i$. However, in the more general cases, it is more likely that our observations depend (possibly in a non-linear fashion) on unobserved latent variables that make the underlying dynamical model evolve in a higher dimensional space $\mathbb{R}^s$ with $s > n$. Under the assumption that the relationship between the observed and unobserved variables can not be decoupled, it is rigorously impossible to find an appropriate *one-to-one* mapping governed by an ODE in the observation space $\mathbb{R}^n$. In the latter case, classical approaches do not apply since no ODE or, more generally, no one-to-one mapping defined in the observation space can represent the time evolution of the observations.

In this context, Takens's theorem states the conditions under which a delay embedding representation guarantees the existence of governing equations in the embedded space [7]. This technique was initially used as a geometrical reconstruction technique of the higher dimensional unobserved limit-cycle. The derivation of a dynamical system from such a representation, on the other hand, encountered large disparities since no explicit relationships between the defined phase space and an ODE formulation have been clearly identified.

The identification of an embedding of the observations parametrized by an ODE was proposed in [8] and appears to be an interesting trade-off between reconstructing the phase space of the unseen dynamical system and forecasting the observations through the parametric ODE. However, this formulation is very limited when considering generalization issues above the limit-cycle described by the observations. From a topological point of view, and without loss of generality, one can expect the ODE representation to i) be bounded, ii) only include the limit-cycle describing the observations in a higher dimensional space with a reasonable attracting region. Unfortunately, those characteristics relate to some physical constraints that define trapping regions of limit-cycles. The optimization criterion as proposed in [8] does not guarantee those elementary constraints which severely affects the generalization quality of the models. In this work, we propose a new implementation of the learning algorithm that allows to enforce prior

Corresponding author: S. Ouala, said.ouala@imt-atlantique.fr
[1]IMT-Atlantique, Lab STICC, UMR CNRS 6285, F-29238, France,
[2]Department of Mechanical Engineering, University of Washington, Seattle, WA 98195,USA

knowledge such as physical constraints. We focus on energy preserving non-linearities and illustrate on a toy model whose the long term boundedness and the attracting region of the revealed limit cycle are highly influenced when this type of constraints are satisfied. Regarding the data driven identification of climate and ocean dynamics, we believe that this work provides an initial playground for learning consistent models in terms of long term forecast through the implementation of physical constraints issued from prior knowledge of the conservation laws governing the dynamics.

This paper is organized as follows. Section II reviews the Neural Embedding for Dynamical Systems technique as proposed in [8]. Section III introduces the new optimization criterion of the model which includes energy preservation constraints. The numerical experiments are presented in Section IV. We further discuss our contributions in section V.

II. Neural Embedding of Dynamical Systems

This section summaries the Neural Embedding of Dynamical Systems —NbedDyn— proposed in [8].

Let us consider a dynamical system governed by and autonomous ODE:

$$\dot{\mathbf{z}}_t = f_H(\mathbf{z}_t) \tag{1}$$

For most applications, the true state $\mathbf{z}_t \in \mathbb{R}^s$ of the system is unknown and we are only provided a series of observations $\{\mathbf{x}_t\}$:

$$\mathbf{x}_t = h(\mathbf{z}_t) \tag{2}$$

Where $h : \mathbb{R}^s \rightarrow \mathbb{R}^n$ is an observation operator that does not satisfy the conditions [9] under which the predictable deterministic dynamics expressed in the space of $\mathbf{z}$ is still deterministic in the observation space.

The NbedDyn technique tackles this problem by searching an augmented latent space, where the latent states are governed by diffeomorphic flows and can be mapped to the observations $\mathbf{x}_t$. For any given operator h of a deterministic dynamical system, Takens's theorem [7] guarantees that such augmented space exists. However, instead of using a delay embedding, NbedDyn defines a d_E-dimensional augmented latent space with states ($d_E > n$) $\mathbf{X}_t \in \mathbb{R}^{d_E}$ as follows:

$$\mathbf{X}_t{}^T = [\mathbf{x}_t^T, \mathbf{y}_t^T] \tag{3}$$

where $\mathbf{y}_t \in \mathbb{R}^{d_E - n}$ presents the information of the unobserved components of the true latent state $\mathbf{z}_t$.

The corresponding dynamics and observation operator are defined as:

$$\dot{\mathbf{X}}_t = f_\theta(\mathbf{X_t}) \tag{4}$$

$$\mathbf{x}_t = G(\mathbf{X_t}) \tag{5}$$

where the dynamical operator f_θ belongs to a family of operators parametrized by a parameter vector θ. Using an integration scheme, we can associate f_θ with an one-step-ahead diffeomorphic mapping:

$$\Phi_{\theta,t}(\mathbf{X}_{t-1}) = \mathbf{X}_{t-1} + \int_{t-1}^{t} f_\theta(\mathbf{X}_{t-1}) \tag{6}$$

From Eqs. (4), (5) and (6), we have a state space model:

$$\begin{cases} \mathbf{X}_t &= \Phi_{\theta,t}(\mathbf{X}_{t-1}) \\ \mathbf{x}_t &= G(\mathbf{X_t}) \end{cases} \tag{7}$$

with G a projection matrix that satisfies $\mathbf{x}_t = G(\mathbf{X_t})$. Given an observation time series $\{\mathbf{x}_0, \ldots, \mathbf{x}_T\}$, the Neural Embedding of Dynamical Systems model minimizes the forecasting error of the observations with respect to the model parameters and the augmented states as follows

$$\begin{aligned} \hat{\theta}, \mathbf{y}_{1:T} = \arg\min_{\theta} \min_{\{\mathbf{y}_t\}_t} & \sum_{t=1}^{T} \|\mathbf{x}_t - G\left(\Phi_{\theta,t}\left(\mathbf{X}_{t-1}\right)\right)\|^2 \\ & + \lambda \|\mathbf{X}_t - \Phi_{\theta,t}(\mathbf{X}_{t-1})\|^2 \end{aligned} \tag{8}$$

with λ a trade-off parameter.

The ODE operator f_θ is stated as a linear quadratic neural network and the corresponding flow map $\Phi_{\theta,t}$ is a neural network based on a numerical integration scheme formulation (typically a 4^{th}-order Runge-Kutta scheme).

III. Constrained Dynamical Embedding

The dynamical model f_θ is expressed as a linear quadratic model. This particular architecture is suitable for the identification of reduced order models of incompressible flows as it can be seen as a low dimensional approximation of the Navier-Stokes equation. Formally, we can formulate the operator f_θ as follows

$$\dot{\mathbf{X}}_i = c_i + \sum_{j=1}^{d_E} l_{i,j}\mathbf{X}_j + \sum_{j=1}^{d_E}\sum_{k=1}^{d_E} b_{i,j,k}\mathbf{X}_j\mathbf{X}_k \tag{9}$$

with $c_i, l_{i,j}, b_{i,j,k}$ are the trainable coefficients of the dynamical operator (θ), the time index t of $\mathbf{X}$ is omitted for simplicity.

Regarding the data-driven identification of the parameters θ, the minimization of the cost function in (8) does not guarantee f_θ to satisfy elementary conservation constraints present in the true underlying system which severely affects the generalization performance of the model. This is a classical issue in most data driven representations. If the provided data is not big enough for the model to learn these constraints one should

explicitly enforce them within the optimization criterion of the model. Here, we will focus on energy preserving non-linearities as this constraint is known to help long-term boundness of reduced order models of non-compressible flow [10], however, the general framework proposed here applies to any prior knowledge (known coefficients, existing symmetries ...etc.) on this specific parametrization of the network f_θ.

Let us consider the evolution of the fluctuation energy $K = \frac{1}{2}\sum_{i=1}^{d_E} X_i^2$ of the system described by f_θ. The time derivative of this quantity can be written as:

$$\begin{aligned}\dot{K} &= [\nabla_X \mathbf{K}]^T \dot{\mathbf{X}} = \sum_{i=1}^{d_E} \mathbf{X}_i f_{\theta,i} \\ &= \sum_{i=1}^{d_E} c_i \mathbf{X}_i + \sum_{i,j=1}^{d_E} l_{i,j}\mathbf{X}_i\mathbf{X}_j + \sum_{i,j,k=1}^{d_E} b_{i,j,k}\mathbf{X}_i\mathbf{X}_j\mathbf{X}_k\end{aligned} \tag{10}$$

An energy preserving quadratic non linearity satisfies the constraint:

$$\sum_{i,j,k=1}^{d_E} b_{i,j,k}\mathbf{X}_i\mathbf{X}_j\mathbf{X}_k = 0 \tag{11}$$

i.e. the contribution of the quadratic terms of f_θ to the fluctuation energy should sum up to zero. In this case, the quadratic coefficients are responsible for redistributing the perturbation energy in directions of positive and negative energy growth that are defined by the eigenvalues of the matrix $(l_{i,j}), i, j = 1, ..., d_E$ [11]. It can be shown that for constraint (11) to hold, the sums of the quadratic coefficients over index permutations must be zero:

$$\begin{aligned}&b_{i,j,k} + b_{i,k,j} + b_{j,i,k} + b_{j,k,i} + b_{k,i,j} + b_{k,j,i} = 0, \\ &i, j, k = 1, ..., d_E\end{aligned} \tag{12}$$

From the above definition of the energy-preserving non-linearity and the corresponding energy function, one can think of two distinct ways to enforce this constraint on the approximate model, either through enforcing the constraint described in (12) as a penalty term over the quadratic coefficients in the loss function (8) or penalizing the quadratic energy expressed by (11). The latter implementation of the constraint is avoided in this work since the loss function is optimized with respect to both the parameters of the model and the latent states. This will result in an expression (11) that is not necessarily minimized due to the constraint over the quadratic weights (12) as the latent states in (11) are also trained to minimize the quadratic energy. This represents an issue in the sense that our energy preserving constraint will depend on the latent states and thus, for large deviations from the spanned manifold, this constraint will no longer be satisfied. Finally, the following criterion is considered:

$$\begin{aligned}\hat{\theta}, \mathbf{y}_{1,...,T} = \arg\min_\theta \min_{\{\mathbf{y}_t\}_t} &\sum_{t=1}^{T} \|\mathbf{x}_t - G\left(\Phi_{\theta,t}\left(\mathbf{X}_{t-1}\right)\right)\|^2 \\ &+ \lambda\|\mathbf{X}_t - \Phi_{\theta,t}(\mathbf{X}_{t-1})\|^2 \\ \text{s.t. } \{\ &b_{i,j,k} + b_{i,k,j} + b_{j,i,k} + b_{j,k,i} + b_{k,i,j} + b_{k,j,i} = 0\end{aligned} \tag{13}$$

the constrained optimization problem is solved by using the equality constraint as a penalty term in the loss function.

IV. NUMERICAL EXPERIMENTS

Considered system : Lorenz-63 dynamical system is a 3-dimensional model governed by the following ODE:

$$\begin{cases} \frac{dz_{t,1}}{dt} &= \sigma\left(z_{t,2} - z_{t,2}\right) \\ \frac{dz_{t,2}}{dt} &= \rho z_{t,1} - z_{t,2} - z_{t,1}z_{t,3} \\ \frac{dz_{t,3}}{dt} &= z_{t,1}z_{t,2} - \beta z_{t,3} \end{cases} \tag{14}$$

Under parametrization $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$, this system involves chaotic dynamics with a strange attractor [12].

We simulate Lorenz-63 state sequences using the LOSDA ODE solver [13] with an integration step of 0.01. We assume that only the first Lorenz-63 variable is observed $\mathbf{x}_t = \mathbf{z}_{t,1}$. We apply the proposed framework to this experimental setting using a training sequence of 10000 time-steps.

Proposed model : Regarding the proposed framework, we tested the model for a dimension of the latent space equal to 3. The neural-network parametrization for operator f_θ is a simple linear quadratic model. We compare in this work the model optimized using the initial criterion (8) as proposed in [8] and the new criterion with the energy preserving constraint (13).

Forecasting performances of the proposed data-driven models: We further evaluate the performances of the learning criterion based on the comparison of the foretasted limit-cycles. Table I reports the Lyapunov spectrum and the Lyapunov dimension of the data driven models of the proposed NbedDyn representation compared with the true spectrum and dimension of the Lorenz 63 system. As demonstrated in [8], when the initial condition is inside the spanned manifold of the augmented states, the dynamical model optimized using criterion (8) gives trajectories that are bounded (the sum of the Lyapunov exponents is negative) and with topological characteristics that are very similar to the true Lorenz 63 model. However, when the initial condition is far from the spanned manifold, the model optimized by the equation (8) diverges to infinity (the

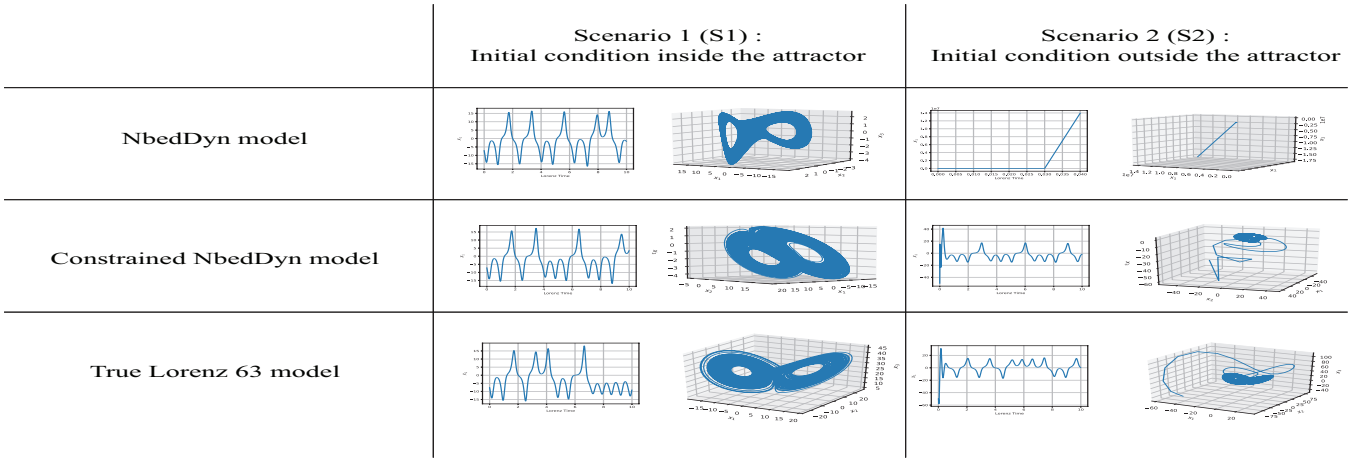


Fig. 1: ***Forecasting performances of the data driven models under different initial conditions:*** *first row, NbedDyn model as proposed in [8]; second row, proposed constrained NbedDyn model; third row, True lorenz 63 model.*

sum of the Lyapunov exponents is positive). From a machine learning perspective, this is the direct consequence of a poor generalization performance to states that are far from the attractor spanned by the training data. From a dynamical systems point of view, our model contains several attracting regions of chaotic and unstable solutions and when the initial condition is far from the spanned attractor, the state evolution is dominated by positive energy growth which makes our model diverge to infinity. The constrained model in the other hand, satisfies elementary preservation constraints that are present in the actual Lorenz 63 system and leads to a much more stable behavior with a larger attracting region of the chaotic limit-cycle.

Qualitative analysis of the proposed schemes: We also illustrate these conclusions through the forecasting examples in Figure 1. When starting from an initial condition inside the attractor, both the NbedDyn and the Constrained NbedDyn models end up with a forecasted limite cycle that is similar to the true Lorenz attractor. When starting from an initial condition that is far from the the spanned attractor, the classical NbedDyn as proposed in [8] diverge to infinity. By contrast, adding energy preserving constraints to the model significantly improves the generalization performances to states beyond the attractor spanned by the training data.

V. CONCLUSION

In this work, we address the data-driven identification of dynamical representations of partially observed systems. We propose to include physical constraints to the data driven models as prior knowledge of the dynamics. The reported forecasting performance for Lorenz-63 dynamics illustrates clearly the importance of such an approach as only enforcing energy preserving nonlinearity constraints significantly improves the generalization performances of the model far from the attractor spanner by the training data.

Model		Exponents	Dimension
NbedDyn	S1	(0.889, 0.0, -14.21)	2.063
	S2	NaN	NaN
Constrained NbedDyn	S1	(0.811, 0.0, -12.66)	2.064
	S2	(0.828, 0.0, -12.67)	2.064

TABLE I: ***Forecasting performance the data-driven models****: full Lyapunov spectrum and Lyapunov dimension of the NbedDyn model as proposed in [8] and with the additional energy preserving constraint on the quadratic terms of f_θ. The simulation S1 is carried with respect to an initial condition inside of the spanned attractor of the augmented states $\mathbf{X}_t$, the simulation S2 is performed with respect to an initial condition far from the attractor. The Lyapunov spectrum of the true Lorenz 63 system is (0.91, 0.0, -14.57) and it's dimension is estimated to be 2.064 [14]*

ACKNOWLEDGMENTS

This work was supported by GERONIMO project (ANR-13-JS03-0002), Labex Cominlabs (grant SEACS), CNES (grant OSTST-MANATEE), Microsoft (AI EU Ocean awards) and by MESR, FEDER, Région Bretagne, Conseil Général du Finistère, Brest Métropole and Institut Mines Télécom in the framework of the VIGISAT program managed by "Groupement Bretagne Télédétection" (BreTel).

REFERENCES

[1] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proceedings of the National Academy of Sciences*, vol. 113, pp. 3932–3937, Apr. 2016.

[2] A. Braakmann-Folgmann, R. Roscher, S. Wenzel, B. Uebbing, and J. Kusche, "Sea Level Anomaly Prediction using Recurrent Neural Networks," *arXiv:1710.07099 [cs]*, oct 2017. arXiv: 1710.07099.

[3] H. C. C. B. P. A. C. F. G. L. Ouala Said, Fablet Ronan, "Neural network based kalman filters for the spatio-temporal interpolation of satellite-derived sea surface temperature," 2018.

[4] S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Sparse identification of nonlinear dynamics with control (sindyc)," *IFAC-PapersOnLine*, vol. 49, no. 18, pp. 710–715, 2016.

[5] R. Fablet, S. Ouala, and C. Herzet, "Bilinear residual neural network for the identification and forecasting of geophysical dynamics," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 1477–1481, Sep. 2018.

[6] D. Nguyen, S. Ouala, L. Drumetz, and R. Fablet, "Em-like learning chaotic dynamics from noisy and partial observations," *SciRate*, Mar. 2019.

[7] F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980* (D. Rand and L.-S. Young, eds.), (Berlin, Heidelberg), pp. 366–381, Springer Berlin Heidelberg, 1981.

[8] S. Ouala, D. Nguyen, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier, and R. Fablet, "Learning Latent Dynamics for Partially-Observed Chaotic Systems," *arXiv e-prints*, p. arXiv:1907.02452, Jul 2019.

[9] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, pp. 579–616, Nov 1991.

[10] J.-C. Loiseau and S. L. Brunton, "Constrained sparse galerkin regression," *Journal of Fluid Mechanics*, vol. 838, pp. 42–67, 2018.

[11] M. Schlegel and B. R. Noack, "On long-term boundedness of galerkin models," *Journal of Fluid Mechanics*, vol. 765, pp. 325–352, 2015.

[12] E. N. Lorenz, "Deterministic Nonperiodic Flow," *Journal of the Atmospheric Sciences*, vol. 20, pp. 130–141, Mar. 1963.

[13] A. C. Hindmarsh, "ODEPACK, a systematized collection of ODE solvers," *IMACS Transactions on Scientific Computation*, vol. 1, pp. 55–64, 1983.

[14] J. C. Sprott, *Chaos and Time-Series Analysis*. New York, NY, USA: Oxford University Press, Inc., 2003.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

LEARNING THE HIDDEN DYNAMICS OF OCEAN TEMPERATURE WITH NEURAL NETWORKS

Ibrahim Ayed[*] [1,2], Emmanuel de Bézenac[*] [1], Arthur Pajot[1], Julien Brajard[3, 4], Patrick Gallinari[1,5]

***Abstract*—Climatic phenomena often exhibit complex spatio-temporal physical processes described by a state whose evolution may be unknown, and which very often not fully measurable. In this work, we formulate a data-driven continuous framework where we model the dynamics as a differential equation parametrized by a neural network, optimized through partial observations of the system. We apply our method on two challenging dynamical systems which are crucial for climatic modeling: the classical incompressible Navier-Stokes equations and a more realistic state-of-the-art dataset of ocean simulations. This allows us to show that not only does our model consistently outperform classical baselines for forecasting observations, it also recovers the unobserved state dynamics when given access to the initial state. Moreover, in the absence of the initial we show that our method remains very accurate when only observations are given as input and that it produces an interpretable hidden state even in this case.**

I. INTRODUCTION

A. Partially observed differential equations

Let us consider a dynamical system describing a real-world physical process. There generally is a certain set of variables of interest $\{a_i\}$ which spatiotemporal dynamics needs to be followed. The *state* X_t ($t \in \mathbb{R}_+$ stands for the time) of the dynamical system can then be defined as a vector-valued function of space $x \in \Omega$ such that there exists a function F such that:

$$\forall t, \ \frac{dX_t}{dt} = F(X_t) \tag{1}$$

and that, for all i, $a_i(t)$ can be calculated easily from X_t[1]. Thus, X_t is sufficient to describe its own temporal dynamics. For example, in the incompressible Navier Stokes equations for which a state can be the concatenation of the density and velocity fields of the fluid as those are enough for the evolution of the system, while pressure, which can be an additional variable of interest, can be computed from those variables.

With the availability of very large amounts of data captured via diverse sensors and recent advances of statistical methods, a new data-driven paradigm for modelling dynamical systems is emerging, where relations between the states are no longer handcrafted, but automatically discovered based on the available observations. This problem can be approached by considering some class of admissible functions $\{F_\theta\}$, and looking for a θ such that the solution X^θ of :

$$\frac{dX_t}{dt} = F_\theta(X_t) \tag{2}$$

fits the measured data. This approach has motivated some recent work for exploiting machine learning in order to solve differential equations. For example, [1] parameterizes F_θ as sparse regression over a set of predefined candidate differential terms, [2], [3], [4] or [5] use statistical models such as Gaussian processes and neural networks to model F_θ and learn a solution to the corresponding equation. However, previous methods have essentially considered the case where the state of the system X_t is fully-observed at all times t while, for many real-world applications, the entire state of the system is not fully visible to external sensors: one usually only has access to low-dimensional projections of the state, *i.e. observations*. Intuitively, the latter can be seen as what is readily and easily measurable; this means that, in contrast with the ideal case where the full state can be observed at all times with perfect certainty, there is an important loss of information. This issue is a major one in many fields within applied sciences [6], [7].

More formally, the available data is only a projection of the complete state X_t. This observation process can be modelled with an operator $\mathcal{H}_t$, linking the system's state X_t to the corresponding observation Y_t. In the following,

[*]Equal Contribution. Corresponding author: I. Ayed, ibrahim.ayed@lip6.fr [1]Sorbonne Université, UMR 7606, LIP6, F-75005 Paris, France [2]Theresis lab, Thales, Thales Research & Technology Route Départementale, 91120 Palaiseau [3]Sorbonne Université, CNRS-IRD-MNHN, LOCEAN, Paris, France [4]Nansen Environmental and Remote Sensing Center, Bergen, Norway [5]Criteo AI Lab, Paris, France

[1]In other words, $a_i(t) = f_i(X_t)$ for a certain deterministic measuring function f_i

$\mathcal{H}_t = \mathcal{H}$ is supposed known, constant in time and differentiable. Let us note that, generally, the observation process represents a considerable loss of information compared to the case where X is available, as the measurements may be sparse, and low-dimensional.

B. Reconstructing a state from observations

One question which arises in the partially observable setting is whether it is possible to reconstruct the dynamics of a system given only observations. In 1981, the Takens embedding theorem [8] proved that this is indeed possible in many cases with reasonably mild assumptions and this result was then refined over the years to cover most of the interesting use cases [9]. In practice, the theorem implies that, for almost any reasonable observation function $\mathcal{H}$ and corresponding observations at discrete times $(Y_{t_i})_i$, there is a k such that we can reconstruct the state X_{t_j} from the k observations $(Y_{t_{j-k+1}}, \cdots, Y_{t_j})$.

Let us stress that, if only observations Y are available, without additional constraints, we can only hope to reconstruct the state up to a smooth diffeomorphism. Indeed, with our definition above, for any ϕ a smooth diffeomorphism of $\mathbb{R}^d$, a straightforward calculation shows that $f_i^\phi = f_i \circ \phi^{-1}$ and $F^\phi = \mathrm{d}\phi_\cdot \circ F(\cdot)$ verify the conditions so that $\phi \circ X$ is also a valid state. Thus, for any given system, there is an infinity of possible states describing it which are conjugated through a diffeomorphism.

Thus, our problem can be framed as the following: Using only observations Y at training, estimate a model which produces future states X_t for $t \geq 0$ and takes as initial input $\{Y^{(-k)}, \breve{X}_0\}$, where $Y^{(-k)} = (Y_{-k+1}, ..., Y_0)$ are k past observations and $\breve{X}_0$ is a proxy for the initial state X_0, and produces future states X_t for $t \geq 0$.

This general formulation covers two important cases that will be treated in the experiments:

- when $\breve{X}_0 = X_0$, we theoretically do not have to use past observations as we already have a perfect initial state;
- when $\breve{X}_0$ is empty, we have to rely exclusively on past observations to construct a state.

For the problem to be meaningful, we suppose that the observations are given by the function $\mathcal{H}$ fall under the conditions of Takens theorem.

II. Methodology

In this section, we outline the learning algorithm, making use of classical tools from continuous-time optimal control.

A. Optimization Problem

Our goal is to learn the differential equation driving the dynamics of a smooth state function X for which we only have supervision over observations Y through a fixed operator $\mathcal{H}$. In order to ensure our dynamical system at least explains the observations, we define a cost functional of the form :

$$\mathcal{J}(Y, \widetilde{Y}) = \int_0^T \|Y_t - \widetilde{Y}_t\|^2 \mathrm{dt} \tag{3}$$

Here, Y is a spatio-temporal field representing observations of the studied system, $\widetilde{Y}$ the output of the system, and $\|\cdot\|$ the norm associated to the standard dot product over space.

Since the state X_t is constrained to follow the dynamics described by equation 2, starting from its initial condition X_0, the optimization problem is in fact a constrained one :

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \mathbb{E}_{Y \in \text{Dataset}} \left[\mathcal{J}(Y, \mathcal{H}(X))\right] \\ \text{subject to} \quad & \frac{dX_t}{dt} = F_\theta(X_t), \\ & X_0 = g_\theta(Y^{(-k)}, \breve{X}_0) \end{aligned} \tag{4}$$

where F_θ is a smooth vector valued function defining the trajectory of X, and g_θ gives us the initial condition X_0 from the input[2] $(Y^{(-k)}, \breve{X}_0)$. In particular, if a full initial state is given as input to the system, g_θ can be taken as independent of any parameter and doesn't need to be learned.

For any θ, we assume that F and g are chosen such that there always exists a unique solution to the equation given as a constraint in (4) up to time T. In the following, we will call such a solution X_t^θ.

B. Training

In order to use gradient descent, one possible method is to use the *adjoint state equation*:

$$\frac{\partial}{\partial\theta}\mathcal{J}(Y, \mathcal{H}(X^\theta)) = -\int_0^T \left\langle \lambda_t, \partial_\theta F_\theta(X_t^\theta) \right\rangle \mathrm{dt} - \langle \lambda_0, \partial_\theta g_\theta \rangle \tag{5}$$

where λ is solution of :

$$\partial_t \lambda_t = A_t \lambda_t + B_t \tag{6}$$

solved backwards, starting with $\lambda_T = 0$, and where :

$$A_t = -(\partial_X F_\theta(X_t^\theta))^\star$$

and

$$B_t = 2(\partial_X \mathcal{H}(X_t^\theta))^\star(\mathcal{H}(X_t^\theta) - Y_t)$$

where $M^\star$ denotes the adjoint operator of linear operator M.

[2] In other words, θ parameterizes both the dynamics through F and the initialization through g.

When training, for a given value of θ, we can solve the forward (2) to find X^θ. Then, λ can be solved backwards as its equation only depends on X^θ which gives us all necessary elements to calculate the gradient of $\mathcal{J}$. This gives us the following iterative algorithm to solve the optimization problem, starting from a random initialization of θ to reach the optimal value $\theta^\star$:

1) If needed, estimate initial state X_0^θ with g_θ;
2) Solve the forward state equation (2) to find X^θ;
3) Solve the backward adjoint equation (6) to find the corresponding λ;
4) Update θ in the steepest descent direction using equation (5).

At inference, we use the learned $g_{\theta^\star}$ to compute an initial state then simply solve the forward equation with the learned[3] $F_{\theta^\star}$.

In practice, discretizing this procedure in order to implement this procedure can be done either by using classical numerical solvers for the forward and backward equations or by solving the forward equation with a differentiable solver which can then be combined with backpropagation to estimate the gradient.

III. EXPERIMENTS

A. Practical details

We test our model on two datasets: the incompressible Navier Stokes equations and the Glorys2v4 which is a more challenging and complex simulation of ocean dynamics. We decompose the training simulations into training sequences of fixed length, using 6 timesteps for the target sequence. In practice, the cost functional $\mathcal{J}$ is estimated on a minibatch of sequences from the dataset and optimized using stochastic gradient descent. For all datasets, the split between train, validation and test sets is made so that each split only includes sequences generated by *different, independently sampled* initial conditions. Thus, the scores reported and images shown are computed over sequences in conditions the models have never seen.

Throughout *all* the experiments, F_θ is a standard residual network [10], with 2 downsampling layers, 6 residual blocks, and bilinear up-convolutions instead of transposed convolutions. To discretize the forward equation (2) in time, we use a simple three steps Euler scheme. For the spatial discretization, we use the standard gridlike discretization induced by the dataset. The weights of the residual network θ are initialized using an orthogonal initialization. Our model is trained using an exponential scheduled sampling scheme with exponential decay, using the Adam optimizer, with a learning rate set to 1×10^{-5}. We use the Pytorch deep learning library [11].

To evaluate our model's performance we consider the quality of the predictions, using the renormalized mean-squared error between generated and ground-truth observations, averaged over the time sequence, and spatial coordinates.

To evaluate the quality of the hidden states, we use cosine similarity between the model's hidden state u and the true hidden state of the system v

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\Omega|}\sum_{x\in\Omega}\frac{\langle u(x), v(x)\rangle}{\|u(x)\|\,\|v(x)\|} \tag{7}$$

For the velocity vector field representation, color represents the angle, and the intensity the magnitude of the associated vectors.

Moreover, we propose two variants of our models :

- **Ours (with $X_0 = \breve{X}_0$)**. This variant receives X_0 as an input and only learns F_θ.
- **Ours, Joint Training (with $\breve{X}_0$ empty)** . This variant receives only observations $Y^{(-k)}$ as input and learns g_θ as well as F_θ.

and compare them to 2 different baselines:

- **PKnI [12]** This hybrid model uses an advection-diffusion equation to link the velocity with the observed temperatures, and uses a neural network to estimate the velocities.
- **PRNN [13]** This is heavy-weight, state of the art model used for video prediction tasks. It is based on a Spatiotemporal LSTM that models spatial deformations and temporal variations simultaneously.

B. Forecasting the Navier Stokes equations

TABLE I
RELATIVE MSE AND COSINE SIMILARITY SCORES FOR THE NAVIER STOKES EQUATIONS

MODEL	H=5		H=10		H=50	
	MSE	COSINE	MSE	COSINE	MSE	COSINE
OURS	**0.118**	**0.798**	**0.180**	**0.679**	**0.628**	**0.483**
OURS, JOINT TRAINING	0.152	0.201	0.243	0.192	0.650	0.183
PKNI	0.194	0.243	0.221	0.207	0.752	0.098
PRNN	0.170	×	0.227	×	0.719	×

Figure 1 shows a sample of the predictions of our system over the test set for the Navier Stokes equations. Our model consistently forecasts observations with very good accuracy and the quantitative evaluation in table I corroborate this claim. More surprisingly, it also

[3] In particular, it is important to note that no further updates or corrections are made and no additional observations are needed.

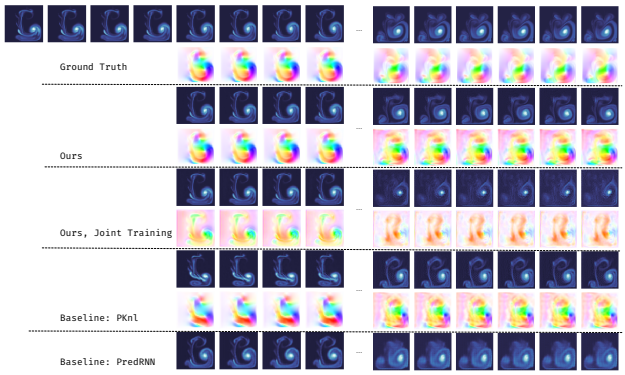


Fig. 1. Forecasting the Navier Stokes equations 30 time-steps ahead with different models, starting from a given initial condition.

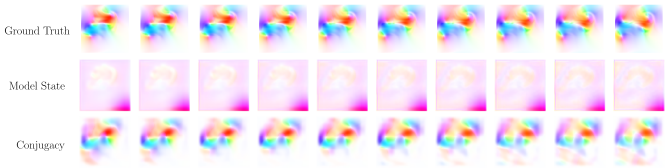


Fig. 2. Example of a sequence of hidden states transformed by the calculated conjugacy.

accurately forecasts the dynamics of the velocity even though it was not observed during training in the case where $\breve{X}_0 = X_0$.

Regarding the state learned by the Jointly Trained model, we analyze its structure by considering $Z = g_\theta(Y^{(-k)})$ the representation it learns and then, following the considerations of section I-B, given that there must exist a certain smooth diffeomorphism ϕ which transforms Z into the true state X, we learn ϕ by parameterizing it with a neural network and by taking a small subset of the Navier Stokes dataset as a training set Figure 2 shows an example of output from the conjugacy we learned: It allows us to transform the non-structured hidden states of the jointly trained model into interpretable states corresponding to the canonical representation. It also means that the learned state Z is equivalent to the canonical state X. From a quantitative point of view, after 5 predictions, the average cosine similarity over the whole test set goes from 0.192 in the jointly trained representation to 0.582 when transformed by conjugacy. This means that, when a model is trained without any available states, if a small dictionary of states can be constituted later, it is possible to learn the diffeomorphism and produce meaningful full states.

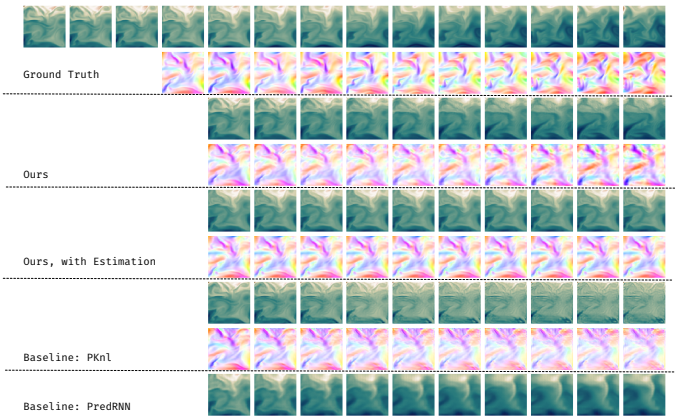


Fig. 3. Forecasting Sea Surface Temperatures 10 time-steps ahead with different models, starting from a given initial condition.

C. Ocean circulation and Sea Surface Temperatures

The second and more challenging dataset is taken from a very complex simulation which uses reanalyzed real data, encompassing a vast portion of ocean dynamics. Moreover, data is taken from a small zone of the ocean which implies that the boundary conditions change in time: This is a limit case of our framework as it violates the stationarity hypothesis. In order to accommodate for those difficulties, we modify g_θ to include a proxy for part of the state as well as previous observations:

$$g_\theta = E_\theta(Y^{(-L)}, \tilde{w}_0) + \begin{pmatrix} Y_0 \\ \breve{w}_0 \\ 0 \end{pmatrix} \tag{8}$$

where E_θ is an encoder neural network. Using E_θ allows us to encode available information from the observations $Y^{(-L)}$ which is not contained in $\breve{w}_0$, which is a proxy for the velocity field, nor in Y_0. For E_θ, we use the UNet architecture [14]. This variant shows the potential of our method to be used in settings of varying difficulties.

Figure 3 and table II show that our models are accurate both for observations and hidden states. However, there is a clear drop in performance when compared to the other two datasets, even at horizon 10.

TABLE II
RELATIVE MSE AND COSINE SIMILARITY SCORES FOR THE GLORYS2V4 DATASET

MODEL	H=5		H=10	
	MSE	COSINE	MSE	COSINE
OURS	**0.306**	0.671	**0.402**	0.589
OURS, EST.	0.364	**0.718**	0.490	**0.670**
PKNI	0.411	0.448	0.494	0.368
PRNN	0.423	XX	0.546	XX

REFERENCES

[1] S. H. Rudy, S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Data-driven discovery of partial differential equations," *Science Advances*, vol. 3, p. e1602614, Apr. 2017.

[2] M. Raissi, P. Perdikaris, and G. E. Karniadakis, "Machine learning of linear differential equations using gaussian processes," *Journal of Computational Physics*, vol. 348, pp. 683 – 693, 2017.

[3] M. Raissi, "Deep hidden physics models: Deep learning of nonlinear partial differential equations," *Journal of Machine Learning Research*, vol. 19, 2018.

[4] M. Bocquet, J. Brajard, A. Carrassi, and L. Bertino, "Data assimilation as a deep learning tool to infer ODE representations of dynamical models," *Nonlinear Processes in Geophysics Discussions*, pp. 1–29, 2019.

[5] Z. Long, Y. Lu, X. Ma, and B. Dong, "PDE-Net: Learning PDEs from Data," in *ICML*, pp. 3214–3222, 2018.

[6] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen, "Data assimilation in the geosciences: An overview of methods, issues, and perspectives," *Wiley Interdisciplinary Reviews: Climate Change*, vol. 9, no. 5, p. e535, 2018.

[7] A. Lorenc, "Analysis methods for numerical weather prediction," *Quarterly Journal of the Royal Meteorological Society*, vol. 112, pp. 1177 – 1194, 10 1986.

[8] T. F., "Detecting strange attractors in fluid turbulence.," *Symposium on dynamical systems and turbulence*, pp. 366–381, 1981.

[9] J. C. Robinson, *Dimensions, Embeddings, and Attractors*. Cambridge Tracts in Mathematics, Cambridge University Press, 2010.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778, 2016.

[11] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[12] E. de Bézenac, A. Pajot, and P. Gallinari, "Deep learning for physical processes: Incorporating prior scientific knowledge," in *ICLR*, 2018.

[13] Y. Wang, Z. Gao, M. Long, J. Wang, and P. S. Yu, "Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning," 2018.

[14] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," *CoRR*, vol. abs/1505.04597, 2015.

CROSS-INFORMATION KERNEL CAUSALITY: REVISITING GLOBAL TELECONNECTIONS OF ENSO OVER SOIL MOISTURE AND VEGETATION

Diego Bueso, Maria Piles, Gustau Camps-Valls

***Abstract*—Identifying causal relations from observational data is of wide interest in Earth and climate sciences. Among the many existing approaches, Granger causality (GC) has been perhaps the most widely used approach. Noting its strong linearity and stationarity assumptions, several nonlinear alternatives to GC have been proposed based on kernel methods. Here we formalize the field by introducing a new cross-information kernel Granger causality (XKGC) test, that accounts for nonlinear cross-relation between the involved variables and generalizes the previous nonlinear GC methods. The proposed XKGC is successfully evaluated in toy examples and in the identification of impacts of ENSO on observational global soil moisture and vegetation data.**

I. INTRODUCTION

Causal discovery from observational Earth and climate data is of paramount relevance. Compared to other data-driven *machine learning* methods, such as probabilistic modeling [1], kernel machines [2], or deep learning [3], which mainly focus on prediction and classification, *causal inference* methods aim at discovering and explaining the causal structure of the underlying system [4], [5]. Causality promises to revolutionize the field of Earth sciences with applications in hypothesis testing, model-data intercomparison and attribution of extreme impacts [6].

Among the many existing methods and approaches, Granger causality (GC) [7] is the most widely used approach. GC was a first attempt to formalize quantitatively the causal relation between time series. Noting the (strong) linearity assumption in GC, many nonlinear extensions have been proposed. Non-linear relations using kernel methods were originally introduced in [8], while later [9] presented an alternative kernel-based test in combination with a filtering approach to avoid overfitted solutions. Recently, alternative nonlinear probabilistic GC based on Gaussian processes (GPs) [1] have been also introduced [10], [11]. In all cases, the autoregressive (AR) models simply implement nonlinear kernel autoregression schemes, instead of linear autoregression, but often do not account *explicitly* for nonlinear cross-relations between the involved variables.

In this paper we formalize the field of kernel methods for GC, and introduce several alternative kernel functions that account for nonlinear cross-relations and alleviate the nonstationary issue. The proposed kernel function here is rooted on standard density estimation [12], and measures sample similarities with the past and future of each variable but also between them.

We show empirical evidence of performance in a toy example, and in a real experiment with spatio-temporal Earth data. Using seven years of global soil moisture (SM) and vegetation optical depth (VOD) data derived from the SMOS satellite, we illustrate how the proposed XKGC provides an efficient alternative to uncovering causal dry/wet footprints of ENSO globally.

II. CROSS-KERNEL GRANGER CAUSALITY

A. Notation and Granger causality

The intuition behind GC is to test whether the past of X helps in predicting the future of Y from the past of Y, and for that one builds univariate and bivariate AutoRegressive (AR) models:

$$y_{t+1} = \sum_{k=0}^{p} a_k y_{t-k} + \varepsilon_t^y,$$

$$y_{t+1} = \sum_{k=1}^{p} a_k y_{t-k} + \sum_{l=1}^{q} b_l x_{t-l} + \varepsilon_t^{y|x},$$

and computes a GC test as the ratio of model fitting errors:

$$\delta_{x\to y} = \log(\mathbb{V}[\varepsilon_t^y]^2/\mathbb{V}[\varepsilon_t^{y|x}]^2).$$

The time embeddings p and q are fixed to define $\mathbf{y}_t = [y_t, y_{t-1}, \ldots, y_{t-p}]^\top$ and $\mathbf{x}_t = [x_t, x_{t-1}, \ldots, x_{t-q}]^\top$, and the vector coefficients $\mathbf{a} = [a_1, \ldots, a_p]^\top$ and $\mathbf{b} = [b_1, \ldots, b_q]^\top$ are estimated normally by least squares . The residual errors are defined for unrestricted (ε_t^y) and restricted ($\varepsilon_t^{y|x}$) cases separately. Yet, the main

Research funded by the European Research Council (ERC) under the ERC-CoG-2014 SEDAL project (grant agreement 647423), and by projects TEC2016-81900-REDT and RTI2018-096765-A-100 (MCIU/AEI/FEDER, UE). M. Piles is supported by a Ramón y Cajal contract (MICINN).

Authors are with the Image Processing Laboratory, Universitat de Valéncia, Spain. http://isp.uv.es/, e-mail: diego.bueso@uv.es.

problems in GC is to find a regression model that represent faithfully the dependence with the conditional variable X.

B. Cross-kernel Granger causality

Nonlinear GC has been mainly approached with either transfer entropy measures [13] or by kernel methods [8], [9]. We first briefly review the traditional kernel GC methods, and then introduce a cross-information kernel for generalization and enhanced sensitivity.

a) Stacked kernel: The standard kernel GC (KGC) approach considers a kernel-based AR modeling [9]. The method works with concatenated feature vectors $\mathbf{z}_t = [\mathbf{y}_t, \mathbf{x}_t] \in \mathbb{R}^{p+q}$, and then defines two feature maps $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$ to a reproducing kernel Hilbert space $\mathcal{H}$ (RKHS) of (possibly infinite) dimensionality H where $\mathbf{y}_t$ and $\mathbf{z}_t$ are mapped to, and which are endorsed with reproducing kernels k and ℓ, respectively. The kernel (nonlinear) regression models are thus:

$$y_{t+1} = \mathbf{a}_H^T \phi(\mathbf{y}_t) + \varepsilon_t^y, \qquad y_{t+1} = \mathbf{b}_H^T \psi(\mathbf{z}_t) + \varepsilon_t^{y|x},$$

where now $\mathbf{a}_H, \mathbf{b}_H \in \mathbb{R}^{H\times 1}$. Secondly, one defines the span using the representer's theorems $\mathbf{a}_H = \boldsymbol{\Phi}^\top \boldsymbol{\alpha}$ and $\mathbf{b}_H = \boldsymbol{\Psi}^\top \boldsymbol{\beta}$, where $\boldsymbol{\Phi}, \boldsymbol{\Psi} \in \mathbb{R}^{n\times H}$ contain all instances n. Then, by linear algebra operations, the AR models can be defined in terms of kernel functions only:

$$y_{t+1} = \boldsymbol{\alpha}^\top \mathbf{k}_t + \varepsilon_t^y, \quad y_{t+1} = \boldsymbol{\beta}^\top \boldsymbol{\ell}_t + \varepsilon_t^{y|x},$$

where $\mathbf{k}_t = [k(\mathbf{y}_1, \mathbf{y}_t), \ldots, k(\mathbf{y}_n, \mathbf{y}_t)]^\top$ and $\boldsymbol{\ell}_t = [\ell(\mathbf{z}_1, \mathbf{z}_t), \ldots, \ell(\mathbf{z}_n, \mathbf{z}_t)]^\top$ contain all evaluations of k and ℓ at time t.

b) Summation kernel: An alternative builds *implicit* AR models in RKHS [8] such that:

$$y_{t+1} = \mathbf{a}_H^T \phi(\mathbf{y}_t) + \varepsilon_t^y$$

$$y_{t+1} = \mathbf{a}_H^T \phi(\mathbf{y}_t) + \mathbf{b}_H^T \psi(\mathbf{x}_t) + \varepsilon_t^{y|x},$$

which leads to the following kernel ARMA models:

$$y_{t+1} = \alpha^\top \mathbf{k}_t + \varepsilon_t^y, \quad y_{t+1} = \alpha^\top \mathbf{k}_t + \beta^\top \boldsymbol{\ell}_t + \varepsilon_t^{y|x},$$

where now $\boldsymbol{\ell}_t := [\ell(x_1, \mathbf{x}_t), \ldots, \ell(x_n, \mathbf{x}_t)]^\top$. The summation kernel is more appropriate when large embeddings p and q are needed to capture long memory, as it avoids constructing large dimensional feature vectors $\mathbf{z}$ by concatenation. However, the cross-information between X and Y random variables is missing, as pointed out in [14].

c) Explicit cross-kernel: In the context of system identification of nonlinear systems, we proposed in [15] a full kernel-based ARMA framework. To account for the cross-information, one explicitly defines a *joint feature mapping* $\boldsymbol{\phi}(\mathbf{x}_t, \mathbf{y}_t)$ for the second AR model:

$$y_{t+1} = \mathbf{a}_H^T \boldsymbol{\phi}(\mathbf{y}_t) + \varepsilon_t^y, \qquad y_{t+1} = \mathbf{b}_H^T \boldsymbol{\phi}(\mathbf{x}_t, \mathbf{y}_t) + \varepsilon_t^{y|x},$$

where

$$\boldsymbol{\phi}(\mathbf{x}_t, \mathbf{y}_t) = [\mathbf{A}_1 \boldsymbol{\varphi}(\mathbf{y}_t), \mathbf{A}_2 \boldsymbol{\varphi}(\mathbf{x}_t), \mathbf{A}_3(\boldsymbol{\varphi}(\mathbf{y}_t) + \boldsymbol{\varphi}(\mathbf{x}_t))]^\top,$$

and $\boldsymbol{\varphi}$ is a nonlinear feature map into an RKHS $\mathcal{H}$, and $\mathbf{A}_i$, $i = 1, 2, 3$, are three linear transformations from $\mathcal{H}$ to $\mathcal{H}_i$. The corresponding kernel function can be readily computed as:

$$\begin{aligned} n(\mathbf{z}_t, \mathbf{z}_t') &= \boldsymbol{\phi}(\mathbf{x}_t, \mathbf{y}_t)^\top \boldsymbol{\phi}(\mathbf{x}_t, \mathbf{y}_t) \\ &= \boldsymbol{\varphi}(\mathbf{y}_t)^\top \mathbf{R}_1 \boldsymbol{\varphi}(\mathbf{y}_t') + + \boldsymbol{\varphi}(\mathbf{x}_t)^\top \mathbf{R}_2 \boldsymbol{\varphi} \mathbf{x}_t') \\ &\quad + \boldsymbol{\varphi}(\mathbf{y}_t)^\top \mathbf{R}_3 \boldsymbol{\varphi}(\mathbf{x}_t') + \boldsymbol{\varphi}(\mathbf{x}_t)^\top \mathbf{R}_3 \boldsymbol{\varphi}(\mathbf{y}_t'), \\ &= n_1(\mathbf{y}_t, \mathbf{y}_t') + n_2(\mathbf{x}_t, \mathbf{x}_t') \\ &\quad + n_3(\mathbf{y}_t, \mathbf{x}_t') + n_4(\mathbf{x}_t, \mathbf{y}_t'), \end{aligned}$$

where $\mathbf{R}_1 = \mathbf{A}_1^\top \mathbf{A}_1 + \mathbf{A}_3^\top \mathbf{A}_3$, $\mathbf{R}_2 = \mathbf{A}_2^\top \mathbf{A}_2 + \mathbf{A}_3^\top \mathbf{A}_3$, and $\mathbf{R}_3 = \mathbf{A}_3^\top \mathbf{A}_3$. Note that the new kernel function considers all cross-terms relations between the time series, and still works with the original time embeddings. We call this the cross-kernel GC (XKGC) method, which generalizes all previous kernel GC methods.

C. On the kernel functions

Among the great many kernel functions available [16], the squared exponential kernel, $k_\sigma(\mathbf{a}, \mathbf{b}) = \exp(-\|\mathbf{a} - \mathbf{b}\|^2/(2\sigma^2))$ is the preferred choice because it summarizes all higher order moments of similarity and it only requires selecting the length-scale parameter σ. In this work, however, we chose the standard Nadayara-Watson (NW) density estimator as a kernel function $k_{\text{NW}}(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^n y_i k_\sigma(\mathbf{a}, \mathbf{b}_i) / \sum_{j\neq i}^n k_\sigma(\mathbf{a}, \mathbf{b}_j)$, where the denominator ensures sum-to-one densities. To test the robustness of the non-parametric method, the statistical significance is assessed with surrogate data methods by selecting a random subsection of the data of approximately 13 of the predictor variable length [17]. The threshold is chosen as the higher causal strength estimated from the surrogate time series.

III. EXPERIMENTAL RESULTS

We show performance of the proposed method in a simulation example with controlled nonlinear coupled processes, and in assessing causal relations and teleconnection patterns between ENSO, SM and VOD.

A. Simulated data

Here we test our approach with a bivariate dynamic example. We used a non-linear AR model with a multiplicative coupling term between X and Y:

$$x_{t+1} = 3.4x_t(1-x_t^2)\exp(-x_t^2) + \varepsilon_t^x$$
$$y_{t+1} = 3.4y_t(1-y_t^2)\exp(-y_t^2) + 0.5x_t y_t + \varepsilon_t^{y|x},$$

where $\varepsilon_t^x \sim \mathcal{N}(0, \sigma_x^2)$ and $\varepsilon_t^{y|x} \sim \mathcal{N}(0, \sigma_y^2)$. Note in this example the model is coupled in the direction $X \to Y$ but not in the inverse way. We run a standard KGC and our XKGC for different noise ratios, $r = \varepsilon^x/\varepsilon^{y|x}$. We expect a high significance when $\varepsilon_t^x \gg \varepsilon_t^{y|x}$ and a null causal identification when $\varepsilon_t^x \ll \varepsilon_t^{y|x}$. Experiment assumed $p = 2$ for all cases and 50 realizations for each test were averaged.

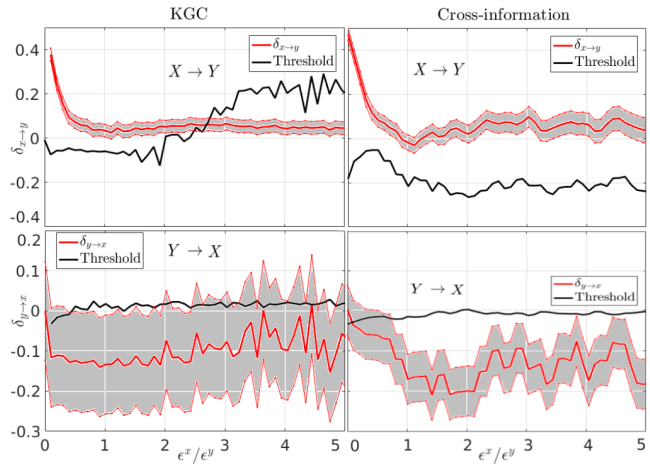


Fig. 1. . Estimated causal strength δ for two kernel methods (KGC and XKGC) and the two causal directions $(\delta_{x\to y}, \delta_{y\to x})$. The threshold was assessed by random surrogates. A δ lower than 0 or the threshold, means a non-significant GC relation.

Results are show in Figure 1. The XKGC achieves a lower false-positive rate in the case of $Y \to X$ independently of the noise rate. In the case of $X \to Y$, for a low Y noise regime, results are virtually similar between methods, while for high Y noise regimes, the standard KGC method is unable to identify the causal relation, unlike the proposed XKGC which finds a low (yet significant) causal relation.

B. ENSO-SM-VOD causal patterns

Here we are interested in uncovering (potentially nonlinear) GC relations between ENSO and two relevant Earth observation variables: soil moisture and vegetation optical depth. Since its launch in 2009, SMOS provides SM and VOD products every 3-days with a spatial resolution of ∼50 km. For this work we used SMOS-IC SM and VOD data [18] covering the first seven years of the mission, after its commissioning phase (from May 2010 to January 2017). To ensure enough coverage and smooth spatio-temporal transitions, ascending and descending orbits were temporally averaged into 5-day bins. Pixels with less than 30% temporal coverage and latitudes higher than 60° were not considered. Alongside SMOS data, in this experiment we use the climate index ENSO4, which is calculated daily based on the Sea Surface Temperature (SST) anomalies measured on the western Pacific Ocean.

Recent studies have improved our knowledge – previously solely based on models– about the global relation between ENSO events and SM long-trends [19], [20]. However, even the most advanced EO-based teleconnection patterns are generally derived from correlation analyses only. In this work, we apply the XKGC causal method to uncover ENSO dry/wet footprints over soil moisture and vegetation. Our analyses involve four main processing steps:

1) *Spatio-temporal dimensional reduction* of the SM and VOD data cubes is performed using the ROCK-PCA method [21]. ROCK-PCA generalizes standard methods (like PCA/EOF, complex EOF, Promax, Varimax and their nonlinear versions) and yields complex temporal components with associated spatial distribution. The interannual components of SM and VOD are selected for this work (see Fig. 2a). Note that there is a spatial phase that defines which mixture of the real and imaginary components is represented in a given pixel. This spatial phase is defined as ϕ_{SM} and ϕ_{VOD} in Fig. 2b, and allows to depict the regions where the causation links are significant (Fig. 2c).
2) *GC relations between ENSO-SM-VOD and their time-lags* are obtained using the proposed XKGC test. To extract the causal link between each pair or variables, we need to consider each combination of spatial phase, within the range $[-\pi, \pi]$. In the case of the SM-VOD causal relations, we have two different phases to explore. However, ϕ_{SM} is almost invariant under ϕ_{VOD} changes, and therefore we only show δ dependent with ϕ_{VOD} (see Fig. 2b). The time embedding was chosen as the maximum $\delta_{x\to y}$ within the first half of the temporal data (250 time samples), and was used as a measure of cause-effect lag.
Figure 3 shows the learned causal graph along their time embeddings, which appear to be dependent with the spatial phase: we found that the estimated lag between ENSO4 and SM is between 50-85 days, 55-70 days for the SM causing VOD, and ∼75 days between ENSO4 causing VOD. All causal indices have a similar GC strength.

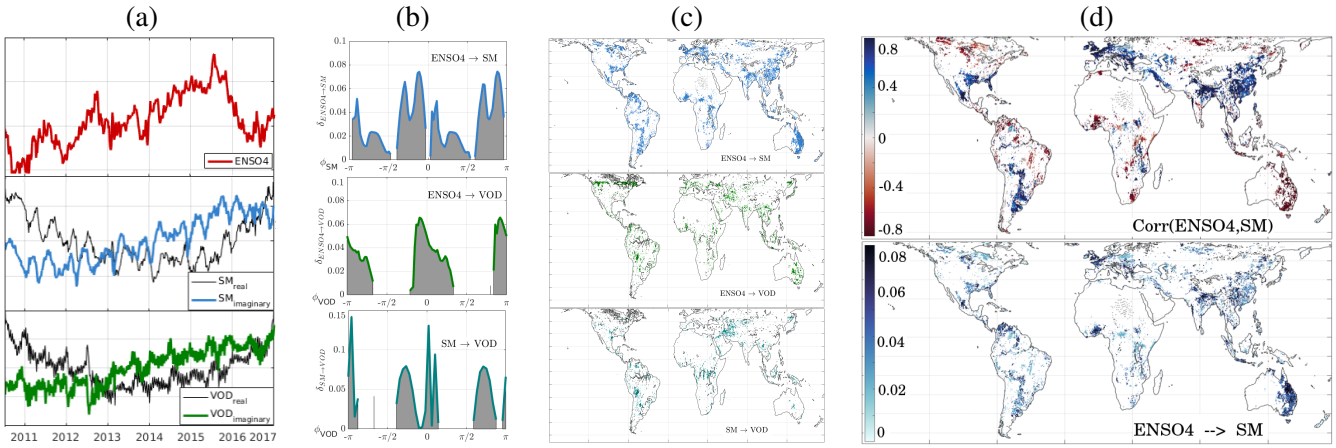


Fig. 2. (a) Time series of ENSO4, SM and VOD to explore their causal relations. SM and VOD are summarized in the interannual component extracted with ROCK-PCA [21], which yields a real (0 or π) and imaginary ($\pm\pi/2$) counterparts. (b) Significant XKGC causal tests δ as a function of the spatial phases. Only one direction is shown for each pair of variables because the other causal direction was not statistical significant. (c) Spatialization of each causal pair of variables (only statistically significant regions are shown, δ below the threshold). (d) 5-day lagged correlation (top) and causal (bottom) maps of ENSO4 and SM inter-annual components.

3) *Spatial teleconnection patterns* are obtained from the relation between spatial phases (ϕ_{SM} or ϕ_{VOD}) for each causal link; they are shown in Fig. 2c. Since we have the relation between spatial phase and the causality index (Fig. 2b), its spatialization just requires the spatial phase map. It is worth noting that regions affected by the ENSO events [20] are represented in the spatial distribution of ENSO4 with SM and VOD, as for example the East coast of Australia [22] or Nigeria [23], but not in the SM-VOD causal map, which indicates a good separation of ENSO impact and the SM-VOD causal link.
4) *Causality uncovers spurious correlations*. We show the 5-day lagged correlation map (the one leading to highest correlation) between ENSO4 and the long-trend SM signal as well as the causal map obtained using XKGC in Fig. 2d. As can be observed, many of the correlations are not causal, even the highest ones $\rho \sim 0.8$, thus suggesting mere spurious associations. In regions where SM is correlated but not caused by ENSO, other potential causes such as global warming or anthropogenic actions could explain the footprints.

IV. CONCLUSION

In this paper we proposed a framework for kernel Granger causality. The proposed cross-information kernel functions account for more complex relations and dynamics, and introduce significant gains in sensitivity. We illustrated the performance with a toy example, and in the reconstruction of global climate teleconnection patterns of ENSO over SM and VOD spatio-temporal observational data. Further work is tied to the inclusion of more variables related to vegetation activity, such as gross primary productivity, radiation and sun-induced fluorescence, as well as to test the algorithm under episodes of vegetation water stress and disturbances.

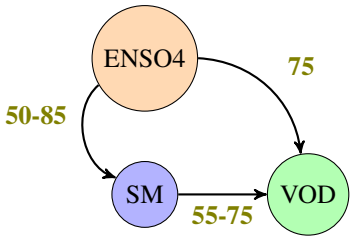


Fig. 3. Reconstructed causal graph with the XKGC method along the time lags ranges (in days) between ENSO4 and the interannual components of SM and VOD.

REFERENCES

[1] G. Camps-Valls, D. Sejdinovic, J. Runge, and M. Reichstein, "A perspective on Gaussian processes for earth observation," *National Science Review*, 2019.
[2] J. Rojo-Álvarez, M. Martínez-Ramón, J. Muñoz-Marí, and G. Camps-Valls, *Digital Signal Processing with Kernel Methods*. UK: Wiley & Sons, Apr 2018.
[3] M. Reichstein, G. Camps-Valls, B. Stevens, J. Denzler, N. Carvalhais, M. Jung, and Prabhat, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, pp. 195–204, Feb 2019.
[4] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT press, 2nd ed., 2000.
[5] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference - Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning Series, Cambridge, MA, USA: MIT, 2017.

[6] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Clymour, M. Kretschmer, M. Mahecha, J. Muñoz-Marí, E. van Nes, J. Peters, R. Quax, M. Reichstein, M. Scheffer, B. Schölkopf, P. Spirtes, G. Sugihara, J. Sun, K. Zhang, and J. Zscheischler, "Inferring causation from time series with perspectives in Earth system sciences," *Nature Communications*, vol. 10, no. 2553, 2019.

[7] C. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, pp. 424–438, 1969.

[8] N. Ancona, D. Marinazzo, and S. Stramaglia, "Radial basis function approach to nonlinear granger causality of time series," *Phys. Rev. E*, vol. 70, p. 056221, Nov 2004.

[9] D. Marinazzo, M. Pellicoro, and S. Stramaglia, "Kernel method for nonlinear granger causality," *Phys. Rev. Lett.*, vol. 100, p. 144103, Apr 2008.

[10] P. Amblard, O. Michel, C. Richard, and P. Honeine, "A Gaussian process regression approach for testing Granger causality between time series data," in *Acoustics, Speech, and Signal Processing, 1988. ICASSP-88., 1988 International Conference on*, pp. 3357–3360, 03 2012.

[11] L. Ambrogioni, M. Hinne, M. A. J. van Gerven, and E. Maris, "GP CaKe: Effective Brain Connectivity with Causal Kernels," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (USA), pp. 951–960, Curran Associates Inc., 2017.

[12] N. Nicolaou and T. G. Constandinou, "A nonlinear causality estimator based on non-parametric multiplicative regression," *Frontiers in Neuroinformatics*, vol. 10, p. 19, 2016.

[13] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, pp. 461–464, Jul 2000.

[14] X. Sun, "Assessing nonlinear granger causality from multivariate time series," in *Machine Learning and Knowledge Discovery in Databases* (W. Daelemans, B. Goethals, and K. Morik, eds.), (Berlin, Heidelberg), pp. 440–455, Springer Berlin Heidelberg, 2008.

[15] M. Martínez-Ramón, J. L. Rojo-Álvarez, G. Camps-Valls, J. Muñoz-Marí, n. Navia-Vázquez, E. Soria-Olivas, and A. R. Figueiras-Vidal, "Support vector machines for nonlinear kernel arma system identification," *IEEE Transactions on Neural Networks*, vol. 17, pp. 1617–1622, Nov 2006.

[16] G. Camps-Valls, J. Verrelst, J. Muñoz-Marí, V. Laparra, F. Mateo-Jiménez, and J. Gómez-Dans, "A survey on Gaussian Processes for Earth observation data analysis," *IEEE Geoscience and Remote Sensing Magazine*, vol. 4, no. 2, pp. 58–78, 2016.

[17] R. Quian Quiroga, A. Kraskov, T. Kreuz, and P. Grassberger, "Performance of different synchronization measures in real data: A case study on electroencephalographic signals," *Phys. Rev. E*, vol. 65, p. 041903, Mar 2002.

[18] R. Fernández-Morán, A. Al-Yaari, A. Mialon, A. Mahmoodi, A. Al Bitar, G. De Lannoy, N. Rodríguez-Fernandez, E. López-Baeza, Y. Kerr, and J.-P. Wigneron, "SMOS-IC: An Alternative SMOS Soil Moisture and Vegetation Optical Depth Product," *Remote Sensing*, vol. 9, no. 5, 2017.

[19] D. G. Miralles, M. J. V. D. Berg, J. H. Gash, R. M. Parinussa, R. A. M. D. Jeu, H. E. Beck, T. R. H. Holmes, C. Jiménez, N. E. C. Verhoest, W. A. Dorigo, and et al., "El Niño−La Niña cycle and recent trends in continental evaporation," *Nature Climate Change*, vol. 4, Aug 2013.

[20] S. Brands, "Which ENSO teleconnections are robust to internal atmospheric variability?," *Geophysical Research Letters*, vol. 44, no. 3, pp. 1483–1493, 2017.

[21] D. Bueso, M. Piles, and G. Camps-Valls, "Nonlinear PCA for Spatio-Temporal Analysis of Earth Observation Data," *IEEE Transactions on Geoscience and Remote Sensing*, 2019.

[22] E. Forootan, Khandu, J. Awange, M. Schumacher, R. Anyah, A. V. Dijk, and J. Kusche, "Quantifying the impacts of ENSO and IOD on rain gauge and remotely sensed precipitation products over Australia," *Remote Sensing of Environment*, vol. 172, 2016.

[23] O. R. Salau, A. Fasuba, K. A. Aduloju, G. E. Adesakin, and A. T. Fatigun, "Effects of changes in enso on seasonal mean temperature and rainfall in nigeria," *Climate*, vol. 4, no. 1, 2016.

MULTI-TASK LEARNING VIA LATENT BASIS TASKS AND CONSTRAINED PRECISION MATRIX

Yumin Liu[1], Auroop Ganguly[2], Jennifer Dy[1]

***Abstract*—Many interesting climate variables such as observational temperatures, precipitations and air pressure can be regarded as location-based time series. The prediction at each location separately can be viewed as a single task learning (STL). However, this may not be a good idea in many cases since in real world many tasks are related to each other and share some common information. By dealing with different tasks jointly, we may gain some benefits. Multi-task learning (MTL) provides us with tools to take advantage of share information between tasks. In this paper we proposed a MTL method that enables us to estimate the relationship between tasks and learn the basis tasks. This method assumes that the weights of an observed task is a linear combination of several latent basis tasks and that the task relationships can be learnt by imposing a regularized precision matrix. Experiments on different data sets shown that this method outperforms competing methods.**

I. MOTIVATION

Climate change has growing impact on human activities and people are collecting climate variables at a larger scale involving more and more geological locations. The most common idea to predict/forecast those variables is to view each location as an individual task and deal with each task separately, such as fitting a regression for each location. This is a single task learning (STL) procedure. However, it may miss some information since very often the variables in different locations are correlated, especially for variables in those locations that are near to each other. For example, when predicting temperatures in an area, we naturally assume that the temperatures in the close neighbourhood locations maybe similar to each other since we know that temperature tend to vary smoothly over small areas. In those cases Multi-task learning (MTL) can help to gain benefits from sharing information between tasks. In MTL multiple tasks are modeled and learnt at the same time and it can improve generalization performance by utilizing the common information between related tasks and distinguish differences between specific tasks [1].

The essential problem of MTL is to capture the shared structure. However, in climate science most relationships between tasks are not explicitly nor rigorously known and need to be learnt or estimated. A general way to do this is to constrain the parameters matrix with some kind of regularization such as L1 norm, L2 norm or trace norm. Variations of the regularization can induce different sharing structures, such as reducing the number of common features for all the tasks [2], [3], grouping in the MTL [4]–[7], and sparsity [8], [9]. Task relationship can also be learnt through a shared covariance function in the context of Gaussian Process [10].

To learn the shared structure, Zhang *et al.* [11] proposed a method for learning both positive and negative task correlations in MTL. In their formulation, the objective function is convex and therefore a global optimum can be obtained. Gonçalves *et al.* [8] proposed a similar formulation which jointly estimate the task relationship structure and the individual task parameters. By imposing L1 norm on both the relation structure matrix and the parameter matrix, this method allows a sparse relationship among tasks and sparse feature dependency for each task. Kumar *et al.* [12] proposed an idea that the parameter vector of each observed task is a linear combination of several latent basis tasks and allow tasks to selectively share information by overlapping between groups.

In this paper, inspired by [8] and [12], we incorporate the benefit of both the latent basis tasks pattern and the learnt task relationship structure. The method can estimate the task parameters and task relationships simultaneously and tasks can "partially" share information between tasks. This formulation assumes that each observed task is a linear combination of a set of latent basis tasks and the coefficients of the linear combination are possibly sparse for each task, which means that each

Corresponding author: Yumin Liu, yuminliu@ece.neu.edu
[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, USA [2]Department of Civil and Environmental Engineering , Northeastern University, Boston, USA

task consists of a few basis tasks instead of all the basis tasks. In addiction, by imposing a Gaussian prior on the coefficient matrix and regularizing on the Gaussian precision matrix, we can capture the task relationship structure. The proposed method can learn the potential basis tasks and the relationship between the tasks and may better capture the nature of the tasks.

II. METHOD

In this section, we will present our formulation for the method and an algorithm to update the parameters.

A. Formulation

Consider a linear regression problem of totally T observed tasks. For each task t, we have an input predictor matrix $\boldsymbol{X}_t \in R^{n_t \times D}$ where n_t is the number of data points for task t and D the dimension of features; and an output predictant vector $\boldsymbol{y}_t \in R^{n_t \times 1}$. Let $\boldsymbol{w}_t \in R^{D \times 1}$ be the regression weight vector for task t, and $\boldsymbol{\varepsilon}_t \in R^{n_t \times 1}$ a vector of i.i.d multivariate Gaussian noise. Then we have

$$\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{w}_t + \boldsymbol{\varepsilon}_t, \ t = 1, \cdots, T \tag{1}$$

Let $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_T]$ be the $D \times T$ weight matrix whose columns $\boldsymbol{w}_t$ are the regression weight vectors of the tasks. Following [12], we assume that there are K latent basis tasks ($K \ll T$) and each observed tasks is a linear combination of the latent basis tasks. Specifically, $\boldsymbol{W}$ can be decomposed as

$$\boldsymbol{W} = \boldsymbol{L}\boldsymbol{S} \tag{2}$$

where $\boldsymbol{L} \in R^{D \times K}$ is a latent basis task weight matrix whose columns are weight vectors for each latent basis, and $\boldsymbol{S} \in R^{K \times T}$ is a coefficient matrix whose columns are coefficients of the linear combination of latent basis for each task. We assume that $\boldsymbol{L}$ has low Frobenius norm such that each column has low $L2$ norm to avoid overfitting. The relationship between different tasks is determined by $\boldsymbol{S}$. $\boldsymbol{S}$ may be sparse such that each observed task consists of a subset of the latent basis tasks instead of all the basis tasks. We further assume that each row $\hat{\boldsymbol{s}}_k$ of $\boldsymbol{S}$ is generated from a multivariate Gaussian distribution with mean $\boldsymbol{0}$ and a sparse precision matrix $\boldsymbol{\Omega} \in R^{T \times T}$. If $\boldsymbol{\Omega}_{mn} = 0$, then the m-th and n-th features in the each $\hat{\boldsymbol{s}}_k$ are conditionally independent given all other features, which means that the m-th and n-th columns of $\boldsymbol{S}$ are conditionally independent.

Given the assumptions above, by Bayes' Theorem, the posterior probability of the weights $\boldsymbol{W}$ is proportional to

$$\begin{aligned} P(\boldsymbol{W} \mid (\boldsymbol{X}, \boldsymbol{Y}), \boldsymbol{\Omega}) &= \frac{P((\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{W}) P(\boldsymbol{W} | \boldsymbol{\Omega})}{P(\boldsymbol{X}, \boldsymbol{Y})} \\ &\propto P((\boldsymbol{X}, \boldsymbol{Y}) | \boldsymbol{W}) P(\boldsymbol{W} | \boldsymbol{\Omega}) \\ &= \prod_{t=1}^{T} \prod_{i=1}^{n_t} p(y_t^i \mid \boldsymbol{x}_t^i, \boldsymbol{w}_t) \prod_{k=1}^{K} p(\hat{\boldsymbol{s}}_k \mid \boldsymbol{\Omega}) \end{aligned} \tag{3}$$

where $(\boldsymbol{X}, \boldsymbol{Y})$ is the set of predictors and predictants. For the likelihood we assume

$$p(y_t^i \mid \boldsymbol{x}_t^i, \boldsymbol{w}_t) = N(\boldsymbol{w}_t^T \boldsymbol{x}_t^i, 1) \tag{4}$$

and for the prior we assume

$$p(\hat{\boldsymbol{s}}_k \mid \boldsymbol{\Omega}) = N(\boldsymbol{0}, \boldsymbol{\Omega}^{-1}) \tag{5}$$

Substituting Eq. (2), (4) and (5) into (3) and taking negative logarithm, we get

$$\begin{aligned} -logP &\propto \frac{1}{2} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \left(y_t^i - \boldsymbol{w}_t^T \boldsymbol{x}_t^i \right)^2 - \frac{K}{2} log \mid \boldsymbol{\Omega} \mid + \frac{1}{2} \sum_{k=1}^{K} \hat{\boldsymbol{s}}_k^T \boldsymbol{\Omega} \hat{\boldsymbol{s}}_k \\ &\propto \sum_{t=1}^{T} \parallel \boldsymbol{y}_t - \boldsymbol{X}_t^T \boldsymbol{L} \boldsymbol{s}_t \parallel^2 - K log \mid \boldsymbol{\Omega} \mid + tr(\boldsymbol{S} \boldsymbol{\Omega} \boldsymbol{S}^T) \end{aligned} \tag{6}$$

where $\boldsymbol{s}_t$ is the t-th column of $\boldsymbol{S}$. We add L1 regularizers of $\boldsymbol{S}$ and $\boldsymbol{\Omega}$ to impose sparsity, and Frobenius norm regularizer on $\boldsymbol{L}$ to avoid overfitting. We also multiply a regularization coefficient λ to the trace term to get better flexibility (it can be regarded as a scaling factor for $\boldsymbol{\Omega}$). Finally, to deal data imbalance of tasks we average the squared error over the number of data points for each task. Combining those together we get the objective function:

$$\begin{aligned} \underset{\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{\Omega}}{\text{minimize}} f(\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{\Omega}) = \sum_{t=1}^{T} \frac{1}{n_t} \parallel \boldsymbol{y}_t - \boldsymbol{X}_t^T \boldsymbol{L} \boldsymbol{s}_t \parallel^2 - K log \mid \boldsymbol{\Omega} \mid + \lambda tr(\boldsymbol{S} \boldsymbol{\Omega} \boldsymbol{S}^T) \\ + \mu \parallel \boldsymbol{S} \parallel_1 + \gamma \parallel \boldsymbol{L} \parallel_F^2 + \beta \parallel \boldsymbol{\Omega} \parallel_1 \end{aligned} \tag{7}$$

where K, λ, μ, γ and β are hyperparameters.

B. Learning Algorithm

In order to solve the objective function (Eq. 7), we update the $\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{\Omega}$ in an iterative way by alternatively solving one variables while fixing the other two.
When $\boldsymbol{L}$ and $\boldsymbol{\Omega}$ are fixed, the objective function with respect to $\boldsymbol{S}$ is a quadratic function with a L1 regularizer, i.e., in Eq. 8. We use the Alternating Direction Method of Multipliers (ADMM) algorithm [13] to solve this function and update $\boldsymbol{S}$.

$$f_{\boldsymbol{L}, \boldsymbol{\Omega}}(\boldsymbol{S}) = \sum_{t=1}^{T} \frac{1}{n_t} \parallel \boldsymbol{y}_t - \boldsymbol{X}_t^T \boldsymbol{L} \boldsymbol{s}_t \parallel^2 + \lambda tr(\boldsymbol{S} \boldsymbol{\Omega} \boldsymbol{S}^T) + \mu \parallel \boldsymbol{S} \parallel_1 \tag{8}$$

When $\boldsymbol{S}$ and $\boldsymbol{\Omega}$ are fixed, the objective function with respect to $\boldsymbol{L}$ is just a quadratic function (Eq. 9) and can be solved in a closed form.

$$f_{\boldsymbol{S}, \boldsymbol{\Omega}}(\boldsymbol{L}) = \sum_{t=1}^{T} \frac{1}{n_t} \parallel \boldsymbol{y}_t - \boldsymbol{X}_t^T \boldsymbol{L} \boldsymbol{s}_t \parallel^2 + \gamma \parallel \boldsymbol{L} \parallel_F^2 \tag{9}$$

When $\boldsymbol{S}$ and $\boldsymbol{L}$ are fixed, the objective function with respect to $\boldsymbol{\Omega}$ (Eq.10) is a so-called "*Sparse Inverse Covariance Selection*" problem, which had been studied in [13] and can also be solved using ADMM method.

$$f_{\boldsymbol{S},\boldsymbol{L}}(\boldsymbol{\Omega}) = -Klog \mid \boldsymbol{\Omega} \mid +\lambda tr(\boldsymbol{S}\boldsymbol{\Omega}\boldsymbol{S}^T) + \beta \parallel \boldsymbol{\Omega} \parallel_1 \quad (10)$$

The overall algorithm is shown in Algorithm 1.

Algorithm 1 Iteratively Update Parameters

1: **Input:**
$\boldsymbol{X}_t, \boldsymbol{y}_t$: predictor matrix and predictant vector
K: number of latent basis tasks
$\lambda, \mu, \gamma, \beta$: regularization coefficients
2: **Output:**
$\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{\Omega}$: output parameter matrices
3: **Initialize:**
i) learn $\boldsymbol{w}_t^0$ for each task using only its own data
ii) let $\boldsymbol{W}^0$ be a matrix containing $\boldsymbol{w}_t^0$ as columns
iii) get top-k singular vectors: $\boldsymbol{W}^0 = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$
iv) initialize $\boldsymbol{L}$ with the first k columns of $\boldsymbol{U}$
v) initialize $\boldsymbol{S} = (\boldsymbol{L}^T\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{W}^0$
vi) initialize $\boldsymbol{\Omega}$ to be an identity matrix $\boldsymbol{I}_{T\times T}$
4: **Compute:**
step $i = 1$
repeat:
$\boldsymbol{S}^{(i+1)} = \underset{\boldsymbol{S}}{\operatorname{argmin}} f_{\boldsymbol{L}^{(i)},\boldsymbol{\Omega}^{(i)}}(\boldsymbol{S})$
$\boldsymbol{L}^{(i+1)} = \underset{\boldsymbol{L}}{\operatorname{argmin}} f_{\boldsymbol{S}^{(i+1)},\boldsymbol{\Omega}^{(i)}}(\boldsymbol{L})$
$\boldsymbol{\Omega}^{(i+1)} = \underset{\boldsymbol{\Omega}}{\operatorname{argmin}} f_{\boldsymbol{L}^{(i+1)},\boldsymbol{S}^{(i+1)}}(\boldsymbol{\Omega})$
until convergence or stopping condition is met
5: **return:**
$\boldsymbol{S} = \boldsymbol{S}^{(i+1)}$, $\boldsymbol{L} = \boldsymbol{L}^{(i+1)}$, $\boldsymbol{\Omega} = \boldsymbol{\Omega}^{(i+1)}$
6: **End of algorithm**

III. EVALUATION

We will evaluate our method using synthetic data, real world benchmark data and climate data. The results are compared with three other methods, i.e., STL, MSSL [8] and GOMTL [12]. For STL, we fit a linear regression for each task $\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{w}_t$ separately and get the regression weights $\boldsymbol{w}_t = (\boldsymbol{X}_t^T\boldsymbol{X}_t)^{-1}\boldsymbol{X}_t^T\boldsymbol{y}_t$ for prediction. For MSSL and GOMTL, their methods can be found in the original papers. The code for MSSL was from the author[1], and the GOMTL code was from this implementation[2] since we didn't find out the authors' implementation.

[1]https://bitbucket.org/andreric/mssl-code/src/default/
[2]https://github.com/wOOL/GO_MTL

A. Synthetic data and benchmark data

In this subsection we present the experiments on four data sets, i.e., two synthetic data sets and two real world benchmark data sets. The first synthetic data set (denoted as *"3 non-overlapped"*) contains 3 non-overlapped groups of tasks, with 10 observed tasks in each group generated from 3 latent basis tasks. Non-overlapping means that tasks in different groups are generated from different basis tasks. Each observed task has 15 training data points and 50 testing data points, with feature dimension being 20. Within each group, the coefficients of the 10 tasks are identical to each other up to a scale. We use the same method with [12] to generate the second synthetic data set (denoted as *"4 overlapped"*). Overlapping means that tasks in different groups have some common basis tasks. We generate 30 observed tasks using 4 latent basis tasks of 20 dimensions. Tasks 1 to 10 are linear combinations of the first two latent basis tasks; tasks 11 to 20 are linear combinations of the second and third latent basis tasks; and tasks 21 to 30 are linear combinations of the third and fourth latent basis tasks. The combination weights for different tasks are different. We set K to be the true values in our method and GOMTL.

The two real world benchmark data sets are the school data and the computer survey data, both have been widely used to evaluate MTL performance in regression [14], [15]. The computer survey data was collected by asking 190 persons about their rating of 20 different kinds of computers regarding 13 characters. The data set consists of 190 tasks with 20 data points per task. The feature of the predictor is 14 dimensional (including an extra dimension of constant 1 to account for the bias term). The school data set consists of the exam scores of 15362 students from 139 schools in London. Thus there are totally 139 tasks with 15362 data points. Each predictor has 27 dimension features (including a feature of constant 1).
We use 5-fold cross validation to select hyperparameters for each method, and then compare the root mean square error (RMSE) of the four methods using the selected optimal hyperparameters. We compare the average RMSE over all the tasks. The results are shown in Table I. From the table, we can see that our proposed algorithm win 3 out of the 4 cases.

The learnt structure of the coefficient matrix $\boldsymbol{S}$ of synthetic data "3 non-overlapped" is shown in the Fig. 1. We can see that the sparse pattern is correctly recovered.

We also compared the performance of our MTL method with STL method in terms of RMSE when

TABLE I: RMSE of synthetic and benchmark data sets. Bold numbers represent best performance. For STL, the results with "$*$" are results with a small identity matrix added to the coefficient matrix to avoid singular issue.

RMSE	3 non-overlapped	4 overlapped	School	Computer
STL	0.98*	2.84*	11.34*	2.24
MSSL	0.60	0.58	11.40	2.17
GOMTL	0.36	1.60	11.72	**1.73**
Proposed	**0.35**	**0.43**	**10.58**	1.78

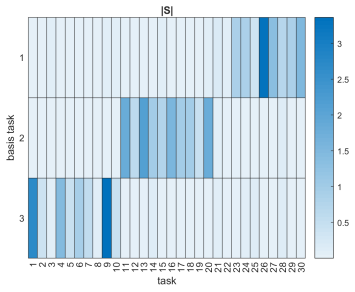


Fig. 1: Absolute values of learnt S matrix structure for 3 non-overlapped data set.

varying the number of training data points. The results are shown in Fig. 2. From the figures we can see that, given enough training data points, both our method and STL can recover the true pattern and have small RMSE. However, our method is much more consistent and performs better than STL, especially when there are very few training data points. For example, if the number of training data points is less than the dimension of the features, then the results of STL are very bad and highly unstable while the results of our MTL are much better. This is due to the fact that the gram matrix $\boldsymbol{X}_t^T \boldsymbol{X}_t$ maybe singular and not invertible. This indicates that our method is better than STL for some climate related problems where there are very few data points. Even with large number of training data points, our method still performs better than STL.

B. *Climate data sets*

We evaluate the proposed method on three climate data sets, of which the first two are from [8] [3], each contains monthly mean surface temperature of South America and North America, respectively. And each with 10 General Circulation Models (GCMs) as predictors ($\boldsymbol{X}$) and an observation as predictant ($\boldsymbol{y}$). The observational data was obtained from the Climate Research Unit (CRU) at the University of East Anglia.

[3] https://www-users.cs.umn.edu/~agoncalv/softwares.html, last accessed July 2019

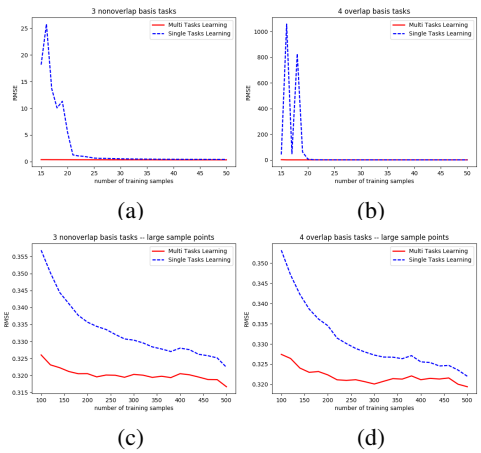


Fig. 2: The impact of the number of training samples. (a)-(d) is non-overlapped/overlapped with small/large number of data points, respectively.

TABLE II: Results for climate data sets. The first number is the RMSE, and the number in the parenthesis is the standard deviation. Bold numbers represent best performance.

RMSE(std)	south_american	north_america	USA
STL	0.9244(0.0067)	2.4920(0.0286)	2.5717
MSSL	0.9264(0.0076)	2.4976(0.0269)	2.4504
GOMTL	0.9263(0.0078)	2.4956(0.0269)	2.2343
Proposed	**0.9133(0.0037)**	**2.4664(0.0178)**	**2.1633**

GCMs are coupled climate models try to simulate the dynamics of the earth system using physical and mathematical equations. They output simulated values for climate variables for a long time period. The spatial grid over longitudes and latitudes is $2.5^\circ \times 2.5^\circ$. We use a time period of 60 years (720 months) from 1901 to 1960. There are 250 and 490 spatial locations/tasks each with 720 data points in the South American and North America, respectively. The regression for each location is regarded as an individual task. We set $K = 10$ and select other hyperparameters using grid search.

For these two data sets, we divide the data set into 5 folds. In each time, we train on the 4 folds and test on the remaining fold and get RMSEs for each task (location) and then take average. Finally, we average over these 5 fold results to get the mean and standard deviation of RMSE. All the comparing methods go through the same procedure. The results are shown in Table II. From the table we can see that both the mean and the range the RMSE of our method are smaller than those of other methods, which indicates our method is better than other methods in terms of both accuracy and variability.

The third climate data set is monthly mean precipitation for the continental United States (USA). The data set contains 18 GCMs of Coupled Model Intercomparison Project 5 (CMIP5) from National Aeronautics and Space Administration (NASA) database [16] [4] as predictor $\boldsymbol{X}$ and the reanalysis data set by the University of Delaware from the National Oceanic and Atmospheric Administration (NOAA) [17] as predictant $\boldsymbol{y}$. There are totally 256 tasks (spatial grid $2^{\circ} \times 2^{\circ}$) with 5 years (60 months) from 2001 to 2004. The first 48 months are as the training data and the other 12 months as the test data. We set $K = 19$ for GOMTL and proposed method and use 5-fold cross validation on the training data to select the other optimal hyperparameters and then retrain the model using the whole training data with the optimal hyperparameters and evaluate on the test data. The RMSE results are shown in the last column of Table II. The learnt coefficients for our method are shown in the Fig.3, which exhibits a sparse pattern.

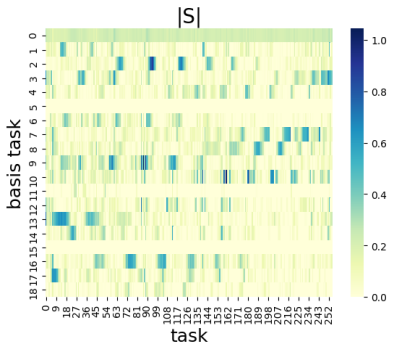


Fig. 3: Absolute values of learnt S matrix structure for the continental USA data set.

IV. CONCLUSION

We proposed a multi-task learning method that incorporates the benefit of learning the latent basis tasks pattern and the task relationship structure. We compared our method with three other methods on two synthetic, two real world benchmark data sets and three climate data sets. The results show that our method outperforms the competing methods. We also show the advantage of multi-task learning over single task learning. This may be very useful in many climate applications where each location can be regarded as a task.

ACKNOWLEDGMENTS

This work was supported by the National Science Foundation CyberSEES project under grant number: NSF CCF-1442728.

[4] https://nex.nasa.gov/nex/resources/348/, last accessed May 2019

REFERENCES

[1] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.
[2] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," in *Advances in neural information processing systems*, pp. 41–48, 2007.
[3] X. Wang, J. Bi, S. Yu, and J. Sun, "On multiplicative multitask feature learning," in *Advances in Neural Information Processing Systems*, pp. 2411–2419, 2014.
[4] L. Jacob, J.-p. Vert, and F. R. Bach, "Clustered multi-task learning: A convex formulation," in *Advances in neural information processing systems*, pp. 745–752, 2009.
[5] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 543–550, Omnipress, 2010.
[6] J. Chen, J. Zhou, and J. Ye, "Integrating low-rank and group-sparse structures for robust multi-task learning," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 42–50, ACM, 2011.
[7] M. T. Bahadori, Q. R. Yu, and Y. Liu, "Fast multivariate spatio-temporal analysis via low rank tensor learning," in *Advances in neural information processing systems*, pp. 3491–3499, 2014.
[8] A. R. Gonçalves, P. Das, S. Chatterjee, V. Sivakumar, F. J. Von Zuben, and A. Banerjee, "Multi-task sparse structure learning," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pp. 451–460, ACM, 2014.
[9] K. Lounici, M. Pontil, A. B. Tsybakov, and S. Van De Geer, "Taking advantage of sparsity in multi-task learning," *arXiv preprint arXiv:0903.1468*, 2009.
[10] E. V. Bonilla, K. M. Chai, and C. Williams, "Multi-task gaussian process prediction," in *Advances in neural information processing systems*, pp. 153–160, 2008.
[11] Y. Zhang and D.-Y. Yeung, "A convex formulation for learning task relationships in multi-task learning," *arXiv preprint arXiv:1203.3536*, 2012.
[12] A. Kumar and H. Daume III, "Learning task grouping and overlap in multi-task learning," *arXiv preprint arXiv:1206.6417*, 2012.
[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.
[14] Z. Kang, K. Grauman, and F. Sha, "Learning with whom to share in multi-task feature learning," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 521–528, Omnipress, 2011.
[15] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.
[16] K. E. Taylor, S. Ronald, and G. Meehl, "An overview of cmip5 and the experiment design," *Bulletin of the American Meteorological Society*, vol. 93, pp. 485–498, 11 2011.
[17] C. J. Willmott, "Terrestrial air temperature and precipitation: Monthly and annual time series (1950-1996)," *WWW url: http://climate. geog. udel. edu/˜ climate/html_pages/README. ghcn_ts. html*, 2000.

USING CAUSAL INFERENCE TO GLOBALLY UNDERSTAND BLACK BOX PREDICTORS BEYOND SALIENCY MAPS

Christian Reimers[1,2], Jakob Runge[2], Joachim Denzler[1,3]

***Abstract*—State-of-the-art machine learning methods, especially deep neural networks, have reached impressive results in many prediction and classification tasks. Rising complexity and automatic feature selection make the resulting learned models hard to interpret and turns them into black boxes. Advances into feature visualization have mitigated this problem but some shortcomings still exist. For example, methods only work locally, meaning they only explain the behavior for single inputs, and they only identify important parts of the input. In this work, we propose a method that is also able to decide whether a feature calculated from the input to an estimator is globally useful. Since the question about explanatory power is a causal one, we frame this approach with causal inference methods.**

I. INTRODUCTION

State-of-the-art machine learning methods, especially deep neural networks, have reached impressive results in many prediction and classification tasks in computer vision and many other fields of science. The benefit of these algorithms for many tasks in Earth system science has been discussed in [1]. One of the main challenges that arises when applying these methods is interpretability. Improved prediction performance comes at the cost of high complexity. This, together with automatic feature selection introduced by deep learning makes the resulting estimators difficult to interpret largely rendering them black boxes.

However, for many applications it is important to identify the features used in a prediction or classification task. This is especially true for applications that are safety and security relevant, for example in medicine or autonomous transportation. It is also true for applications in which predictions cannot be easily verified and, therefore, we need domain experts to determine whether a predicted value makes sense, for example, climate science.

We will discuss some works that aim to make deep learning more explainable in Section II. Most of those methods have at least one of the following drawbacks. First, they only give a local explanation of the estimator, meaning an explanation that is true for a specific input and inputs very close to it. In contrast to that, we aim to produce explanations that are not only true for single inputs but global explanations that are able to explain most inputs.

Second, they can only assign saliency to parts of the input. They cannot be used to identify whether features, for example the variance of the input, that are aggregate functions of the input, are used. In most real-world problems, not the inputs directly are important, but aggregate functions. If our input is, for example, a grid of sea surface temperatures in many different positions, we do not expect a single value to be important towards a prediction task but an aggregate function such as the mean or the variance of the whole grid or the existence of a certain pattern involving multiple values.

Third, they do not handle confounding. They can identify features of the input that are correlated to the output of the classifier but they do not check for a causal link between the input feature and the output. Imagine a classifier that distinguishes two classes, one consisting of red circles and one consisting of green squares. If the classifier only recognizes the shape of the object, there will still be a high correlation between the color of the object and the output of the classifier. Our method respects the label of the input as a confounder.

To understand an estimator, we need to be able to reason about it. These are questions of causality studied in the field of causal inference [2] that has many applications in Earth system science [3]. Methods from this field allow us to understand which features are causing estimations on a global scale. These features do

Corresponding author: C. Reimers, christian.reimers@uni-jena.de
[1]Computer Vision Group, Friedrich Schiller University Jena, Jena
[2]Climate Informatics Group, Institute for Data Science, German Aerospace Center, Jena [3] Michael Stifel Center Jena for data-driven and simulation science, Jena

not have to be part of the input. Our method does not require any information on the estimator but treats it as a black box estimator. Hence, it can also be used in a black box setting if the estimator is non-differentiable or if the inner workings of the estimator are completely unknown.

After we introduce some existing solutions for the problem of identifying relevant features for an automated estimator, we introduce the basic concepts and notations from both machine learning and causal inference that we are using throughout this work. In the fourth section, we describe how a machine learning approach can be phrased as an structural causal model and explain how we can use this to identify features that are causing the estimation of the function resulting from the machine learning approach. In the fifth section, we demonstrate our approach in a toy example. Finally, we discuss open questions and problems of our approach in the last section.

II. Related Work

In this section we introduce existing methods that were developed to explain which features are used by automatic estimators.

In feature visualization, the goal is to create an input image that evokes a maximal response from a specific neuron. If we select a neuron from the output layer, we can find the input that is maximizes the output. This input can be understood as a prototype. Feature visualization was demonstrated in [4], [5] . Feature visualization, however, only visualizes the maximum of the function represented by the estimator. Since the function can be non-concave and multiple maxima might exist, feature visualization is a local explanation around this maximum, while the method presented in this work is a global method explaining all estimations. Further, to apply feature visualization, information on the gradient of the estimator is needed. In contrast, our method can be applied in the black box setting where only input-output pairs are known for the estimator.

The goal of saliency maps is to highlight parts of the input that have a high influence on the output. There are multiple ways to derive the saliency of an input. The first and most straightforward approach is to use the gradient of the loss function depending on every input value or some approximation of this gradient. See [6], [7], [8] for examples of this approach. A slightly more advanced approach is to Taylor-approximate the learned function. A first order Taylor-approximation is demonstrated in [9], [10]. Higher order Taylor-approximations are used in [11], [9]. A third way to generate saliency that can also be used for non-differentiable estimators is substituting parts of the input with a neutral alternative and record the change in the output as demonstrated for example in [12]. The problem of this approach is to find a neutral substitute, since many inputs, for example a black box in an image might already indicate a certain class in a classification problem.

Since they only highlight the important parts in one input, saliency maps are a local explanation method. The method presented in this work is a global explanation method. Furthermore, the method presented in this work can not only identify important parts of the input but also important features that are a aggregate function of these input values. Many algorithms that calculate saliency maps need information on the inner workings of the estimator, such as the gradient. The method presented in this work can be applied to black box estimators.

III. Basics and Notation

A. Machine Learning

Machine learning algorithms are algorithms that can preform tasks without being explicitly programmed but instead are presented with examples and learn from these examples. Machine learning algorithms are mostly used to train functions that preform either regression or classification tasks. For this work we focus on neural networks, even though the method presented here does not make any assumption on the estimator and can be applied to any estimator. Deep neural networks can perform regression and classification. Classification is performed by regressing the probability for every class and then using the maximum likelihood classifier, classifying the input as the most likely class. Therefore, we focus on the regression task as it is implicitly also covering the classification task.

We define a machine learning approach as a pair (T, F) of two functions. The training function T maps a set of labeled training examples $\{(S, Y_S)\}$ onto a set of weights W

$$\begin{aligned} T : \mathcal{P}(\mathbb{S} \times \mathbb{R}) &\to R^m \\ \{(S, Y_S)\} &\mapsto W \end{aligned} \tag{1}$$

and the inference function F maps an input example D and a set of weights W onto a prediction P

$$\begin{aligned} F : \mathbb{S} \times \mathbb{R}^m &\to \mathbb{R} \\ (D, W) &\mapsto P. \end{aligned} \tag{2}$$

If, for example, the machine learning method is a linear estimator, the function T is the optimization process

that maps the training examples on the optimal coefficients and the function F multiplies the coefficients with the inputs D. If the machine learning method is a k-nearest-neighbor approach, the function T is the identity, such that the set of weights is also the labeled training set and the function F is the function that identifies the k-nearest-neighbors of D in W and combines their labels into a single prediction for D.

B. Causal Inference

In this work we only use one task of causal inference. We want to determine whether one variable is causing another variable, given all other causal relation between variables are known. To achieve this, we bring the machine learning approach into the context of a structural causal model.

A structural causal model (SCM) [2] is a triplet (U, V, E) of exogenous variables U, endogenous variables V and a set of functions E. The endogenous variables often represent the observed variables, the exogenous variables represent noise in the system and the functions in E represent the causal mechanisms.

From the functions in E, a directed graph is derived. The vertices in the graph are the endogenous variables V and the graph contains a directed edge from $v_1 \in V$ to $v_2 \in V$ if v_1 is used as the input to the function defining v_2. To be able to evaluate the functions, the graph needs to be a directed acyclic graph (DAG).

Our goal is to identify whether a certain link in the DAG exists. Even if all other links in the DAG are known, we need to employ two assumptions to be able to determine whether a specific link exists. The conditional Markov assumption states that two endogenous variables X, Y are independent given a subset $G \subset V$ if G d-separates $\{X\}$ and $\{Y\}$ in the DAG. The faithfulness assumption states that two endogenous variables are only independent given a subset $G \subset V$ if G d-separates $\{X\}$ and $\{Y\}$ in the DAG. Both of these assumptions are common in the causal inference literature. Together they provide a one-to-one connection between independence and d-separation in the DAG. Even though common, these assumptions are violated in many real applications. We discuss these problems in Section VI.

IV. Method

The goal of this work is to identify whether a specific feature X is used by a black box machine learning method (T, F) for its prediction P.

We proceed by, first, showing that the machine learning approach can be modeled by an SCM and construct a possible DAG for the machine learning approach. The resulting DAG is displayed in Figure 1.

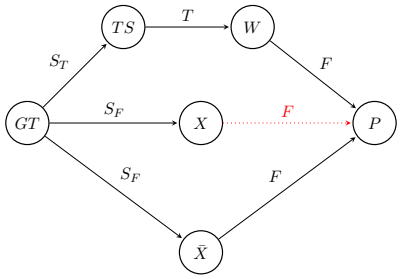


Fig. 1: The DAG of the machine learning approach. In case the feature X is used by the black box machine learning method (T, F) to generate the prediction P, the red link exists.

The endogenous variables we look at are the ground truth GT for the prediction, the labeled training set TS, the set of weights W of the machine learning approach, the feature X for which we want to identify whether it is causing the prediction P, the set $\bar{X}$ of all features independent of X given GT, and the prediction P.

We identify the edges of the DAG from the processes used in the machine learning approach, namely the data generation and sampling processes used to create the training set TS called S_T and the features X and $\bar{X}$ called S_F and the functions F and T described in (2) and (1).

These processes are represented as edges in the DAG. The process S_T is represented as an edge from GT to TS. The process S_F is represented as an edge from GT to X and as an edge from GT to $\bar{X}$. An edge from TS to W represents T and two edges from $\bar{X}$ to P and from W to P represent F.

The only remaining question is, therefore, whether an edge between X and P exists. We know that the prediction cannot cause the input feature. If a causal link between the feature X and the prediction P exists, it is directed from X to P.

Since we made the causal Markov assumption and the faithfulness assumption, we know that conditional independence corresponds to d-separation in the DAG. In the graph we see that $\{GT\}$ d-separates $\{X\}$ and $\{P\}$ in the case where X does not cause P. Further, $\{GT\}$ does not d-separate $\{X\}$ and $\{P\}$ in the case where X does cause P. Hence, to identify whether the feature X is important for the prediction P of the black box F, we check whether

$$X \perp\!\!\!\perp P | GT \quad \text{or} \quad X \not\perp\!\!\!\perp P | GT \tag{3}$$

holds.

TABLE I: Results of the experiment for all three estimators used. The p-values that indicate independence at a confidence level of 0.01 are marked in bold.

Name	MSE	p-value $\bar{A}$	p-value S
F_1	0.0026	0.5137	**0.0014**
F_2	0.0153	0.4576	**0.0091**
F_3	0.0840	**0.0052**	0.3710
F_4	0	0.05912	0.0728

V. EXPERIMENTS

A. Experimental Setup

To demonstrate that the method is able to identify features that are used by the black box predictor for its prediction, we use the following toy data set.

Let α and β be independent latent variables that influence a field $A = (a_{i,j})_{i,j\in\{1,\dots,8\}}$ of observables through

$$a_{i,j} = \alpha \cdot \varepsilon_{i,j} + \beta. \tag{4}$$

Here, $\varepsilon_{i,j} \sim \mathcal{N}(0,1)$ are independent standard normally distributed and independent of α and β. The task is to recover α from A. Note that this task is easy if we know the data creation mechanism but might be non-trivial if we are just presented with the data A and the label α.

For the feature X we test in this toy example two different features. The first is the sample mean

$$\bar{A} = \frac{1}{N}\sum_{i,j} a_{i,j} \tag{5}$$

and the second is the sample standard deviation

$$\bar{S} = \left(\frac{1}{N}\sum_{i,j}(a_{i,j} - \bar{A})^2\right)^{\frac{1}{2}}. \tag{6}$$

We compare four estimators. The first estimator F_1 is the sample standard deviation estimator

$$F_1(A) = \bar{S}. \tag{7}$$

The second estimator F_2 is a fully convolutional neural network. The network consists of three convolutional layers with 4, 16 and 64 kernels of size 2×2 and stride 2×2 followed by one convolutional layer with one kernel of size 1×1. All but the last layer have ReLU activations. We trained the neural network using Tensorflow [13] with gradient descent for 200000 steps using a learning rate of 0.00003. The third estimator F_3 is a linear estimator. To optimize F_2 and F_3 we used a training set of 20000 labeled examples. The fourth estimator F_4 is the oracle estimator that just reports the ground truth data α.

For our experiment we used an 8×8 array for A. The variables α and β are sampled from a uniform distribution. We evaluated every estimator on an identical set of 10000 examples not used for optimizing F_2 and F_3. The results can be observed in Table I. We report for every estimator the mean squared error between the true value of α and the estimated value, the confidence value "p-value $\bar{S}$" for the event

$$F(A) \perp\!\!\!\perp \bar{S}|\alpha \tag{8}$$

and the confidence value "p-value $\bar{A}$" for the event

$$F(A) \perp\!\!\!\perp \bar{A}|\alpha. \tag{9}$$

To calculate this p-values we used the fast conditional independence test presented in [14].

B. Results

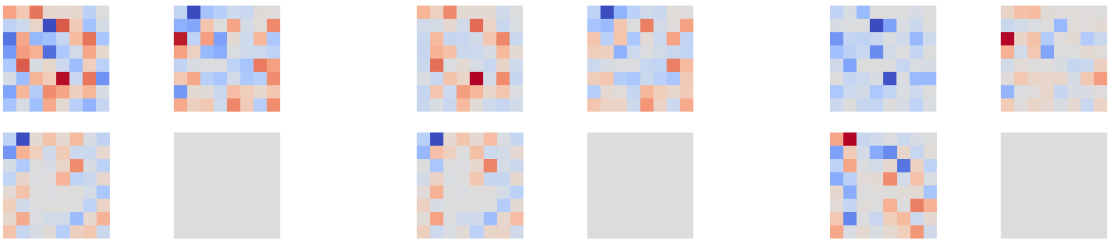
For the optimal estimator F_1 the mean squared error is small. For this estimator, we know that the only used feature is the sample standard deviation. Our method is able to correctly identify that the feature $\bar{S}$ is used by F_1 and the feature $\bar{A}$ is not used by F_1. Further, for the linear estimator F_3 we know that it cannot use the non-linear standard deviation estimate. We find, the linear estimator does not use the sample standard deviation but the sample mean as a feature. It also has the highest mean squared error. For the deep neural network estimator F_2 we observe that the mean squared error is low, and the sample standard deviation is used as a feature. For the oracle F_4 we observe that none of the two features of the input is used for the estimation. This is correct since the oracle reports the correct value independent of the input.

C. Comparison to Related Work

For comparison, in Figure 2 we display the output of the methods described in Section II. We display for all four estimators described in Section V-A the gradient of the output depending on each input (2a), the product of the gradient and the input (2b) and the change of the input if we replace one input by the mean of the other inputs (2c). In our opinion, it is not possible to infer which features are used by the estimator from this methods. The only exception is estimator F_4 which can be identified as independent of the inputs.

VI. DISCUSSION AND OUTLOOK

The toy example showcases two use cases that we imagine for the approach presented in this work. The first use case is to understand fail cases of estimators.



(a) The gradient of the estimator depending on the inputs.

(b) The product of gradient of the estimator and the value of the input.

(c) The difference of the output of the estimator when replacing one of the inputs by the mean of the other inputs.

Fig. 2: In every sub-figure the results of one baseline method is displayed. In the top line of every sub-figure shows the results for estimators F_1 and F_2, in the second line the results for estimators F_3 and F_4. Since the oracle classifier F_4 does not depend on the inputs, the results for F_4 is always zero for all inputs.

In the case of the estimator F_3, our method helps us to understand why it failed and might warn us that it will fail even more for examples with a very different mean. The second use case is to better understand the task at hand. If we have trained an estimator of high quality like F_1 or F_2, we can use the method to identify features that are relevant to solve the task. The experiment on the estimator F_4 demonstrates that conditioning on the ground truth is important and leads to more information than simply checking for independence. The method was able to correctly identify that no feature was used by the estimator F_4, even though the feature $\bar{S}$ is highly dependent with the output of the estimator.

The results on the toy data are very promising. Still, we think that a lot of future work is needed to make this method applicable to a wider range of data and to use it effectively on real-world data.

We assume that the data generation is caused by the ground truth and no further confounding mechanism in the data generation exists. To use this method we need to be able to condition on the data generation process. We assume this drawback can be tackled using the work described in [15], [16], [17] but we leave this for future work. Further, we rely on the causal Markov and the faithfulness assumption. These assumptions can be violated even in simple situations such as an XOR-gate, a trivial function or effects that cancel out. Furthermore, they are very high level assumptions that are hard to validate from other properties of the SCM. Hence, it is difficult to know whether the method of this paper can be used for a given black box predictor. Testing continuous variables for conditional independence, using only samples, is also a hard problem and often additional assumptions have to be made to solve it. Some of the drawbacks of the conditional independence test we used can be found in [14]. When using the approach presented here, one should spend time to decide on an appropriate conditional independence test based on the data used. In this work we only demonstrated a toy example with independent noise. To apply this method to real-world situations in climate science it has to handle dynamic noise and work on sparse data. We leave this for future work. A further drawback of the method is that it has to be presented with candidate features and does not generate features itself. We think, however, that this method can still help in situations in which, due to prior domain knowledge, candidate features already exist or can be generated using other methods.

Throughout this work we use the notation of causality defined by the SCM. One has to be careful when linking this concept to the colloquial concept of ”causes”.

Despite these drawbacks making the method not applicable in some situations, it is still very useful in situations where it can be applied. Its main advantages are that it can provide global explanation for features that are not directly part of the input. Since climate science estimations often depend on features calculated from multiple measurements distributed in space and time and rarely single measurements cause an estimation, these advantages are needed in the field of climate science.

REFERENCES

[1] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, *et al.*, “Deep learning and process un-

derstanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, p. 195, 2019.

[2] J. Pearl *et al.*, "Causal inference in statistics: An overview," *Statistics surveys*, vol. 3, pp. 96–146, 2009.

[3] J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, *et al.*, "Inferring causation from time series in earth system sciences," *Nature communications*, vol. 10, no. 1, p. 2553, 2019.

[4] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," *Distill*, 2017. https://distill.pub/2017/feature-visualization.

[5] D. Erhan, Y. Bengio, A. Courville, and P. Vincent, "Visualizing higher-layer features of a deep network," *University of Montreal*, vol. 1341, no. 3, p. 1, 2009.

[6] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *ICLR 2014 workshop submission*, December 2013.

[7] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*, pp. 818–833, Springer, 2014.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

[9] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek, "On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation," *PloS one*, vol. 10, no. 7, p. e0130140, 2015.

[10] K. R. Mopuri, U. Garg, and R. V. Babu, "Cnn fixations: an unraveling approach to visualize the discriminative image regions," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2116–2125, 2018.

[11] G. Montavon, S. Lapuschkin, A. Binder, W. Samek, and K.-R. Müller, "Explaining nonlinear classification decisions with deep taylor decomposition," *Pattern Recognition*, vol. 65, pp. 211–222, 2017.

[12] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *arXiv preprint arXiv:1702.04595*, 2017.

[13] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015. Software available from tensorflow.org.

[14] K. Chalupka, P. Perona, and F. Eberhardt, "Fast conditional independence test for vector variables with large sample sizes," *arXiv preprint arXiv:1804.02747*, 2018.

[15] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal effect inference with deep latent-variable models," in *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

[16] V. T. Trifunov, M. Shadaydeh, J. Runge, V. Eyring, M. Reichstein, and J. Denzler, "Nonlinear causal link estimation under hidden confounding with an application to time-series anomaly detection," in *German Conference on Pattern Recognition (GCPR)*, 2019.

[17] V. T. Trifunov, M. Shadaydeh, J. Runge, V. Eyring, M. Reichstein, and J. Denzler, "Causal link estimation under hidden confounding in ecological time series," in *Climate Informatics Workshop*, 2019.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

FORECASTING MAXIMA IN CLIMATE TIME SERIES

Israel Goytom[1,2,3] Kris Sankaran[1,3]

***Abstract*—Climate change is already altering the probabilities of weather hazards. Accurate prediction of climate extremes can inform effective preparation against weather-induced stresses. Accurately forecasting extreme weather events is a task that has attracted interest for many years. Classical and to a lesser extent, machine learning-based approaches have handled this issue; however, such systems are hard to tune or scale. While the prediction of extremes has been the subject of investigation across several communities, including meteorology, machine learning, and statistics, it has been subject to far less scrutiny than the prediction of conditional means. In this work, we offer a systematic comparison of existing approaches on prediction of maximum temperature. Further, motivated by this comparison, we propose a method to forecast maxima temperature in weather time series that unifies deep learning with extreme value theory.**

I. MOTIVATION

Weather has an enormous impact in our daily lives. When weather forecasting is effective, we know that burdensome weather events are not coming tomorrow or the day after, either. Extreme weather events such as hurricanes, tornadoes, heavy downpours, heat waves, and droughts affect all sectors of the economy and the environment, impacting people where they live and work (1). According to EM-DAT (International Disaster Database), more than 60 million people were affected only in the year 2018 alone by extreme events. – Forecasting the occurrence of extreme events in time series has attracted interest of researchers for many years (2; 3; 4). Forecasting maxima in weather time series data is essential for extreme weather events, i.e., anticipating high temperature will help people to prepare in advance, forecasting high precipitation might help with flooding events hazards, high-wind speeds with protecting infrastructure. Forecasting maximum surface temperature will help to foretell extreme rainfall, which is the main factor generating floods, landslides, and soil erosion and thus can cause environmental, societal, and economical damages (5).

Forecasting these sources of stress hinges on being able to forecast extremes accurately, and while this problem has been viewed from several angles in the machine learning community, including quantile regression and extreme value forecasting, there have been no systematic comparisons. This work provides a common evaluation of three alternative approaches – direct prediction using an LSTM, a probabilistic LSTM with a likelihood common in extreme value theory, and a quantile regression technique on daily temperature forecasting problem.

Predicting extreme events such as peak wind (6), traffic, (7) and electricity demand (8) has become a common task in both statistics and machine learning community. In statistics, there is a branch known as extreme value theory (9).

Classical methods for extreme weather events forecasting mostly treat the problem as a full-time series prediction problem (9; 10). Alternatively, methods have been developed to model quantiles specifically, including quantile regression and quantile regression forests (11), though these are rarely applied to extreme values. Classical models require hard tuning for the parameters. Long Short Term Memory (LSTM) (12) based forecasting gained popularity due to its end-to-end modeling, automatic feature extraction abilities, and capacity to learn complex interactions.

A combination of classical time series models and machine learning methods have been used to predicting special events (13; 14). Deep convolutional neural networks based classifiers have been used to detect extreme weather (15). (16) proposed a multichannel spatiotemporal encode-decoder convolutional neural network architecture for semi-supervised bounding box prediction in large climate datasets. Recurrent neural networks (RNNs), especially LSTMs, have been used for precipitation nowcasting (17) – when trained on two-dimensional radar map time series, their system

Corresponding author: Israel Goytom, isrugeek@gmail.com
[1]Mila, Montreal, Canada. [2]Ningbo University, Faculty of Science, Ningbo, 352100, China [3]Université de Montréal, Department of Informatics and Operations Research, Montreal, Canada.

is able to outperform the current state-of-art precipitation nowcasting system on various evaluation metrics. Recently, (18) developed an end-to-end forecast model for multi-step time series forecasting that can handle multivariate inputs for extreme events, applying their system to peak travel prediction.

The questions discussed in this paper are:

- To forecast extreme values of the time series, does it help to account for the heavy-tailed distributions expected to arise according to classical statistics theory, or are modern deep learning or quantile regression methods sufficient as they are?
- Alternatively, is there some way to combine the classical theory with modern machine learning in a way that gets the best of both worlds?

Answering these questions will help both the machine learning community, by giving insight into where to invest research effort, and the climate modeling community, as it suggests best practices in a problem of practical importance. We are unaware of any deep learning based methods for climate extreme values or maxima forecasting in weather time series.

The main contributions of this paper are:

- We provide benchmark experiments of modern deep learning, the proposed probabilistic LSTM, and quantile random forest, to evaluate their relative merits on shared tasks.
- We propose an LSTM model with Gumbel-distributed errors, as one way to combine classical theory of extreme values with modern deep learning.

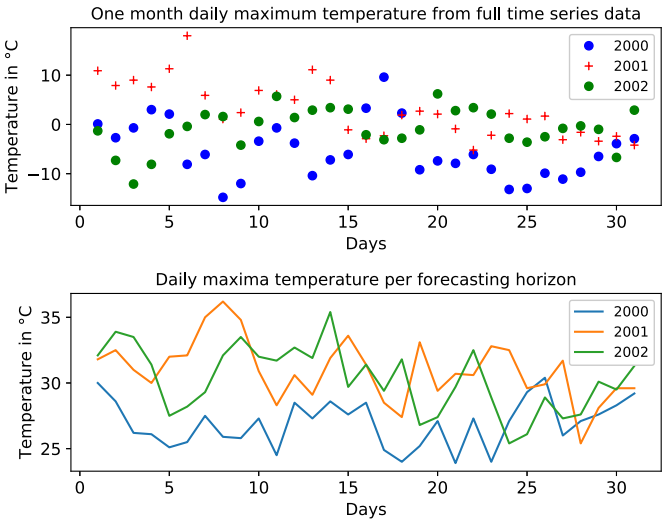


Fig. 1: Extracting maxima values from full time series data. We extract only the maxima from the full time series over the forecasting horizon. This figure is a sample of daily maximum temperature for three years of weather data.

II. METHODS

We consider three models to forecast maxima in weather datasets: LSTM, LSTM with a Gumbel likelihood, and quantile random forest.

A. *Long Short-Term Memory (LSTM)*

Long Short-Term Memory (LSTM) is a type of RNN, capable of learning long-term dependecies that was designed by Hochreiter et al. (12) to address the vanishing and exploding gradient problems of conventional RNNs. The LSTM model discussed in this paper is based on the the original LSTM paper (12) with a hidden layer of LSTM units and an output layer used to make predictions. We provide multivariate data as input and forecast the output maxima. Univariate time-series approaches directly model the temporal domain, they suffer from a frequent retraining requirement (19). Hence, we choose multivariate input, allowing the model to learn from multiple features, not only from the feature being forecasted.

$$
\begin{aligned}
i_t &= \sigma(x_t U^i + h_{t-1} W^i) \\
f_t &= \sigma(x_t U^f + h_{t-1} W^f) \\
o_t &= \sigma(x_t U^p + h_{t-1} W^o) \\
\tilde{C}_t &= \tanh(W x_t U^g + h_{t-1} W^g) \\
C_t &= \sigma(f_t * C_{t-1} + i_t * \tilde{C}_t) \\
h_t &= \tanh(C_t) * o_t.
\end{aligned}
\tag{1}
$$

In Equation 1 i is input gate, f is forget gate and o is output gate. W is the recurrent connection at the previous and current hidden layer while U is the weight matrix connecting the inputs to the current hidden layer. $\tilde{C}$ is a candidate hidden state that is computed based on the current input and the previous hidden state. C is the internal memory of the unit. The output hidden state h_t is computed by multiplying the memory with the output gate as shown in Equation 1.

B. *LSTM + Gumbel-Markov model*

In our next approach we add a Gumbel distribution and Markov stochastic model to the LSTM model. The Markov model is based on (20) and discussed in (21). The cumulative distribution function (CDF) and probability density function (PDF) for the Gumbel

distribution are given in Equation 2 and Equation 3 respectively.

$$F(x;\mu,\beta) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right) \quad (2)$$

$$f(x) = \frac{1}{\beta}\exp\left(\frac{x-\mu}{\beta}\right)\exp\left(-\exp\left(\frac{x-\mu}{\beta}\right)\right) \quad (3)$$

The mode is μ, while the median is $\mu - \beta\ln(\ln 2)$, and the mean is given by : $E(X) = \mu + \gamma\beta$ where $\gamma \approx 0.5772$ is the Euler-Mascheroni constant.

At the mode, where $x = \mu$, the value of $F(x;\mu,\beta)$ becomes $e^{-1} \approx 0.37$ for whatever the value of β.

We use a Markov model to describe the sequence of possible extremes, requiring the distribution of the extreme value to depend only on the state attained at the previous time point,

$$\Pr(X_{n+1} = x \mid X_n = y) = \Pr(X_n = x \mid X_{n-1} = y)$$

For the standard Gumbel distribution where $\mu = 0$ and $\beta = 1$, then CDF states in Equation 2 will be $F(x) = e^{-e^{(-x)}}$ and the PDF states in Equation 3 will be $f(x) = e^{-x}e^{-e^{x}}$.

We optimize the Gumbel likelihood over the features learned by LSTM. The negative log-pdf of a Gumbel distribution (parameterized by μ and β) is

$$-\log p(x_t;\mu,\beta) = \log\beta - \frac{x_t-\mu}{\beta} + \exp\left(-\frac{x_t-\mu}{\beta}\right)$$

Here, we parameterize the mode μ by learned features h_t from the LSTM at the same timepoint. We consider them a linear function of those features, i.e. $\mu(h_t) = w^T h_t$. If we force $\beta = 1$, then this expression becomes

$$-\log p(x_t|h_t;w) = -x_t + w^T h_t + \exp\left(-\left(x_t - w^T h_t\right)\right)$$

For an LSTM the representations $h_t = f_\theta(x_{t-\Delta},\ldots,x_t)$, parameterized by θ, must be learned, along with the Gumbel parameter w. We approach this using maximum likelihood. Specifically, if $x_i := \left(x_{i(t-\Delta)},\ldots,x_{it}\right)$ is the i^{th} window, we minimize

$$-\sum_{i,t}\log p(x_{it}|h_{it};w)$$
$$=\sum_{i,t} -x_{it} + w^T h_{it} + \exp\left(-\left(x_{it} - w^T h_{it}\right)\right).$$

We take a minibatch of x_i and backpropagate through this loss, updating our estimates for θ and w based on the gradient.

C. Quantile Random Forest

Quantile random forests are a variant of random forests that maintain the empirical distribution of all points at leaves in every component tree, as opposed to taking the mean in every leaf, as in standard random forests. This allows the model to provide estimates of arbitrary quantiles at any input x. This is in contrast with standard quantile regression methods – including those based on deep learning – which learn to target specific quantiles by optimizing an asymmetric absolute error loss.

Specifically, to estimate the α-quantile at a position x, the method proceeds as follows. First, grow a collection of trees according to the split criterion in standard random forests. For the t^{th} tree, define the weight, $w_i(x) = \frac{1}{|L_t(x_i)|}$ if x is in the same leaf as x_i in the t^{th} tree, and 0 otherwise, where $|L_t(x_i)|$ is the number of observations in that leaf of the t^{th} tree. That is, observations x_i far from x shouldn't get any weight, and large leaves should be downweighted. Finally, average the weights across trees into a single $w_i(x)$, and use them to estimate the distribution function, $\hat{F}(y|x) = \sum w_i(x)\,\mathbf{1}\{y_i \le y\}$. From this distribution function, any quantile can be extracted.

D. Data

The dataset[1] considered in this work is based on Environment and Climate Change Canada data, the dataset has 148 years of recorded data with 68 features.

Maximum temperature is extracted from the daily temperature during the forecast period Figure 1. We use pushforwards imputation to fill missing values and interpolate values onto evenly spaced timepoints. We prepared the data as maxima for the extreme value which needs to be forecasted and their representative multivariate inputs. An example of a raw dataset is shown in Figure 2 (top). We prepared the training dataset by splitting the raw data into sliding windows (Figure 2, bottom). The input x_i includes the 30 most recent observations, and y_i are the maxima over the next two weeks. We used temporal cross-validation to evaluate in sliding windows.

E. Experiments

The network was trained and tested using NVIDIA Tesla K80 GPU, leveraging with NVIDIA CUDA Toolkit (22). We use the pytorch (23) library for implementation. The code is publicly available[2].

[1] https://montreal.weatherstats.ca/
[2] https://github.com/isrugeek/climate_extreme_values

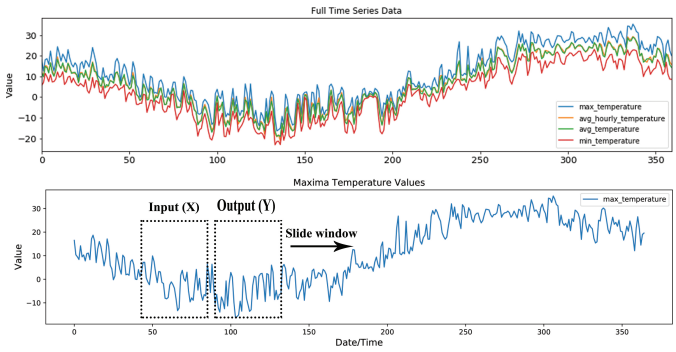
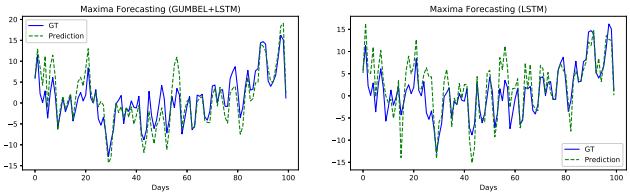


Fig. 2: This figure displays sample maxima from the based on Environment and Climate Change Canada data. Top: Sample input to our model. Bottom: Description of sample creation. We create two sliding windows, one for x_{it} and another for y_{it}.

III. Evaluation

In this section, we present results from the methods we discussed and benchmark them relative to the ground truth. We understood that that LSTM method discussed at the page subsection II-A with more datapoints has improved accuracy, missing points between datapoints will strongly affect the performance of subsection II-A. We notice that interpolation as pre-processing and quantiles in the sample improves the performance. Results from all methods are shown in Table I, and the forecasting results are shown in Figure 3.



(a) (b)

Fig. 3: Ground truth (GT) and prediction comparison of two models on the Canada weather dataset.

We calculate mean absolute error (MAE) see Equation 4 to measure the errors in a set of predictions, when n is number of samples, y actual value (GT) and $\hat{y}$ is the predicted value.

$$MAE = \frac{1}{n}\sum_{t=1}^{n} |y_j - \hat{y}_j| \tag{4}$$

TABLE I: Forecasting error for LSTM, LSTM + Gumbel Markov and quantile random forest model on Canada weather dataset.

METHOD	MAE	RMSE
LSTM	0.3828± 0.0664	0.3075± 0.004
LSTM + GM	**0.3252± 0.0087**	**0.3072± 0.009**
QRF	0.50±0.04	0.411±0.06

IV. Discussion and Future Work

In this paper, we have contrasted existing machine learning and statistical approaches to extreme value modeling, and then we proposed a way to combine the perspectives. We have also evaluated the three models performance in forecasting maximal values on public weather dataset. From our experience (a) LSTM has more trouble on heavy-tailed distributions than the Gumbel-Markov model, (b) directly predicting the maximum of distribution does better than producing a forecast and extracting the maximum from that forecast, (c) a combination of the LSTM + Gumbel-Markov model outperforms LSTM or quantile random forest methods with respect to sample complexity and MAE, and the improvement can be traced to the LSTM's high flexibility and the Gumbel model's ability to deal with heavy tails. In the future, we hope to extend these ideas to the classification of weather extremes, and we will study the effectiveness of our approach on other quantiles and at different time horizons in an extended version of this paper.

References

[1] R. R. Heim, "An overview of weather and climate extremes products and trends," vol. 10, pp. 1–9.

[2] E. J. Kendon, N. M. Roberts, H. J. Fowler, M. J. Roberts, S. C. Chan, and C. A. Senior, "Heavier summer downpours with climate change revealed by weather forecast resolution model," vol. 4, no. 7, pp. 570–576.

[3] A. G. Barnston and S. J. Mason, "Evaluation of iris seasonal climate forecasts for the extreme 15% tails," *Weather and Forecasting*, vol. 26, no. 4, pp. 545–554, 2011.

[4] R. Schnur, "The investment forecast," *Nature*, vol. 415, no. 6871, pp. 483–484, 2002.

[5] G. Panthou, A. Mailhot, E. Laurence, and G. Talbot, "Relationship between surface temperature and extreme rainfalls: A multi-time-scale and

event-based analysis," *Journal of Hydrometeorology*, vol. 15, no. 5, pp. 1999–2011, 2014.

[6] P. Friederichs and T. L. Thorarinsdottir, "Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction," *Environmetrics*, vol. 23, no. 7, pp. 579–594, 2012.

[7] N. Polson and V. Sokolov, "Deep learning for short-term traffic flow prediction," vol. 79, pp. 1–17.

[8] J. Ringwood, D. Bofelli, and F. T. Murray, "Forecasting electricity demand on short, medium and long time scales using neural networks," *Journal of Intelligent and Robotic Systems*, vol. 31, pp. 129–147, 05 2001.

[9] L. d. Haan and A. Ferreira, *Extreme value theory: an introduction*. Springer series in operations research, Springer. OCLC: ocm70173287.

[10] R. W. Katz and B. G. Brown, "Extreme events in a changing climate: Variability is more important than averages," vol. 21, no. 3, pp. 289–302.

[11] N. Meinshausen, "Quantile regression forests," *J. Mach. Learn. Res.*, vol. 7, pp. 983–999, Dec. 2006.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, pp. 1735–1780, Nov. 1997.

[13] R. J. Hyndman and Y. Khandakar, "Automatic time series forecasting: the forecast package for R," *Journal of Statistical Software*, vol. 26, no. 3, pp. 1–22, 2008.

[14] T. Opitz, "Modeling asymptotically independent spatial extremes based on Laplace random fields," *arXiv e-prints*, p. arXiv:1507.02537, Jul 2015.

[15] Y. Liu, E. Racah, Prabhat, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, and W. Collins, "Application of Deep Convolutional Neural Networks for Detecting Extreme Weather in Climate Datasets," *arXiv e-prints*, p. arXiv:1605.01156, May 2016.

[16] E. Racah, C. Beckham, T. Maharaj, Prabhat, and C. J. Pal, "Semi-supervised detection of extreme weather events in large climate datasets," *CoRR*, vol. abs/1612.02095, 2017.

[17] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W. chun Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *NIPS*, 2015.

[18] N. Laptev, J. Yosinski, L. E. Li, and S. Smyl, "Time-series extreme event forecasting with neural networks at uber," p. 5.

[19] L. Ye and E. Keogh, "Time series shapelets: A new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, (New York, NY, USA), pp. 947–956, ACM, 2009.

[20] R. G. Krishnan, U. Shalit, and D. Sontag, "Structured inference networks for nonlinear state space models," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 2101–2109, AAAI Press, 2017.

[21] E. Bingham, J. P. Chen, M. Jankowiak, F. Obermeyer, N. Pradhan, T. Karaletsos, R. Singh, P. Szerlip, P. Horsfall, and N. D. Goodman, "Pyro: Deep Universal Probabilistic Programming," *arXiv preprint arXiv:1810.09538*, 2018.

[22] J. Nickolls, I. Buck, M. Garland, and K. Skadron, "Scalable parallel programming with cuda," *Queue*, vol. 6, pp. 40–53, Mar. 2008.

[23] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

THE IMPORTANCE OF INDUCTIVE BIAS IN CONVOLUTIONAL MODELS FOR STATISTICAL DOWNSCALING

Jorge Baño-Medina[1], Jose M. Gutiérrez [1]

***Abstract*—Statistical downscaling is routinely used to produce regional climate change projections from coarse global model outputs. Along with the success of deep learning in multiple disciplines, recent studies outline the capability of deep neural models as a statistical downscaling technique. In this work, we analyze this problem from a multi-site perspective and highlight the benefits of a deep learning model. We argue that their merits are due to the existence of an inductive bias in multi-site architectures that prevents overfitting in over-parameterized models with no need for dimensionality reduction techniques. We frame the experiment in the largest to date downscaling intercomparison study, called VALUE. The result is a better local reproducibility of multi-site deep neural models in comparison with single-site neural models and VALUE's benchmark methods.**

I. MOTIVATION

In order to study the potential impacts of climate change, and to elaborate suitable adaptation measures, certain socio-economic sectors (e.g., agriculture, energy, health) require regional climate information for the next decades. The projections provided by the Global Climate Models (GCMs) are too coarse to be directly applied in many sectors. Statistical downscaling [1] bridges the gap between the low and the high-resolutions of GCMs and local observations, respectively, by learning empirical functions between large-scale atmospheric variables and a record of climate observations at a local scale. The perfect prognosis approach builds these relationships using reanalysis data.

Classical techniques including generalized linear models (GLMs) or analogs [2], and machine learning models, such as neural networks [3] or support vector machines [4], are among the most common statistical methods used by the downscaling community. Despite there have been numerous intercomparison studies, to date no method outstand against the others in terms of temporal reproducibility or spatial consistency.

However, in the last decade, the advances in the development of neural networks (e.g., stochastic gradient descent, new learning algorithms [5], ReLu activation function [6] and computing infrastructures) have made architectures with many layers tractable, becoming the state-of-the-art in other disciplines (e.g., image recognition [7]). Thus, despite convolutional architectures have existed since the 90s [8], it was not until recently when they first appeared as statistical downscaling models thanks to the deep learning machinery [9][10]. In particular, in [10] a variety of deep neural models were intercompared in order to shed light on the role of each element in the downscaling process. This was carried out applying the validation framework developed by the European COST-action VALUE [11], which represents the largest to date downscaling intercomparison study.

Due to the difficulties of standard models to downscale continental domains (e.g. Europe) several methods contributing to VALUE considered subdomains and performed dimensionality reduction on the predictors, mainly principal component analysis or nearest neighbour selection. Under this scenario, [10] showed that convolutional models can automatically treat high-dimensional spatial domains without any previous feature selection/extraction process. Furthermore, the non-linear spatial patterns learned by the convolutions resulted in a better reproduction of the local variability than the top-ranked methods in VALUE [12].

In this study we show that the multi-site character of neural models (i.e., downscaling to more than 1 site at a time) favours the simultaneous treatment of high-dimensional domains without leading to overfitting. The main idea behind multi-task learning in neural networks (see [13] for a review) is that the common properties are learned in a shared representation working as an

Corresponding author: J. Baño-Medina, bmedina@ifca.unican.es
[1]Meteorology Group. Instituto de Física de Cantabria, IFCA (CSIC - Univ. de Cantabria), Santander, 39005, Spain

inductive bias to the net and consequently improving generalization [14].

Therefore we extend the work done in [10] and investigate the influence of multi-site arquitectures over the implicit regularization of deep learning models in a statistical downscaling context.

II. DATA

Here we build on the research done in [10] and frame the experiment in the VALUE intercomparison project [12]. Some VALUE methods divided Europe in eight regions for a better characterization of regional climates and a simplification of the input space. Though the eight regions were treated simultaneously as a whole in [10], for simplicity in this study we focus on the domain covering the Iberian peninsula (see Figure 1). According to VALUE, we use the following predictor variables from the ERA-Interim reanalysis [15]: temperature, geopotential, specific humidity and zonal and meridional wind at 250,500 750 and 1000 hPa, resulting into 20 variables per gridpoint. On the other hand we use precipitation of the E-OBS dataset [16] as the response variable (i.e., predictand) on a daily scale. Thus, the objective is to learn empirical functions linking the resolution of ERA-Interim (i.e., 2°) to the E-OBS resolution (0.5° in this example). The cross-validation experiment defined in VALUE indicates the splitting of the data in 5 chronological folds: 1979-1984, 1985-1990, 1991-1996, 1997-2002, 2003-2008. In contrast to [10], where only the 5^{th} fold was used as test set, in this study we reconstruct the 5 folds permitting the analysis of the performance of the methods in the same experimental framework proposed in VALUE.

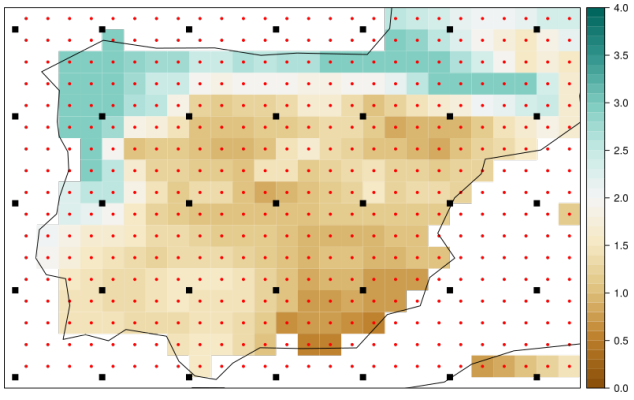


Fig. 1. Climatology of the precipitation in mm/day for the period 1978-2008 over the VALUE Iberian Peninsula subdomain. The black squares represent the predictor's gridpoints (i.e., a total of 35) whereas the red dots indicate the predictand's resolution (i.e., a total of 296 over land).

III. METHODS

In order to evaluate the implicit regularization of neural-based multi-site models in statistical downscaling we start from the intercomparison study done in [10]. According to the latter, an architecture made of 3 convolutional hidden layers with 50,25 and 1 filters, respectively, outperforms the VALUE benchmark methods and other deep learning models in terms of local reproducibility. This success was achieved partly due to the spatial features learned by the convolutions but, in order to study the role played by the inductive bias, we build single-site versions of the architecture described (i.e., everything is identical except from the output layer which reduces to a single gridbox).

Thus, the net uses as input layer 20 feature maps, corresponding to the 20 variables indicated in Section II and fully connect the last convolutional layer with the output layer. Due to the discrete-continuous nature of precipitation, the net minimizes the negative-loglikelihood of a Bernouilli-Gamma distribution as done in [3] and [17]. Therefore, the output layer consists of 3 feature maps (i.e., 3 neurons in single-site mode) matching the parameters p (i.e., probability of rain in a Bernouilli distribution), α (i.e., shape parameter of a Gamma distribution) and β (i.e., scale parameter of a Gamma distribution). The mean daily rainfall for a given day i, can be recovered as the expected value of the conditional Gamma distribution where $\mu_i = \alpha\beta$.

Moreover, we compare both single-site and multi-site models with 2 methods that ranked among the best in a recent contribution to the VALUE experiment (see [18]): a generalized linear model and analogs with a moving window of 4 and 25 neighbours, respectively (e.g., the predictor's gridpoints selected are the 4 closest to the predictand's localization in the case of the GLM).

In terms of reproducibility, we rely on climate4R [19], which is a set of R packages designed to handle climate data and promotes transparent climate data access. Moreover, the downscaling phase can be done with the climate4R library downscaleR [18] for the analogs and the GLM methods. On the other hand, the deep learning models rely on the R version of Keras. The latter facilitates the implementation of the deep learning's new technological developments. To train the models we perform early-stopping as a Keras's callback with a patience of 15 epochs and a learning rate of 5e-4 using the Adam optimizer.

IV. EVALUATION

In Figure 2, we can observe the downscaling done with the CNN-MS (Convolutional Neural Network Multi-Site) for a particular day, using as reference the ERA-Interim's total precipitation. Whereas reanalysis' rainfall pattern lacks from enough resolution, neural-based downscaling is able to capture local structures in the precipitation pattern.

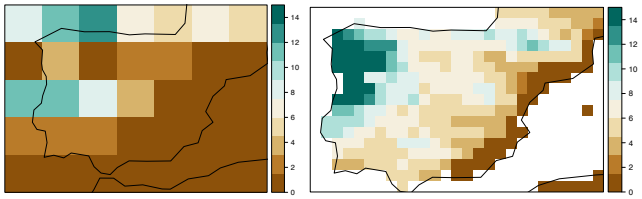


Fig. 2. ERA-Interim's (left) and downscaled (right) total precipitation (mm/day) over Iberia for 01-01-1979.

Multi-site neural models achieve optimum scores in comparison with benchmark methods (i.e., GLM and analogs) and their equivalent single-site networks, in terms of Roc Skill Score (ROCSS), Root Mean Squared Error (RMSE) and spearman correlation (see Table I). On the other hand, the climatological relative bias is better adjusted by the GLM and analogs than by convolutional models, being very close to 0. Though multi-task architectures are very little biased this is especially relevant for single-site models whose relative values overpass 0.5 for most of the folds.

The CNN-MS reproduces slightly better the occurrence of precipitation than the linear model and far better than the single-site model (CNN-SS), according to the ROCSS index. Thus, when downscaling is wanted for single sites, convolutional models suffer from the same high-dimensional issues than classical methods (e.g. GLM). In order to take advantage of the ability of CNNs to extract spatial information and to add nonlinearity to the task, multi-site models are needed, introducing an inductive bias to the model which plays a major role to prevent overfitting. The same result can be observed for the RMSE, where CNN-MS obtains lower values than classical methods and especially than CNN-SS, which again suffers from overfitting. The success in the reproducibility of the occurrence of precipitation, measured by the ROCSS, and in the amount of precipitation, measured by the RMSE, traduces also into better correlation values of multi-site models than the rest of the methods tested.

In order to visualize a more detailed description of the advantages of multi-site models over single-site ones, we plot their differences in the validation indices

ROCSS	**fold 1**	**fold 2**	**fold 3**	**fold 4**	**fold 5**
analogs	0.61	0.58	0.59	0.58	0.57
GLM	0.87	0.86	0.87	0.87	0.86
CNN-SS	0.77	0.75	0.74	0.75	0.78
CNN-MS	**0.90**	**0.88**	**0.89**	**0.89**	**0.89**
RMSE	**fold 1**	**fold 2**	**fold 3**	**fold 4**	**fold 5**
analogs	4.18	4.17	4.1	4.19	4.14
GLM	4.62	4.25	4.03	4.13	4.16
CNN-SS	4.75	4.75	11.84	5.22	5.18
CNN-MS	**3.83**	**3.64**	**3.69**	**3.79**	**3.48**
Sp. Cor	**fold 1**	**fold 2**	**fold 3**	**fold 4**	**fold 5**
analogs	0.58	0.57	0.58	0.57	0.56
GLM	0.72	0.70	0.71	0.71	0.70
CNN-SS	0.72	0.72	0.73	0.73	0.71
CNN-MS	**0.76**	**0.73**	**0.76**	**0.75**	**0.74**
Rel. bias	**fold 1**	**fold 2**	**fold 3**	**fold 4**	**fold 5**
analogs	**-0.04**	**0.05**	-0.11	-0.07	-0.06
GLM	0.11	0.09	**-0.02**	**0.05**	**0.05**
CNN-SS	0.65	0.68	0.17	0.67	0.55
CNN-MS	0.32	0.16	0.15	0.2	0.16

TABLE I
RESULTS FOR THE VALIDATION INDICES. THE BEST RESULTS PER INDEX AND FOLD ARE IN BOLD.

per gridpoint (see Figure 3). Therefore, we observe how the improvement in the ROCSS (Figure 3a) and in the correlation (Figure 3c) of CNN-MS with respect to CNN-SS is a generalized situation, attaining higher values over the majority of the Iberia region. There are very few isolated gridpoints who show the opposite behaviour such as the one located in the Balearic Islands. We hypothesize that this may be due to its very particular local climatology what prevents it from benefiting from multi-site mode. For the amount of precipitation, we observe negative values (Figure 3b), indicating lower errors for the CNN-MS over CNN-SS, especially in southern and northeastern Iberia.

V. CONCLUSIONS

According to the results described in Section IV, we can conclude that:

1) The deep learning machinery is suitable to benefit from high-dimensional large-scale information avoiding the implicit loss of information of dimensionality reduction techniques.
2) If a convolutional model is designed to downscale only at 1 site, then it suffers from the same issues than traditional approaches. Under this scenario, deep learning adds little value to the task.
3) When the interest is to downscale to multiple sites, then multi-task convolutional architectures outperform classical and single-site models in terms of local reproducibility, thanks to the existence of an inductive bias.

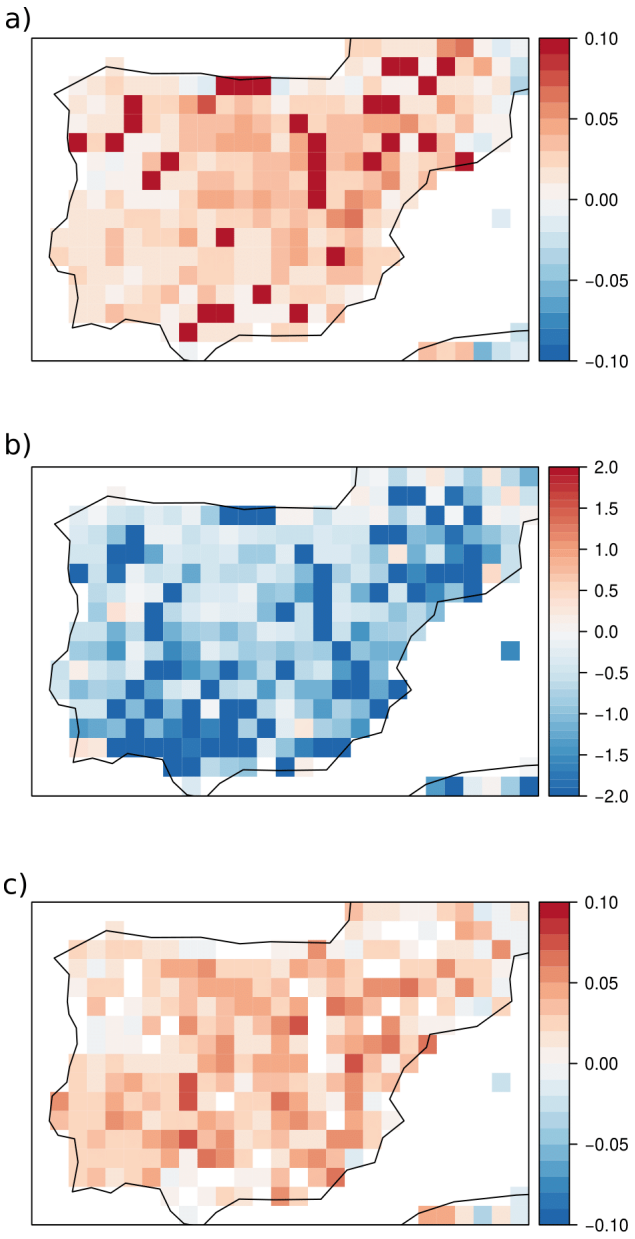


Fig. 3. Differences in the a) ROCSS, b) RMSE and c) Spearman correlation over the Iberian region between the CNN-MS and the CNN-SS models.

Further work will consist in investigating the limits of the predictand's domain or how the addition of completely different climatological sites affects the downscaling. This will be done under the international initiative CORDEX (Coordinated Regional Climate Downscaling Experiment), where one of the objectives is to provide downscaled climate change projections over the entire globe.

ACKNOWLEDGMENTS

We acknowledge the E-OBS dataset from the EU-FP6 project UERRA (http://www.uerra.eu) and the Copernicus Climate Change Service, and the data providers in the ECA&D project (https://www.ecad.eu). Funding for the authors was provided by the project MULTI-SDM (CGL2015-66583-R, MINECO/FEDER).

REFERENCES

[1] D. Maraun and M. Widmann, *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, Jan. 2018. Google-Books-ID: AMhJDwAAQBAJ.

[2] E. Zorita and H. von Storch, "The Analog Method as a Simple Statistical Downscaling Technique: Comparison with More Complicated Methods," *Journal of Climate*, vol. 12, pp. 2474–2489, Aug. 1999.

[3] A. J. Cannon, "Probabilistic Multisite Precipitation Downscaling by an Expanded BernoulliGamma Density Network," *Journal of Hydrometeorology*, vol. 9, pp. 1284–1300, Dec. 2008.

[4] S. Tripathi, V. V. Srinivas, and R. S. Nanjundiah, "Downscaling of precipitation for climate change scenarios: A support vector machine approach," *Journal of Hydrology*, vol. 330, pp. 621–640, Nov. 2006.

[5] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[6] A. F. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," *arXiv:1803.08375 [cs, stat]*, Mar. 2018. arXiv: 1803.08375.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, pp. 84–90, May 2017.

[8] Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.

[9] T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, "DeepSD: Generating High Resolution Climate Change Projections through Single Image Super-Resolution," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*, (Halifax, NS, Canada), pp. 1663–1672, ACM Press, 2017.

[10] J. Baño-Medina, R. Manzanas, and J. M. Gutiérrez, "Configuration and Intercomparison of Deep Learning Neural Models for Statistical Downscaling," *Submitted to Geoscientific Model Development*, 2019.

[11] D. Maraun, M. Widmann, J. M. Gutiérrez, S. Kotlarski, R. E. Chandler, E. Hertig, J. Wibig, R. Huth, and R. A. Wilcke, "VALUE: A framework to validate downscaling approaches for climate change studies," *Earth's Future*, vol. 3, pp. 1–14, Jan. 2015.

[12] J. M. Gutiérrez, D. Maraun, M. Widmann, R. Huth, E. Hertig, R. Benestad, O. Roessler, J. Wibig, R. Wilcke, S. Kotlarski, D. San Martn, S. Herrera, J. Bedia, A. Casanueva, R. Manzanas, M. Iturbide, M. Vrac, M. Dubrovsky, J. Ribalaygua, J. Prtoles, O. Rty, J. Risnen, B. Hingray, D. Raynaud, M. J. Casado, P. Ramos, T. Zerenner, M. Turco, T. Bosshard, P. tpnek, J. Bartholy, R. Pongracz, D. E. Keller, A. M. Fischer, R. M. Cardoso, P. M. M. Soares, B. Czernecki, and C. Pagé, "An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment," *International Journal of Climatology*, vol. 39, pp. 3750–3785, 2019.

[13] S. Ruder, "An Overview of Multi-Task Learning in Deep Neural Networks," *arXiv:1706.05098 [cs, stat]*, June 2017. arXiv: 1706.05098.

[14] R. Caruana, "Multitask Learning," p. 35.

[15] "The ERAInterim reanalysis: configuration and performance of the data assimilation system - Dee - 2011 - Quarterly Journal of the Royal Meteorological Society - Wiley Online Library."

[16] R. C. Cornes, G. van der Schrier, E. J. M. van den Besselaar, and P. D. Jones, "An Ensemble Version of the E-OBS Temperature and Precipitation Data Sets," *Journal of Geophysical Research: Atmospheres*, vol. 123, pp. 9391–9409, Sept. 2018.

[17] P. M. Williams, "Modelling Seasonality and Trends in Daily Rainfall Data," p. 7.

[18] J. Bedia, J. Baño-Medina, M. N. Legasa, M. Iturbide, R. Manzanas, S. Herrera, D. San Martín, A. S. Cofiño, and J. M. Gutiérrez, "Statistical downscaling with climate4R: Contribution to the VALUE intercomparison experiment," *Submitted to Geoscientific Model Development*, 2019.

[19] M. Iturbide, J. Bedia, S. Herrera, J. Baño-Medina, J. Fernndez, M. D. Fras, R. Manzanas, D. San-Martn, E. Cimadevilla, A. S. Cofio, and J. M. Gutiérrez, "The R-based climate4r open framework for reproducible climate data access and post-processing," *Environmental Modelling & Software*, vol. 111, pp. 42–54, Jan. 2019.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

RAINFALL EVENT ANALYSIS IN THE NORTH OF TUNISIA USING THE SELF-ORGANIZING MAP

Derouiche Sabrine[1,2], Mallet Cécile[1], Bargaoui Zoubeida[2]

Abstract— **during the last decade, scientific research has shown a growing interest in quantifying the impact of climate variability on rainfall and water resources. In Tunisia, the variability of rainfall is at the origin of natural disasters extremely expensive in human lives and responsible for innumerable material damages. So, the analysis of rain events characteristics is an essential step for improving our understanding of spatial and temporal variation in precipitation. For this study, 70 raingauge stations in Northern Tunisia are used over 50 years (1959-2008). It is proposed to adopt a seasonal analysis (December-January- February) with a separation of daily rainall time series into events after the determination of the minimum inter event time MIT. This data transformation give us 6 rainfall variables: (1) rainfall event number, (2) total event duration, (3) Average precipitation, (4) total precipitation, (5) Average intensity, (6) Average duration. Those variables are clustered by using the method of Self Organizing Map SOM and give 4 type of seasons.**

I. MOTIVATION

The rainfall pattern in the Mediterranean is characterized by an important spatial and temporal variability. This variability is mainly due to its position (between 30°N and 45° N) which is directly influenced by subtropical high pressures and low mid-latitude pressures [1]. Tunisia, is located between the longitude 6-12 E and the latitude 30-38 N. This is a climatic transition zone between the temperate European domain, north of the Mediterranean and the southern African subtropical domain. Studies of rainfall patterns in Tunisia are often based on annual, seasonal or monthly rainfall averages [2], [3] and [4]. Since precipitation is an intermittent phenomenon that appears in the form of events, the study of the variability of events characteristics, proposed in this study, is well suited to the analysis of precipitation variability. The variability of the events of a single one season (December-January- February) is analysis. The DJF is considered the wet season in Tunisia.

[1] LATMOS/CNRS/UVSQ/Université Paris-Saclay, 11 boulevard d'Alembert, 78280 Guyancourt, France. [2]LMHE/ENIT/University of Tunis El Manar, Farhat HACHED EL MANAR BP 37, LE BELVEDERE 1002 Tunis, Tunisia.

II. DATA AND METHOD

This study is based on a daily rainfall database from 1959 to 2008 collected from the General Direction of Water Resources of the Ministry of Agriculture and Water Resources over 70 rain gauge stations distributed on the northern part of the Tunisian territory (Fig.1).

Fig.1: Rainfall stations distribution

The time series in rain gauge stations is broken down into a separate rain event by a dry period called Minimum Inter event Time MIT. In the literature, there are several statistical method to determinate the MIT such as the autocorrelation analysis and the coefficient of variation analysis [5] and [6]. The method adopted in this study is the autocorrelation analysis. This technic is based on the statistical independence of rain events [7]. Once the MIT is estimated. Two rain events are considered independent when the non-rain or the Inter Event Time IET in the time series is greater than the MIT. If it is smaller than the MIT the rain event is considered as a single event

Descriptive variables of precipitation events calculated for the same season for each station during 50 years are described in Tab.1.

Tab.1: Variables are characterizing rain events (DJF)

Nom	Symbol	Unit	Formula
Event number	EN	Events	
Total duration	TD	days	$TD = \sum_{i}^{NE} TD_i$
Precipitation	P	mm	
Average precipitation	MP	mm/event	$MP = \frac{P}{EN}$
Average duration	MD	days/event	$MD = \frac{TD}{EN}$
Average intensity	MI	mm/day	$MI = \frac{P}{TD}$

In order to classify the data, to analyze the link between those variables and to understand the structure of the rainfall variables the Self Organizing Map (SOM) is used. A SOM is an unsupervised learning algorithm based on artificial neural networks that produce a low-dimensional representation of a high-dimensional input dataset [8] and [9]. SOMs can be used for a variety of operations in exploratory data analysis, such as clustering, data compression, non-linear projection and pattern recognition. In this paper, we run the SOM tool in MATLAB using the SOM algorithm as described by Vesanto et al. [10]. The training of Kohonen map is done starting from the matrix of input data constituted of 3500 observations (50 years * 70 stations) and six variables described in (Tab.1). Many training have been performed for different SOM parameters in order to have a better quality of the map and minimal quantification and topological errors. SOM parameters finally obtained are described in (Tab.2). The learning of the map is very sensitive to neighborhood radius. If the radius is very small there is a risk of losing the data structure and also we can get big topological error. Also, the choice of the neuron number influences the results. If neurone number is very large or close to the number of initial data, there is a risk of overfitting but if it is small the quantification error will be important. It has some empirical techniques to introduce an optimal number of referent vectors [8]. Usually, the SOM is used combined with another classification technique especially the Hierarchical Agglomerative clustering HAC. This second classification is applied to the referents vectors trained by SOM algorithm to reduce the number of cluster, to allow a better understanding of the data and also to extract the information in a relevant way. The principle of HAC is the agglomeration of all data starting with individual elements. The algorithm of HAC is iterative, it merges at each stage the closest classes based on a criterion of similarity or dissimilarity (usually Euclidean distance). There are several agglomeration or linkage strategies used, the most known are single, complete, centroid, average, and Ward strategy [11]. In this study the approach of Ward is adopted .Ward's method minimizes the increase in total within-cluster sum of squared error. This increase is proportional to the squared Euclidean distance between cluster centers [12]

Tab.2: SOM parameters

parameters	
Neuron number	320
Neighborhood function	Gauss
Fine-tuning	
Epoch number T	5000
Initial and Final radius of training	[3 0.5]

III. EVALUATION AND DISCUSSION

As explain previously we have firstly to define the right value of the MIT to avoid merging two independent events. The autocorrelation analysis is done for all stations. The 4 stations used in (Fig.3) are chosen randomly, they are distributed so that they cover the studied area. The correlograms (Fig.3) give the autocorrelation coefficient of daily rainfall data for the season (DJF) over 50 years from 1959 to 2008 with the approximate 95%-confidence intervals for a white noise process. Outside the blue lines, the autocorrelation coefficient are considered significant.

Generally, for a four days lag or more the correlations are no longer significant. The significance of autocorrelation values with a three day lag differs from rain gauge station to another. The coefficient for a lag of 2 days is usually small (less than 0.1) and significant. For a 1 day step, the autocorrelation values vary between 0.2 and 0.45 (a high coefficient). This statistical study showed that the daily rainfall time series are decorrelated after two days. The six variables (Tab.1) corresponding to each season are thus computed in using a MIT equal to two days to define the events.

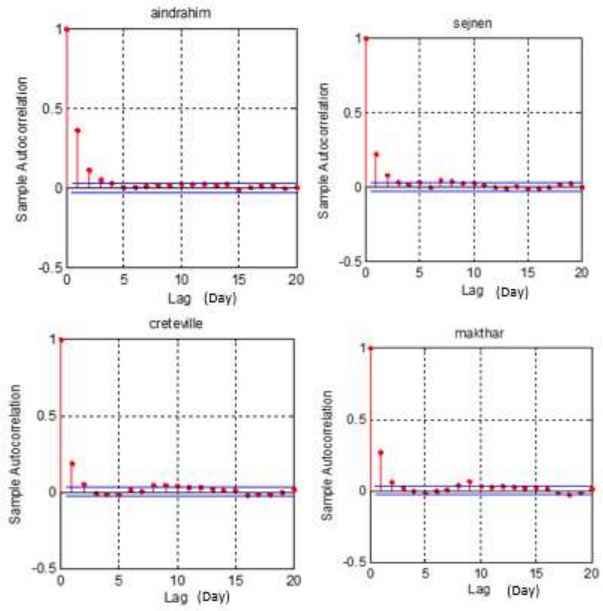


Fig.3: Autocorrelation Analysis

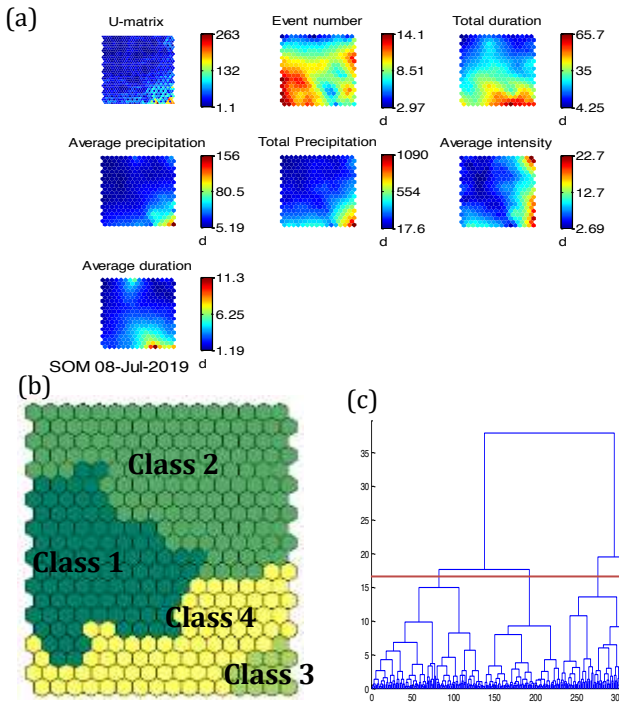


Fig.4: (a) Variables projection in topological Map. (b) Classes delimitation in topological map. (c) Dendrogram

The visualization in space of the topological map (Fig. 4a) of the values of the rainfall variables, corresponding to each neurone, makes it possible to study their relations. The U matrix in (Fig.4a) is a very important visualization for the interpretation of structures of the data. It presents the distance (the similarity) between neurons pairwise. It allowed us to distinguish that in the lower right of the map there is a group of observation less dense, which are very dissimilar compared to other data and which represent extreme situations. The outcome of the HAC is called a dendrogram (Fig.4c). This visualization shows the cluster and sub-cluster relationships and the order in which the clusters were consolidated. The closeness of the clusters can be depicted by lengths of the limbs, and the data items can be clustered by cutting the dendrogram (Fig.4c). So, we can group the corresponding seasons into four clusters and the difference between clusters is visualized in the distribution of (Fig.5). Using the delimitation of clusters in topological map (Fig.4b) and the observation the of the variables projection (Fig.4a) the four clusters are characterized.

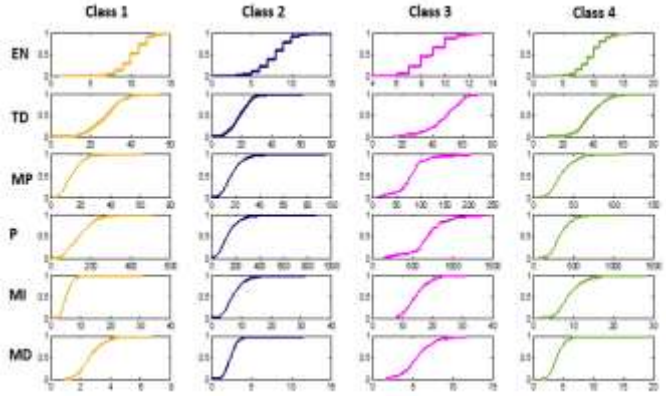


Fig.5: Empirical distribution of classes for the 6 rainfall variables

Class 1 is characterized by a low amount of precipitation (the total of the season, the average of events, or the average per day are weak). However, the event number and the total duration is strong. The class 1 represents the dry seasons with intermittent low rainfall. Class 2 is characterized by a weak values of all variables compared to other classes. So, this class represent the very dry seasons with very low amount of precipitation and very low number of event or rainy days. Class 3 groups very exceptional seasons, characterized by some strong and long events, with an important quantity of rainfall and a strong intensity. This class represent the very wet seasons with extreme events. Class 4 is characterized by a good quantity of precipitation but less than the class 3, also the mean intensity and the total duration is also strong. But the event number is the most important compared to other clusters. This class represent the wet seasons with intermittency of precipitation in the season.

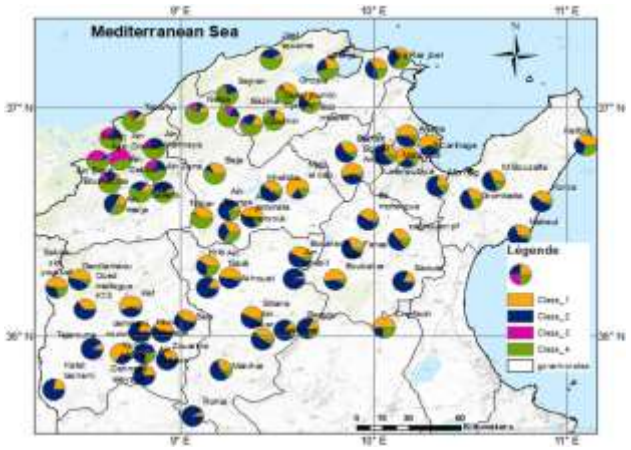


Fig.6: Spatial distribution of classes

The (Fig.6) shows the percentages of each class for each rain gauge station over 50 years. It shows that the rainy seasons, classes 3 and 4, are generally located in the northern part of the region (pink and green colors) and the dry seasons, classes 1 and 2 (yellow and blue colors), are located in the southern part of the study area. The wet area is directly influenced by North West flux coming from the Atlantic during the winter season. The exceptional stations with pink color are located in a forest area. In the south of the studied area the dry seasons (blue) dominates and this region is known by a warm desert climate. For the temporal distribution of classes, the (Fig.7) represent the percentage of stations in each cluster over the years.

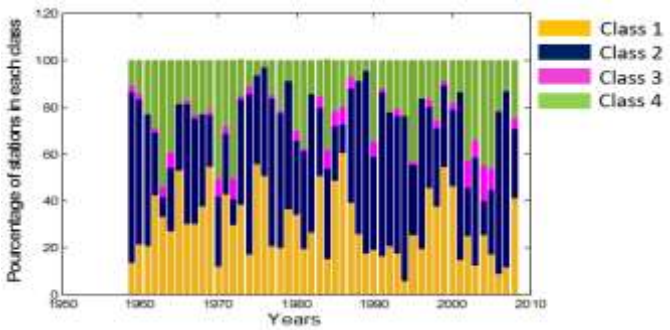


Fig.7: Annual distribution of classes

This figure allows to identify some exceptional seasons like 1963, 1970, 1973, 2002, 2003, 2004 and 2005 seasons where the class 3 and 4 dominate. We observe also the predominance of class 1 and 2 for more than 90 % of stations in 1975, 1976, 1988 and 1989 seasons.

The SOM combined with the HAC clustering for 6 rainfall descriptors allow to identify 4 rainfall situations in DJF seasons. The frequency of each class differ from a rain gauge station to another (Fig.6) and depends on the topography of the area, the proximity to the sea, local meteorological conditions and Mediterranean and Atlantic flux.

The definition of rain event from daily rain gauge data allow the analysis of event characteristics, and the intermittence of precipitation that can be useful to study the teleconnection with global atmospheric circulation and also the variables extracted from event can be used to study various topics like meteorology and hydrology. This classification may be useful in agriculture field, to predict the production especially the cereal and olive.

REFERENCES

[1] J-F Rysman, S. Verrier, Y. Lemaître, E. Moreau. "Space-time variability of the rainfall over the western Mediterranean region: A statistical analysis". Journal of Geophysical Research: Atmospheres, *American Geophysical Union*, 118 (15), pp.8448-8459, 2013.

[2] A.Merzougui and M.Slimani, ''Régionalisation des lois de distribution des pluies mensuelles en Tunisie'', *Hydrological Sciences Journal*, 57:4, 668-685, DOI: 10.1080/02626667.2012.670702, 2011.

[3] A.Douguedroit. ''Precipitation in Tunisia (1951-1980)''. In: Méditerranée, troisième série, tome 66. Recherches climatiques en régions méditerranéennes II, pp. 23-33, 1988.

[4] L.Hénia, ''Les précipitations pluvieuses dans la Tunisie tellienne''. Publ. de l'Université de Tunis, Deuxième série Géographie, vol. 14, 1980.

[5] A. Sharad Parchure, S. Kumar Gedam. "Precipitation Regionalization Using Self-Organizing Maps for Mumbai City, India". *Journal of Water Resource and Protection, 10, 939-956.* DOI: 10.4236/jwarp.2018.109055,2018.

[6] J.G. Joo, J.H. Lee, H.D Jun, J.H .Kim, D.J. Jo." Inter-Event Time Definition Setting Procedure for Urban Drainage Systems". *Water*, 6, 45-58. DOI: 10.3390/w6010045, 2014.

[7] N. Akrour, A. Chazottes, S .Verrier, C. Mallet and L.Barthes . Simulation of yearly rainfall time series at microscale resolution with actual properties: Intermittency, scale invariance, and rainfall distribution, Water Resour. Res., 51, 7417– 7435, doi:10.1002/2014WR016357, 2015.

[8] T.Kohonen, ''Self-Organizing Maps''. Third Edition, Springer, Berlin. https://doi.org/10.1007/978-3-642-97610-0, 1995.

[9] T.Kohonen, Essentials of the Self-Organizing Map. Neural Networks,37,5265.https://doi.org/10.1016/j.neunet.2012.09.01,2013.

[10] J.Vesanto, J.Himberg, E.Alhoniemi and J.Parhankagas. SOM Toolbox for Matlab 5, Report A57. http://www.cis.hut.fi/projects/somtoolbox/, 2000.

[11] F.Murtagh, P. Contreras, Methods of hierarchical clustering. Comput.Res.Repository.abs/1105.0121(2011).http://arxiv.org/abs/1105.0121. 2011

[12] Murtagh, F. & Legendre, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? Journal of classification31, 274–295, 2014.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

PREDICTING INTERANNUAL VARIABILITY OF CLIMATE USING DEEP LEARNING

Changlin Jiang[1,2], Balasubramanya T. Nadiga[1,3], Amir B. Farimani[2]

***Abstract*—Given that the field of near-term prediction of climate is in a nascent stage of development, we examine a deep learning approach to the problem. Preliminary work using a Long Short-Term Memory network architecture with added encoding and decoding is found to be capable of predicting an Earth System Model's leading modes of global temperature variability with prediction lead times of upto a year. Related issues and further extensions are discussed.**

I. MOTIVATION

Predictability of the climate system arises (a) from natural variability of climate, i.e., variability internal to the climate system under conditions of constant external forcing and (b) from the response of the climate system to varying external forcing. Following [1], these predictabilities are termed predictabilities of the first and second kind respectively.

A prediction of the first kind involves being able to accurately track the future evolution of the climate system after estimating its current state. Therefore, the skill of such a prediction is limited on the one hand by errors and uncertainties in the model that is used to approximate the evolution of the climate system, and on the other hand by how errors and uncertainties in the initial condition evolve. Likewise, the skill in a prediction of the second kind is affected not only, again, by model error but also by errors and uncertainties in specifying external forcing.

In the framework of comprehensive earth system modeling (ESM), a host of reasons, including the scientific challenges involved in being able to estimate the state of the climate system with sufficient accuracy and the complex, multiscale and chaotic dynamical nature of the climate system which complicates the process of accounting for uncertainty in the future evolution of errors in the initial state estimate, make predictions of the first kind more difficult than being able to model the response of the climate system to secular changes in external forcing such as due to greenhouse gases (e.g., see [2], [3]). As such, our ability to produce longer term projections, projections that are contolled by external forcing related predictability, is better developed than our ability to produce near-term (interannual) predictions—predictions that are controlled by natural variability related predictability. It should, however, also be noted that the response of the climate system to external forcing can be modulated by natural variability, leading to the response to external forcing being amplified or mitigated on certain time scales of natural variability.

Remaining in the framework of ESM, while initialized predictions of climate seek to augment the external forcing related predictability that is realized in uninitialized long term projections by predictability related to natural variability, there are a number of issues that remain to be resolved before such initialized predictions are skillful. For example, in many ESMs, observation based initialization in the presence of model bias leads to a rather rapid departure of the initialized prediction trajectory from observations necessitating post-processing of the predictions before they can show any skill at all. For these reasons, we concern ourselves with a statistical approach to the problem of near-term predictions of climate in the present article; in particular, we examine a deep learning approach.

II. METHOD

Data and preprocessing: Given the shortness of the observed climate record, we aim to examine the issues involved in a modeling context: we consider the surface air temperature distribution (at a nominal resolution of about a degree) in the pre-industrial control (piControl; a simulation in which external forcing is held fixed) run of the Community Earth System Model (CESM2). This data is publicly available from the CMIP archive at https://esgf-node.llnl.gov/projects/cmip6 and its mirrors and spans a period of 1200 years. After removing the

[1]Los Alamos National Lab., Los Alamos, NM
[2]Carnegie Mellon University, Pittsburg, PA
[3]Corresponding author: B.T. Nadiga, balu@lanl.gov

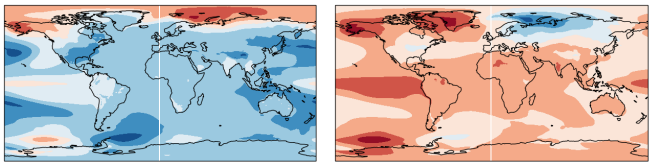


Fig. 1. First two empirical orthogonal functions of surface air temperature variability in the pre-industrial control run of CESM2. PCA is used as a dimension reduction strategy in this study.

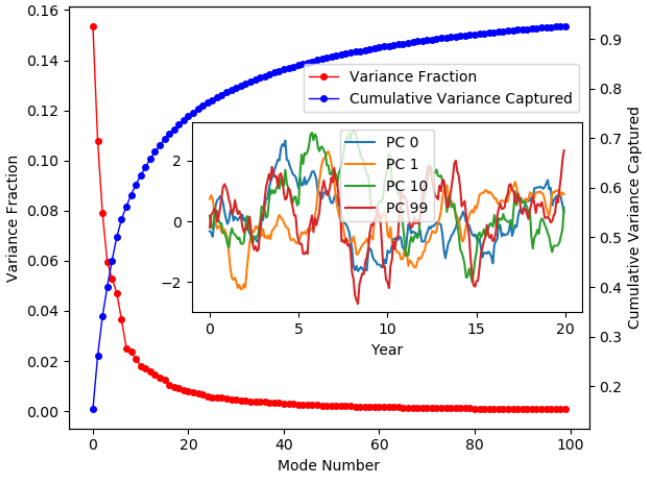


Fig. 2. Fraction of variance captured by individual modes and the cumulative variance captured are shown in the main panel. The top 100 modes are retained in this study and explain in excess of 90% of the variance. The inset shows time variations of the normalized principal components over a 20 year period.

annual seasonal cycle by using a 12 month moving-window average, we perform an empirial orthogonal function (EOF) analysis (equivalently principal component analysis PCA) in order to effectively reduce the dimension of the system being studied. The spatial pattern of the first couple of EOFs is shown in Fig. 1. The main panel of Fig. 2 shows the fraction of variance captured by individual modes and the cumulative variance captured. The top 100 modes are retained in this study and explain in excess of 90% of the variance. The inset in Fig. 2 shows time variations of four (of the 100) normalized principal components over an initial period of 20 years.

Input, Target, Data Augmentation and the Loss Function: We define the problem as follows: given a context_len (equi-spaced, monthly) time sequence of vectors, X with shape (context_len, num_channel) where the vector at a given time comprises the amplitudes of the EOF modes, we are interested in predicting the state of the vector over the following predict_len months, Y with shape (predict_len, num_channel). The left hand panel of Figure 3 shows a schematic of the data. we randomly slice the total data into multiple small sequences. Each slice contains the context and real prediction. In the training part we apply a "sliding window" [4] technique, that is, the target begins with predict_len steps after the input. The model is then optimized to minimize the MSE loss between target and model output, and final model prediction is retrieved from the last predict_len steps. The MSE loss function achieves two goals: ensuring that the model can successfully reconstruct the high-dimensional data in the input, and making accurate prediction. In the testing part, we feed the input into the pre-trained network, and isolate the model prediction from model output. In order to train a more robust model, we conduct data augmentation by using different context lengths and and mix the variable-length data together.

Network Architecture and Optimization: Given that Long Short-Term Memory (LSTM)[5] networks are suitable for the forecasting nature of the problem on hand, we use such an architecture with a further encode-decode [6] capability. First, since the inputs in each mini-batch have different lengths, we pad them to the max length of each mini-batch. Next, the input goes through a embedding layer before the LSTM layer. This linear embedding acts as an "encoder" to extract the hidden feature from the high dimensional input. The output of LSTM is then fed into a similar linear unembedding layer in order to map back to the high dimensional output. Since we apply a zero padding, we introduce a padding mask layer to rule out the influence of zero padding. The model is optimized with classical Adam optimizer[7] and plateau learning rate decay.

III. EVALUATION

In the suite of experiments we present, the model is trained to predict future temperatures based solely on temperatures over the previous five to ten years. In this experiment the initial 80% of the dataset is used to train the model and its ability to predict future temperatures is evaluated over the last 20% of the dataset. The deep learning model is formulated using the pytorch 1.1 framework and the training is performed on NVIDIA 1080 Ti GPUs.

Figure 5 establishes the validity of the approach. These two figures show the root mean square error (RMSE) and the anomaly correlation coefficient (ACC) for each of the 100 EOFs considered as a function of the prediction lead time. Here error is refers to the difference between the predictions of the trained/learnt model and the actual values realized in the detailed

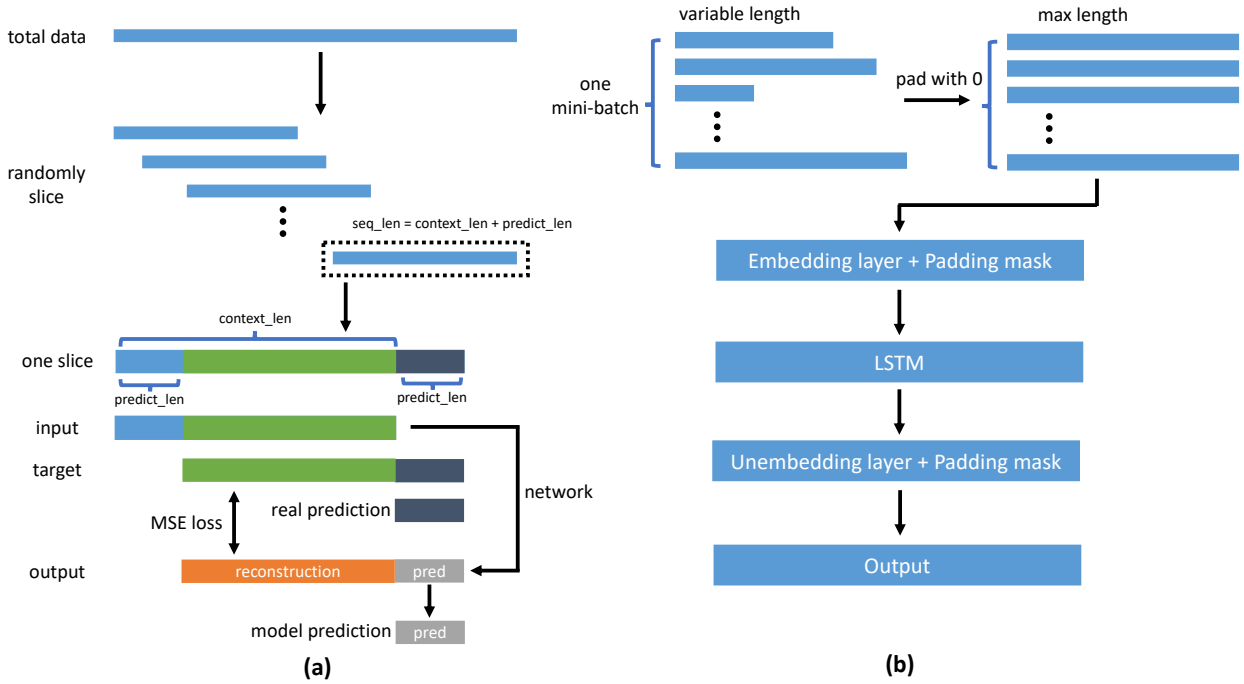


Fig. 3. Dataset generation and network architecture. (a) Dataset generation The total data is randomly sliced into smaller sequences. Each slice consists of variable-length context sequence and fixed-length prediction sequence. Regions which have the same values are represent by the same color. (b) **Network Architecture** Variable length data is padded to the max length of each mini-batch.

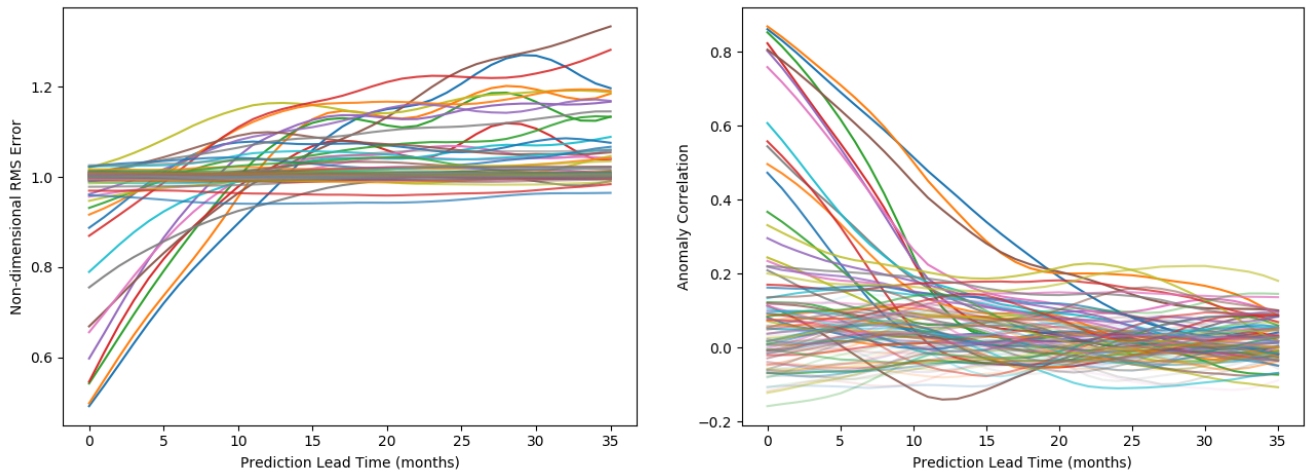


Fig. 4. Non-dimensional root mean square error and anomaly correlation coefficient as a function of prediction lead time for each of the EOFs considered. Any skill in the predictions in limited to the top 10 to 20 EOFs and to lead times of about a year or less.

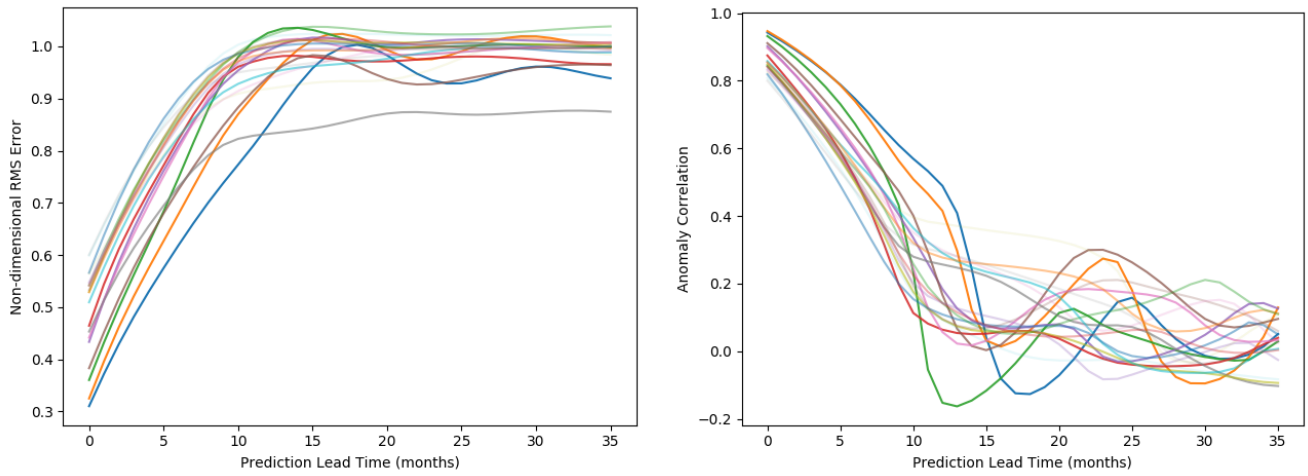


Fig. 5. The experiments are repeated on retaining only the top 20 EOFs and after using the normalized time coefficient time series. Qualitatively similar results with slight quantitative improvement in skill further validates the approach.

ESM simulation on pre-processing; likewise ACC refers to Pearson's correlation between them. Further, the RMSE is non-dimensionalized by the RMS value of the distribution of actual values. Consequently, the predictions have no skill when the non-dimensionalized RMSE reaches a value of about 1. Likewise, the predictions have no skill when the ACC begins to approach zero. In these figures, in general, the higher the mode number, the lesser the skill of the prediction. Further, the measures of skill are seen to decay from initial high values till they are no better than climatology at around a year. That is, the method has significant skill only in predicting the behavior of the top 10 to 20 modes and for lead times up to about a year. To further verify the work, the same experiment is repeated retaining only the top 20 EOFs and now using normalized time series of the coefficients. These results are shown in Fig. **??** and are seen to be qualitatively the same with some quantitative improvements in skill.

This preliminary study points to various issues that we are currently examining and we hope to report on in the future. For example, the the simple data partitioning method that we have used does not account for possible changes in the behavior of the model with time; a better strategy could account for such changes. We have investigated a single network architecture in this study; other appropriate structures need to be considered. We have formulated the prediction problem solely in terms of surface temperature; will a formulation that involves other variables be more skillful? After all, we know a priori that surface temperature is affected by numerous other processes. And so on. Finally, how can the data requirements be mitigated so that the approach can be made relevant to the actual climate system and does the method give us further insight into predictability of climate itself? We exoect that further studies along the lines of the present one will shed light on these issues.

This work was supported by the LDRD program at the Los Alamos National Lab. We thank the WCRP's WG on Coupled Modelling for CMIP, NCAR for producing and making available their model output, and U.S. Department of Energy's PCMDI for the Earth System Science Portals.

REFERENCES

[1] E. Lorenz, "Climatic predictability," *The physical basis of climate and climate modelling*, pp. 132–136, 1975.

[2] G. A. Meehl, L. Goddard, J. Murphy, R. J. Stouffer, G. Boer, G. Danabasoglu, K. Dixon, M. A. Giorgetta, A. M. Greene, E. Hawkins, *et al.*, "Decadal prediction: can it be skillful?," *Bulletin of the American Meteorological Society*, vol. 90, no. 10, pp. 1467–1486, 2009.

[3] G. A. Meehl, L. Goddard, G. Boer, R. Burgman, G. Branstator, C. Cassou, S. Corti, G. Danabasoglu, F. Doblas-Reyes, E. Hawkins, *et al.*, "Decadal climate prediction: an update from the trenches," *Bulletin of the American Meteorological Society*, vol. 95, no. 2, pp. 243–267, 2014.

[4] F. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[5] F. A. Gers, J. Schmidhuber, and F. A. Cummins, "Learning to forget: Continual prediction with lstm," *Neural Computation*, vol. 12, pp. 2451–2471, 2000.

[6] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014.

MACHINE LEARNING OF COMMITTOR FUNCTIONS FOR PREDICTING HIGH IMPACT CLIMATE EVENTS

Dario Lucente, Stefan Duffner, Corentin Herbert, Joran Rolland, Freddy Bouchet

***Abstract*—There is a growing interest in the climate community to improve the prediction of high impact climate events, for instance ENSO (El-Niño–Southern Oscillation) or extreme events, using a combination of model and observation data. In this note we explain that, in a dynamical context, the relevant quantity for predicting a future event is a committor function. We explain the main mathematical properties of this probabilistic concept. We compute and discuss the committor function of the Jin and Timmerman model of El-Niño. Our first conclusion is that one should generically distinguish between states with either intrinsic predictability or intrinsic unpredictability. This predictability concept is markedly different from the deterministic unpredictability arising because of chaotic dynamics and exponential sensibility to initial conditions. The second aim of this work is to compare the inference of a committor function from data, either through a direct approach or through a machine learning approach using neural networks. We discuss the consequences of this study for future applications to more complex data sets.**

I. INTRODUCTION

The low frequency modes of variability of the climate system, for instance ENSO [1]–[9], have a huge impact on nature and human societies, through their local or global signatures. Rare events, such as heat waves, floods, or hurricanes, may also have a huge impact [10]–[14]. Predicting the occurrence of such events is thus a major challenge [10], [11]. Because the dynamics of the climate system is chaotic, one usually distinguishes between time scales much shorter than a Lyapunov time[1] for which a deterministic weather forecast is relevant, and time scales much longer than a mixing times beyond which any deterministic forecast is irrelevant and only climate averaged or probabilistic quantities can be predicted. However, for most applications cited above, the largest interest is for intermediate time scales for which some information, more precise than the climate averages, might be predicted, but for which a deterministic forecast is not relevant. We call this range of time scales *the predictability margin*. As far as applications are concerned, many cases of "medium-range forecast" are within the "predictability margin" range. As a paradigmatic example, we study in this work the probability that El-Niño might occur following year. Another example could be: What is the probability of a heat wave of a given amplitude to happen next summer, given the state of the atmosphere, ocean, and soil moisture, in Spring? For such questions, the system really behaves in a stochastic way as can be seen from the top panel of figure (1), which shows the behavior of El-Niño3 anomaly.

We stress in this work that the prediction problem at the predictability margin is of a probabilistic nature. Indeed, such time scales might typically be of the order of the Lyapunov time scale or larger, where errors on the initial condition and model errors limit our ability to compute deterministically the evolution. However, we stress that the Lyapunov time scale, a global quantity, is clearly not the relevant dynamical quantity for this predictability problem. By contrast, at the predictability margin, the predictability clearly depends on the current state of the system. What is then the relevant mathematical concept? The first aim of this work is to introduce in the field of climate the notion of the committor function [15], [16]. A committor function is the probability that an event will occur or not in the future, as a function of the current state of the system. For the El-Niño case, this committor function will be the probability that an observable $\mathcal{O}$ of the system reaches a given threshold within a time T [17]. The definition and mathematical properties of committor functions are introduced in section III.

The first result of this work is to demonstrate, using

[1]The Lyapunov time can be defined as $t_L = \frac{1}{\lambda}$ where $\lambda = \lim_{t\to\infty} \frac{1}{t}\log(\frac{\delta x(t)}{\delta x(0)})$ and δx is the distance between two different trajectories.

the committor function, that a predictability margin exists for El-Niño. This demonstration is performed within the Jin and Timmermann model, a low dimensional model proposed to explain the decadal amplitude changes of El-Niño [4], [5] (section II). From the computed committor function for the Jin and Timmerman model (section III), we obtain the second main result of this work. This result is the characterisation of regions of the phase space with qualitatively different predictability properties. For example for the intermediate stochasticity regime at the predictability margin, we delineate 4 regions, see (Fig. 4b) that will be explain in section III. First, two regions of perfect predictability, where the event will occur with probability 0 or 1, respectively. Second, regions with good predictability properties where a value of the probability $0 < q < 1$ can clearly be predicted with very mild dependance with respect to initial condition. We call this area the *probabilistically predictable region*. Third, regions which are unpredictable in practice, because the strong dependance with respect to the initial condition prevent any practical prediction, either deterministic or probabilistic. The existence of such features, and especially the new and most interesting *probabilistically predictable region*, should be generic for most prediction problems in climate dynamics.

We will explain that committor functions solve Dirichlet problems. However such partial differential equations are extremely difficult to solve especially for high-dimensional systems. Could we compute it directly from data? There is currently a growing interest to estimate relevant dynamical quantities directly from available data, for instance using machine learning techniques [7], [8], [18]–[21]. The second aim of this work is to propose a machine learning approach, using neural networks, to compute committor functions (section IV). The third result of this work is the demonstration of practicality of this approach on the example of a simple dynamics. We conclude by discussing the feasibility of the computation of a committor function using neural networks for the Jin and Timmerman model, and for more complex data sets related to other climate applications.

II. The Jin and Timmermann model for ENSO

The El-Niño phenomenon consists in an increase of the Sea Surface Temperature in the eastern equatorial Pacific Ocean and it is caused by a large-scale interaction between the equatorial Pacific Ocean and the global atmosphere. El-Niño is also related with the Southern-Oscillation phenomenon and the global phenomenon is called El-Niño–Southern Oscillation (ENSO). In order to explain this phenomenon, in 1997 Jin introduced a simple dynamical model that accounts for the recharge-discharge mechanism which is at the basis of ENSO [2], [3]. This model was later extended by Timmermann [4] and was related to the decadal amplitude changes of ENSO [5].

This model features the evolution of three variables:

1) T_1, the Sea Surface Temperature in the western equatorial Pacific Ocean,
2) T_2, the Sea Surface Temperature in the eastern equatorial Pacific Ocean,
3) h_1, the thermocline depth anomaly in the western Pacific.

The equations can be either deterministic or stochastic, a source of stochasticity being the variable wind stress related to the Walker circulation [4], [5]. After a change of variables [22], from physical to dimensionless ones, the equations, introduced in [4], [5], read

$$\begin{aligned} x' &= \rho\delta(x^2 - ax) + x(x + y + c - c\tanh(x + z)) - D_x(x, y, z)\xi_t, \\ y' &= -\rho\delta(x^2 + ay) + D_y(x, y, z)\xi_t, \\ z' &= \delta(k - z - \frac{x}{2}), \end{aligned} \tag{1}$$

where x is $T_1 - T_2$ divided by a reference temperature, y is related to T_1, and z is related to the thermocline depth h_1 (see [22]). The term $\xi(t)$ is a Gaussian white noise, $D_x(x, y, z) = [(1 + \rho\delta)x^2 + xy + cx(1 - \tanh(x + z))]\sigma$ and $D_y(x, y, z) = \rho\delta x^2\sigma$. The control parameters $[\delta, \rho, c, k, a, \sigma]$ are related to physical quantities [22]. We first describe the phenomenology when the noise is switched off ($\sigma = 0$). For some parameter values, the system has only one attractor, a periodic orbit with oscillations that increase up to strong El-Niño events [22]. For other parameters the system has two different attractors: one periodic attractor that contains strong El-Niño events and one strange attractor without El-Niño events [6]. These two attractors are intertwined with each other as illustrated in figure (2). Figure (1) shows a qualitative comparison of the eastern Pacific sea surface temperature anomalies for the periodic attractor with the El-Niño3 index. Both the measurements and the model display positive temperature anomaly excursions with a return time of approximately 20 years. For this dynamics, we define a strong El-Niño event as any situation when x becomes larger than the threshold $\epsilon = -1$. Following [6], we have chosen these values for the parameters $[\delta, \rho, c, k, a] = [0.225423, 0.3224, 2.3952, 0.4032, 7.3939]$.

The level of stochasticity is controlled by the noise amplitude σ. For small σ, the dynamics can switch from

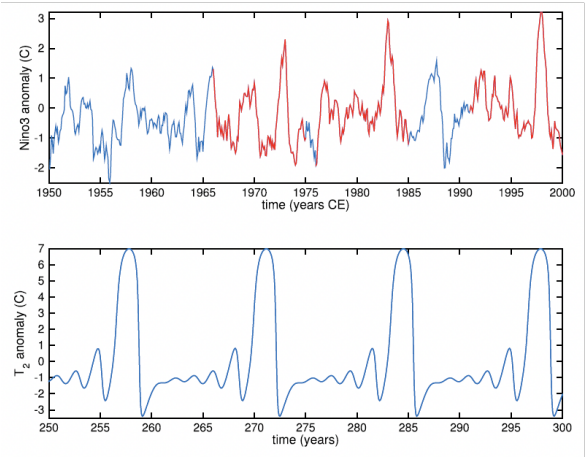


Fig. 1: Top plot: observed sea surface temperature anomalies, spatially averaged over the Niño-3 region. Bottom: eastern Pacific sea surface temperature anomalies simulated with the Jin and Timmermann model (figures from [22]).

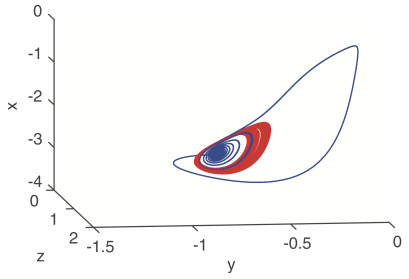


Fig. 2: The two intertwined attractors of the Jin and Timmermann model (periodic attractor in blue and chaotic one in red). Plot from [6].

one attractor to the other. The occurence of the next El-Niño event is then stochastic. Such a mode switching between attractors also occurs when the parameter a is time periodic, mimicking a seasonal forcing [6]. For large values of σ, the dynamics is completely dominated by the noise and the distinction between the two attractors becomes meaningless.

III. COMMITTOR FUNCTIONS

For a stochastic or a deterministic system, the probability $q(\mathbf{x})$ that a trajectory starting from the point $\mathbf{x}$ reaches a set B of the phase space before another set A, as a function of the initial condition $\mathbf{x}$ is called a committor function [15], [16], [23]–[27]. It is thus a function on the phase space of the system.

Let us be more specific for the El-Niño prediction problem within the Jin and Timmerman model. $\mathbf{x} = (x, y, z)$ is the vector of the model phase space and $\{\mathbf{X}(t)\}_{0 \leq t \leq T}$ is one realisation of the dynamics. We consider the two sets $A = \{(\mathbf{x}, t)\,|t \geq T\}$ and $B = \{(\mathbf{x}, t)\,|x > \epsilon\}$. The definitions of A and B are such that these sets are two different regions of the phase space spanned by the new variable $\boldsymbol{\eta} = (\mathbf{x}, t)$. We define the first hitting time of a set C as

$$\tau_C(\mathbf{x}) = \inf\{t : (\mathbf{X}(t), t) \in C | \mathbf{X}(0) = \mathbf{x}\}, \tag{2}$$

The committor function $q(\mathbf{x})$ is the probability that the first hitting time of set B is lower than the first hitting time of set A, i.e.:

$$q(\mathbf{x}) = \mathbb{P}(\tau_B(\mathbf{x}) < \tau_A(\mathbf{x})). \tag{3}$$

For the Jin and Timmerman model, the committor is thus the probability that the variable x reaches the threshold ϵ before time T.

When the dynamics is a stochastic differential equation, which the case of the Jin and Timmerman model, one can prove that the committor function $q(\mathbf{x})$ is the solution of the Dirichlet problem [16], [24]

$$\mathcal{L}q(\mathbf{x}) = 0 \text{ with } \mathrm{q}(\mathbf{x}) = 0 \text{ if } \mathbf{x} \in \mathrm{A} \text{ and } \mathrm{q}(\mathbf{x}) = 1 \text{ if } \mathbf{x} \in \mathrm{B}, \tag{4}$$

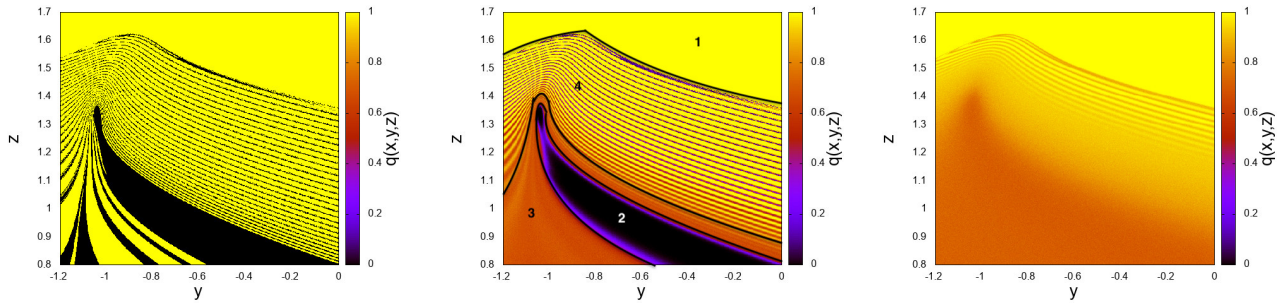
where $\mathcal{L}$ is the infinitesimal generator of the stochastic process ($\mathcal{L}$ is the adjoint of the Fokker-Planck operator)

$$\mathcal{L} = \sum_i a_i(\mathbf{x}) \frac{\partial}{\partial x_i}(\cdot) + \sum_{ij} D_{ij}(\mathbf{x}) \frac{\partial^2}{\partial x_i \partial x_j}(\cdot). \tag{5}$$

Using a numerical simulation, one can simply generate N different trajectories of length T with initial condition $\mathbf{X}(0) = \mathbf{x}$. An estimate of the committor function $q(\mathbf{x})$ is simply $\frac{N_1}{N}$, where N_1 is the number of trajectories that reached the threshold ϵ.

For the Jin and Timmermann model, we have chosen a value of T slightly larger than the period of the periodic attractor. With this value we are at the predictability margin, with a value of T of the order of the Lyapunov time, and of the order of the natural periodicity of El-Niño. This situation is thus analogous to trying to predict wether El-Niño will occur during next winter in the real climate dynamics. We also note that, since the averaged time required to switch from one attractor to the other one is greater than the period of periodic trajectories, each trajectory starting in one point of the periodic attractor almost certainly will reach the threshold $\epsilon = -1$.

Figure (3) shows the committor function q, for different values of σ. As q is a function of 3 variables (x, y, z), we have chosen to represent a cut of q in the plane $x = -2.831$. Fig. (3a) shows q for the deterministic dynamics ($\sigma = 0$). For deterministic



(a) $\sigma = 0$ (b) $\sigma = 0.00005$ (c) $\sigma = 0.001$

Fig. 3: Colour plot of the committor q versus (y, z) for $x = -2.8310$, for $\sigma = 0$, 0.00005 and 0.001 respectively.

dynamics, as the future is completely determined by the initial condition, q is equal to either 0 or 1. We see 3 regions. First, for larger values of z, in a large yellow area all trajectories reach the threshold and $q = 1$. Second, in a thick black band, no trajectory reaches the threshold and $q = 0$. Those two first regions are areas where the occurence or not of El-Niño is easily predicted. Third, everywhere else, we see very fine filaments of alternating yellow and black values. In this area, because of the sensitive dependance on the initial conditions, a small change of the initial conditions lead to a different outcome, and the occurence of El-Niño is very difficult to predict.

When adding stochasticity, one clearly see by comparison of figures (3a), (3b) and (3c) that the effect of a small noise blurs the visible structures of the deterministic case. For larger noise values, figure (3c) shows that while the deterministic predictability is lost for most initial points ($q \neq 0$ and $q \neq 1$), the committor function is smooth nearly everywhere. This means that the occurence of El-Niño is probabilistically predictable (the value of the probability can be determined in practise as it does not depend wildly on the initial condition).

The most interesting case is probably the one with the intermediate stochasticity value $\sigma = 0.00005$. In figure (3b), we delineate 4 regions. First, two regions of perfect predictability, where the event will occur with probability 0 (region 2) or 1 (region 1), respectively. Second, regions 3) with good predictability properties where a value $0 < q < 1$ can clearly be predicted with very mild dependance with respect to initial conditions. We call this area the *probabilistically predictable region*. Third, regions 4) which are unpredictable in practice, because the strong dependance with respect to the initial condition prevent any practical prediction, either deterministic or probabilistic. While regions 1), 2) and 4) are reminiscent of their deterministic counterparts, region 3) is not. It is a region where the stochasticity is large enough to smooth out the deterministic values of q. This occurs even at very low value of the stochasticity, probably in relations to regions leading to extremely unstable parts of the phase space, for instance for trajectories passing close to unstable fixed points or orbits. The existence of such features, and especially the new and most interesting *probabilistically predictable region*, region 3), should be generic for most prediction problems in climate dynamics.

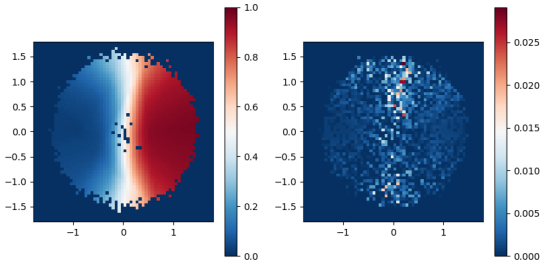


Fig. 4: Left: committor function for a two well gradient dynamics estimated using a neural network. Right: Error in the estimation of the committor function.

IV. LEARNING THE COMMITTOR FUNCTION WITH A NEURAL NETWORK

Estimating the committor function from its definition (3) requires a huge amount of data that is expected to increase exponentially with the phase space dimension. In order to cope with this issue, it could be very interesting to use committor regression, from data, using neural networks. In order to test this idea, we first note that the committor estimation amounts at regressing the parameter of a spatially dependent Bernoulli outcome between occurence B with probability $q(\mathbf{x})$ and occurence A with probability $1 - q(\mathbf{x})$. As a consequence a natural

loss function to be optimised by a neural network is the log-likelihood

$$C = \frac{1}{N}\sum_{n=1}^{N}\{y_n \log\left[1+\exp(-s(\mathbf{X_n}))\right] + (1-y_n)\log\left[1+\exp(s(\mathbf{X_n}))\right]\}$$

where q and s are related through the logistic function $q(\mathbf{x}) = 1/\left[1+\exp(-s(\mathbf{x}))\right]$. The data $\{(\mathbf{X}_n, y_n)\}_{1\leq n\leq N}$ couples each represent a phase space point $\mathbf{X}_n$ corresponding to an initial condition of the dynamics and a value y_n equal to either 1 if the trajectory reaches B before A, or 0 otherwise. The function s is determined as the minimiser of C. The committor function q is then computed from the relation between q and s.

We have tested this approach on a simple gradient stochastic dynamics with two degrees of freedom. The deterministic part of the dynamics has two point attractors. In a limit of small noise, we look for the committor function defined as the probability that the trajectory reaches a neighbourhood of the first attractor before reaching the neighbourhood of the second one. For this simple example, we trained the neural network with simulated trajectories. The neural network model had a standard 3-layer fully connected Multilayer Perceptron (MLP) architecture (32, 64 and 1 neurons respectively) and Rectified Linear Unit (ReLU) activation functions for the hidden layers. Figure (4) shows the estimated committor and the computed error, using as a benchmark a committor computed from its definition (3) and using an extremely long data set. This figure clearly demonstrates the ability of the neural network to learn precisely the committor for this simple example.

In future works, we will use this machine learning approach to learn the committor function for El-Niño, within the Jin and Timmerman model. Our main aim will be to demonstrate the efficiency of this approach and to estimate the required amount of data for genuine climate applications.

ACKNOWLEDGMENTS

This work has received funding through the ACADEMICS grant of the IDEXLYON, project of the Université de Lyon, PIA operated by ANR-16-IDEX-0005. The computation of this work were partially performed on the PSMN platform of ENS de Lyon. During this project, we benefitted from scientific discussions with Patrice Abry, Pierre Borgnat and Charles Edouard Brehier.

REFERENCES

[1] H. A. Dijkstra, *Nonlinear climate dynamics*. Cambridge University Press, 2013.
[2] F.-F. Jin, "An equatorial ocean recharge paradigm for enso. part i: Conceptual model," *Journal of the atmospheric sciences*, vol. 54, no. 7, pp. 811–829, 1997.
[3] F.-F. Jin, "An equatorial ocean recharge paradigm for enso. part ii: A stripped-down coupled model," *Journal of the Atmospheric Sciences*, vol. 54, no. 7, pp. 830–847, 1997.
[4] A. Timmermann, F.-F. Jin, and J. Abshagen, "A nonlinear theory for el niño bursting," *Journal of the atmospheric sciences*, vol. 60, no. 1, pp. 152–165, 2003.
[5] A. Timmermann and F.-F. Jin, "A nonlinear mechanism for decadal el niño amplitude changes," *Geophysical Research Letters*, vol. 29, no. 1, pp. 3–1, 2002.
[6] J. Guckenheimer, A. Timmermann, H. Dijkstra, and A. Roberts, "(un) predictability of strong el niño events," *Dynamics and Statistics of the Climate System*, vol. 2, no. 1, p. dzx004, 2017.
[7] Q. Y. Feng and H. A. Dijkstra, "Climate network stability measures of el niño variability," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 3, p. 035801, 2017.
[8] P. D. Nooteboom, Q. Y. Feng, C. López, E. Hernández-García, and H. A. Dijkstra, "Using network theory and machine learning to predict el ni\˜ no," *arXiv preprint arXiv:1803.10076*, 2018.
[9] J. Ludescher, A. Gozolchiani, M. I. Bogachev, A. Bunde, S. Havlin, and H. J. Schellnhuber, "Very early warning of next el niño," *Proceedings of the National Academy of Sciences*, vol. 111, no. 6, pp. 2064–2066, 2014.
[10] F. Ragone, J. Wouters, and F. Bouchet, "Computation of extreme heat waves in climate models using a large deviation algorithm," *Proceedings of the National Academy of Sciences*, vol. 115, no. 1, pp. 24–29, 2018.
[11] C. B. Field, V. Barros, T. F. Stocker, and Q. Dahe, *Managing the risks of extreme events and disasters to advance climate change adaptation: special report of the intergovernmental panel on climate change*. Cambridge University Press, 2012.
[12] A. AghaKouchak, D. Easterling, K. Hsu, S. Schubert, and S. Sorooshian, *Extremes in a changing climate: detection, analysis and uncertainty*, vol. 65. Springer Science & Business Media, 2012.
[13] S. C. Herring, M. P. Hoerling, T. C. Peterson, and P. A. Stott, "Explaining extreme events of 2013 from a climate perspective," *Bulletin of the American Meteorological Society*, vol. 95, no. 9, pp. S1–S104, 2014.
[14] D. Coumou and S. Rahmstorf, "A decade of weather extremes," *Nature climate change*, vol. 2, no. 7, p. 491, 2012.
[15] E. Vanden-Eijnden, "Transition path theory," in *Computer Simulations in Condensed Matter Systems: From Materials to Chemical Biology Volume 1*, pp. 453–493, Springer, 2006.

[16] E. Weinan, W. Ren, and E. Vanden-Eijnden, "Transition pathways in complex systems: Reaction coordinates, isocommittor surfaces, and transition tubes," *Chemical Physics Letters*, vol. 413, no. 1-3, pp. 242–247, 2005.

[17] T. Lestang, F. Ragone, C.-E. Bréhier, C. Herbert, and F. Bouchet, "Computing return times or return periods with rare event algorithms," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2018, no. 4, p. 043213, 2018.

[18] S. Giffard-Roisin, M. Yang, G. Charpiat, B. Kégl, and C. Monteleoni, "Fused deep learning for hurricane track forecast from reanalysis data," 2018.

[19] S. Giffard-Roisin, D. Gagne, A. Boucaud, B. Kégl, M. Yang, G. Charpiat, and C. Monteleoni, "The 2018 climate informatics hackathon: Hurricane intensity forecast," 2018.

[20] S. Giffard-Roisin, M. Yang, G. Charpiat, B. Kégl, and C. Monteleoni, "Deep learning for hurricane track forecasting from aligned spatio-temporal climate datasets," 2018.

[21] Q. Li, B. Lin, and W. Ren, "Computing committor functions for the study of rare events using deep learning with importance sampling," 2018.

[22] A. Roberts, J. Guckenheimer, E. Widiasih, A. Timmermann, and C. K. Jones, "Mixed-mode oscillations of el nino–southern oscillation," *Journal of the Atmospheric Sciences*, vol. 73, no. 4, pp. 1755–1766, 2016.

[23] J.-H. Prinz, M. Held, J. C. Smith, and F. Noé, "Efficient computation, sensitivity, and error analysis of committor probabilities for complex dynamical processes," *Multiscale Modeling & Simulation*, vol. 9, no. 2, pp. 545–567, 2011.

[24] E. H. Thiede, D. Giannakis, A. R. Dinner, and J. Weare, "Galerkin approximation of dynamical quantities using trajectory data," *The Journal of Chemical Physics*, vol. 150, no. 24, p. 244111, 2019.

[25] C. Schütte and M. Sarich, "A critical appraisal of markov state models," *The European Physical Journal Special Topics*, vol. 224, no. 12, pp. 2445–2462, 2015.

[26] C. Schütte, F. Noé, J. Lu, M. Sarich, and E. Vanden-Eijnden, "Markov state models based on milestoning," *The Journal of chemical physics*, vol. 134, no. 20, p. 05B609, 2011.

[27] L. J. Lopes and T. Lelièvre, "Analysis of the adaptive multilevel splitting method on the isomerization of alanine dipeptide," *Journal of computational chemistry*, vol. 40, no. 11, pp. 1198–1208, 2019.

TOWARDS UNSUPERVISED SEGMENTATION OF EXTREME WEATHER EVENTS

Adam Rupe[1], Karthik Kashinath[2], Nalini Kumar[3], Victor Lee[3], Prabhat [2], James P. Crutchfield [1]

***Abstract*—Extreme weather is one of the main mechanisms through which climate change will directly impact human society. Coping with such change as a global community requires markedly improved understanding of how global warming drives extreme weather events. While alternative climate scenarios can be simulated using sophisticated models, identifying extreme weather events in these simulations requires automation due to the vast amounts of complex high-dimensional data produced. Atmospheric dynamics, and hydrodynamic flows more generally, are highly structured and largely organize around a lower dimensional skeleton of coherent structures. Indeed, extreme weather events are a special case of more general hydrodynamic coherent structures. We present a scalable physics-based representation learning method that decomposes spatiotemporal systems into their structurally relevant components, which are captured by latent variables known as *local causal states*. For complex fluid flows we show our method is capable of capturing known coherent structures, and with promising segmentation results on CAM5.1 water vapor data we outline the path to extreme weather identification from unlabeled climate model simulation data.**

I. INTRODUCTION

Life across the globe has survived and thrived by adapting to its local weather, including extreme events such as strong winds and floods from cyclones, drought and heat waves from blocking events and large-scale atmospheric oscillations, and critically-needed precipitation from atmospheric rivers. Driven by an ever-warming climate, extreme weather events are changing in frequency and intensity at an unprecedented pace [1], [2]. We need to understand these events and their driving mechanisms to enable communities to continue to adapt and thrive.

High-resolution, high-fidelity global climate models are an indispensable tool for investigating climate change. A multitude of climate change scenarios are now being simulated, each producing 100s of TBs of data. Currently, climate change is assessed in these simulations using summary statistics such as mean global sea surface temperature. This is inadequate for answering detailed questions about the effects of climate change on extreme weather events. Due to the sheer size and complexity of these simulated data sets, it is essential to develop robust and automated methods that can provide the deeper insights we seek.

Recently, supervised Deep Learning (DL) techniques have been applied to address this problem [3], [4], [5], [6]. Further progress however has been stymied by two daunting challenges: reliance on labeled training data and interpretability of trained models. The DL models used in the above studies are trained using the automated heuristics of TECA [7] for proximate labels. This is necessary because, simply put, there currently is no ground truth for pixel-level identification of extreme weather events [8]. While the results in [3] show that DL can improve upon TECA, the results of [6] reach accuracy rates over 97% and thus essentially just reproduce the output of TECA. The supervised learning paradigm of optimizing objective metrics (e.g. training and generalization error) breaks down here [9]; TECA is not ground truth and we do not know how to train a DL model to disagree with TECA in just the right way to get closer to "ground truth".

To avoid this issue, a campaign is currently underway to generate expert-labeled training data [10]. Supervised DL models trained on this data will automate expert-level curation of large climate data sets for extreme weather detection. In this case there too will be challenges. Though an improvement over automated heuristics, expert-labeled data is still not an objective ground truth. Further, while human experts can debate the subtleties of physical characteristics of extreme weather events, the interpretability problem [11] prevents us from probing a trained DL model to determine exactly how and why it identifies (or misidentifies) specific events.

To circumvent these challenges of DL-based ap-

Corresponding author: A. Rupe, atrupe@ucdavis.edu
[1]Complexity Sciences Center and Department of Physics, University of California Davis [2]NERSC, Lawrence Berkeley National Laboratory [3] Intel Corporation

proaches, here we take an alternative physics-based unsupervised approach, complementary to DL.

Whereas DL takes inspiration from the human visual processing system to identify patterns in images without consideration for what constitutes a "pattern", our method builds on a theory that seeks to understand the physical nature of pattern without consideration for the visual system that can readily identify such patterns. When viewing a video of a complex fluid flow we do not track the evolution of each individual pixel. Our vision instead focuses on relatively few collective features, generally referred to as *coherent structures* [12], [13], that the flow organizes around. Beyond fluid flows, coherent structures in spatiotemporal systems can similarly be understood as key organizing features that heavily dictate the dynamics of the full system, and thus provide a natural dimensionality reduction. Understanding the lower-dimensional coherent structures gets us most of the way to understanding and predicting the full higher-dimensional system, and, as with extreme weather, the coherent structures are often the features of interest.

Our approach seeks to understand the physical nature of coherent structures so that we can discover and identify them in spatiotemporal systems. It is difficult however, if not impossible, to give an actionable definition of coherent structures as the solution of a general mathematical coherence principle derived from equations of motion. Identifying and predicting complex emergent behaviors starting from fundamental laws is typically infeasible [14]. For example, despite knowing the equations of hydrodynamics and thermodynamics, which critically govern the dynamics of hurricanes, many aspects of how hurricanes form and evolve are still poorly understood [15].

As a response, research on complex, nonlinear systems shifted to focus directly on system behaviors rather than governing equations.The resulting *behavior-driven theories* (e.g. [16], [17], [18], [19]), which lie at the interface of physics and machine learning, provide a new means of scientific discovery directly from data. Our approach to unsupervised extreme weather event detection is through a behavior-driven theory of coherent structures in spatiotemporal systems. Below we give some basics of the theory then demonstrate its utility by identifying known coherent structures in 2D turbulence simulation data and observational data of Jupiter's clouds from the NASA Cassini spacecraft. Finally, we show promising results on CAM5.1 water vapor data and outline the path to extreme weather event segmentation masks.

II. METHOD: LOCAL CAUSAL STATES

Our behavior-driven theory of coherent structures builds on a more general theory of pattern and structure in natural systems. A quantitative theory of structure need be probabilistic, capturing structure in ensembles of behavior, and algebraic, generalizing from the group-theoretic formalism of exact symmetry to the semigroup algebra of finite-state machines. The mathematical representation of the structure of a system's dynamical behavior is given by a minimal, optimally predictive, stochastic model [20], [21], [22]. For a model to optimally predict with minimal resources that model must capture pattern and structure present in the system's behaviors.

Computational mechanics [23] makes this idea operational through the *causal equivalence relation*;

$$\text{past}_i \sim_\epsilon \text{past}_j \iff \Pr(\text{Future}|\text{past}_i) = \Pr(\text{Future}|\text{past}_j) \ .$$

The equivalence classes over pasts induced by the causal equivalence relation are known as the *causal states* of the system; they are the unique minimal sufficient statistic of the past for optimally predicting the future.

For spatiotemporal systems, *lightcones* are used as local notions of past and futures. Two past lightcones ℓ_i^- and ℓ_j^- are causally equivalent if they have the same conditional distribution over future lightcones;

$$\ell_i^- \sim_\epsilon \ell_j^- \iff \Pr(\mathrm{L}^+|\ell_i^-) = \Pr(\mathrm{L}^+|\ell_j^-) \ .$$

The resulting equivalence classes are called *local causal states* [24]. They are the unique minimal sufficient statistic of past lightcones for optimal prediction of future lightcones. The ϵ-function, which generates the causal equivalence classes, maps from past lightcones to local causal states; $\epsilon : \ell^- \mapsto \xi$. Segmentation is achieved by mapping a spacetime field X to its associated local causal state field $S = \epsilon(X)$: every feature $x = X(\vec{r}, t)$ is mapped to its classification label (local causal state) via its past lightcone $\xi = S(\vec{r}, t) = \epsilon\big(\ell^-(\vec{r}, t)\big)$. Crucially, this ensures the global latent variable field S maintains the same geometry of X such that $S(\vec{r}, t)$ is the local latent variable corresponding to the local observable $X(\vec{r}, t)$.

For real-valued systems, such as the fluid flows considered here, local causal state reconstruction requires a discretization to empirically estimate $\Pr(\mathrm{L}^+|\ell^-)$ [25]. We use K-Means to cluster over lightcones with the lightcone distance metric

$$\mathrm{D}_{\mathrm{lc}}(\mathbf{a}, \mathbf{b}) \equiv \sqrt{(a_1 - b_1)^2 + \ldots + \mathrm{e}^{-\tau d(n)}(a_n - b_n)^2} \ ,$$

where $\mathbf{a}$ and $\mathbf{b}$ are flattened lightcone vectors, $d(n)$ is the temporal depth of the lightcone vector at index n,

and τ is the temporal decay rate ($1/\tau$ can be thought of as a coherence time).

III. RESULTS: STRUCTURAL SEGMENTATION

Because the local causal state latent variables are designed to capture structure in spatiotemporal systems, we call this level of segmentation a *structural segmentation*. That is, the classification assignments are labels of unique local causal states. This is an intermediate step for coherent structure segmentation (e.g. classification assignments of 'cylcone', 'atmospheric river', 'background', etc.), for which an additional layer of analysis is needed on top of the local causal states [26]. From comparison with established Lagrangian Coherent Structure results we will now show that physically meaningful coherent structures are captured by our structural segmentation, and then we outline how extreme weather events may be extracted from structural segmentation of global climate data.

Building on the realization that relatively low-dimensional chaotic attractors underlies turbulent fluid flows [27], [28], the Lagrangian Coherent Structure (LCS) approach is grounded in nonlinear dynamical systems theory and seeks to describe the most repelling, attracting, and shearing material surfaces that form the skeletons of Lagrangian particle dynamics [13]. LCS are conjectured to capture localized, emergent structures that organize the large-scale flow. We directly compare our results with the geodesic and LAVD approaches (described below) on the 2D turbulence data set from [29] and the Jupiter data set from [29] and [30].

There are three classes of flow structures in the LCS framework; elliptic LCS are rotating vortex-like structures, parabolic LCS are generalized Lagrangian jet-cores, and hyperbolic LCS are tendril-like stable-unstable manifolds in the flow. The geodesic approach [13], [30] is the state-of-the-art method designed to capture all three classes of LCS and has a nice interpretation for the structures it captures in terms of characteristic deformations of material surfaces. The Lagrangian-Averaged Vorticity Deviation (LAVD) [31] is the state-of-the-art method specifically for elliptic LCS, but is not designed to capture parabolic or hyperbolic LCS.

A. 2D Turbulence

While still complex and multi-scale, the idealized 2D turbulence data provides the cleanest identification of Lagrangian Coherent Structures using our structural segmentation. Figure 1 (a) shows a snapshot of the vorticity field, and (b) and (c) show corresponding snapshots from structural segmentations using different reconstruction parameter values. To reveal finer structural details that persist on shorter time scales, Figure 1 (b) uses $\tau = 0.8$ and $K = 10$. To isolate the coherent vortices, which persist at longer time scales, Figure 1 (c) was produced using $\tau = 0.0$ and $K = 4$. As can be seen in (b), the local causal states distinguish between positive and negative vortices, so for (c) we modded out this symmetry by reconstructing from the absolute value of vorticity.

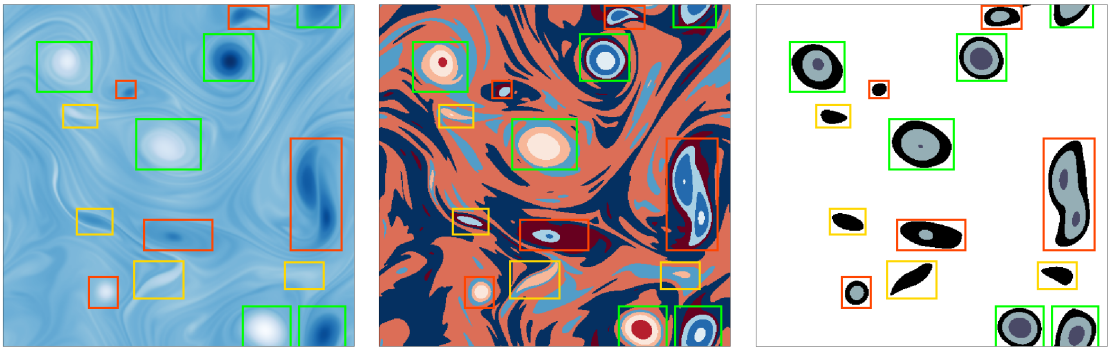
All three images are annotated with color-coded bounding boxes outlining elliptic LCS to directly compare with the geodesic and LAVD LCS results from Figure 9, (k) and (l) respectively, in [29]. Green boxes are vortices identified by both the geodesic and LAVD methods and red boxes are additional vortices identified by LAVD but not the geodesic. Yellow boxes are new structural signatures of elliptic LCS discovered by the local causal states.

Because there is a single background state in (c), colored white, all states not colored white can be assigned a semantic label of `coherent structure` since they satisfy the local causal state definition given in [26] as spatially localized, temporally persistent deviations from generalized spacetime symmetries. Significantly, our method has discovered vortices in the observable field (a) as coherent structures due to the shared geometry with the latent field in (c) where they are identified as locally broken symmetries.

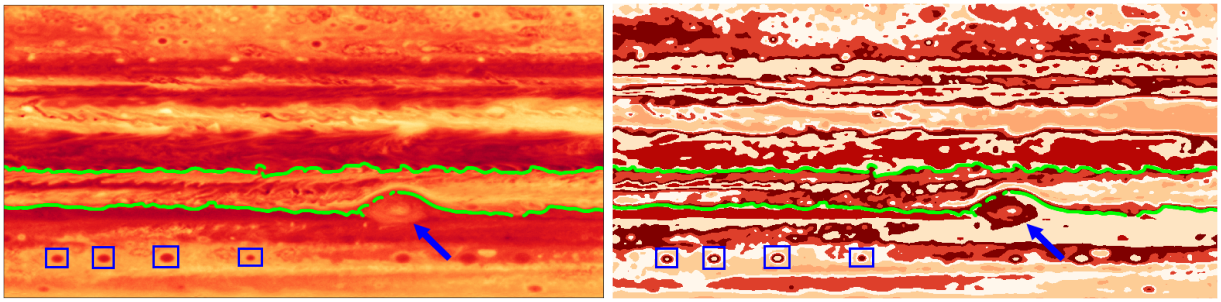
In the finer-scale structural segmentation of (b) we still have states outlining the coherent vortices, as we would expect. If they persist on longer time scales, they will also persist on the short time scale. The additional structure of the background potential flow largely follows the hyperbolic LCS stable-unstable manifolds. Because they act as transport barriers, they partition the flow on either side and these partitions are given by two distinct local causal states with the boundary between them running along the hyperbolic LCS in the unstable direction. For example, the narrow dark blue-colored state in the upper right of (b) indicates a narrow flow channel squeezed between two hyperbolic LCS.

B. Jupiter

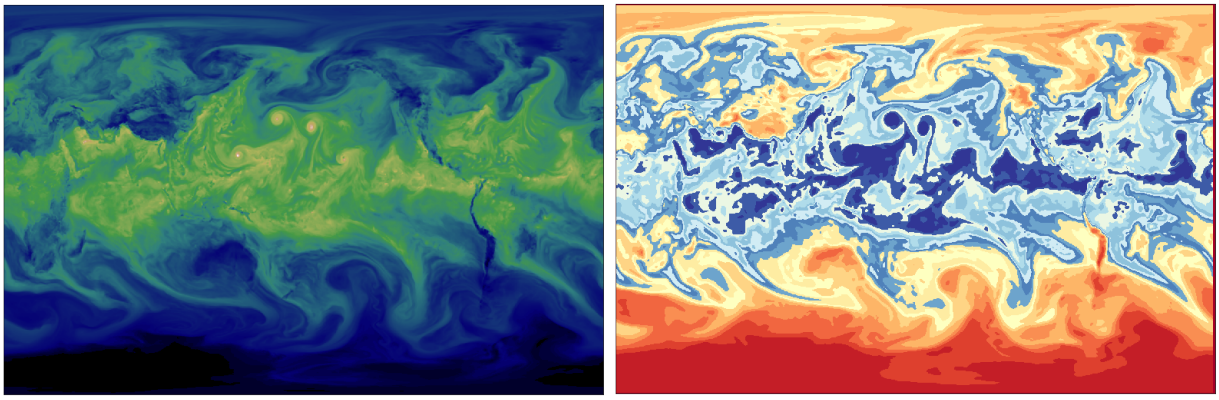
Figure 1 (d) shows a snapshot from the Jupiter cloud data, with corresponding structural segmentation snapshot in (e). The Great Red Spot, highlighted with a blue arrow, is the most famous structure in Jupiter's atmosphere. As it is a giant vortex, the Great Red Spot is identified as an elliptic LCS by both the geodesic and LAVD methods [30], [29]. While the local causal



(a) Turbulence vorticity field (b) Turbulence state field, fine structure (c) Turbulence state field, coarse structure



(d) NASA Cassini Jupiter cloud data in grayscale (e) Jupiter local causal state field



(f) Water vapor field of CAM5.1 climate model simulation (g) Climate local causal state field

Fig. 1. Structural segmentation results for complex fluid flows. Image (a) shows the vorticity observable field for the 2D turbulence data set, with (b) and (c) showing the corresponding latent variable local causal state fields. The segmentation in (c) is tuned to isolate vortices against the background potential flow, while the segmentation in (b) captures additional structure of the background. Image (d) shows the cloud luminosity observable for the Jupiter data set, with the corresponding local causal state field shown in (e). The CAM5.1 water vapor field is shown in (f), with corresponding local causal state field in (g). Each unique color in the latent variable fields (b), (c), (e), and (g) corresponds to a unique local causal state. Full segmentation videos are available on the DisCo YouTube channel [32].

states in (e) do not capture the Great Red Spot as cleanly as the vortices in (b) and (c), it does have the same nested state structures as the turbulence vortices. There are other smaller vortices in Jupiter's atmosphere, most notably the "string of pearls" in the Southern Temperate Belt, four of which are highlighted with blue bounding boxes. We can see in (e) that the pearls are nicely captured by the local causal states, similar to the turbulence vortices in (b).

Perhaps the most distinctive features of Jupiter's atmosphere are the zonal belts. The east-west zonal jet streams that form the boundaries between bands are of particular relevance to Lagrangian Coherent Structure analysis. Figure 11 in [30] uses the geodesic method to identify these jet streams as shearless parabolic LCS, indicating they act as transport barriers that separate

the zonal belts. The particular segmentation shown in (e) captures a fair amount of detail inside the bands, but the edges of the bands have neighboring pairs of local causal states with boundaries that extend contiguously in the east-west direction along the parabolic LCS transport barriers. Two such local causal state boundaries are highlighted in green, for comparison with Figure 11 (a) in [30]. The topmost green line, in the center of (d) and (e), is the southern equatorial jet, shown in more detail in Figure 11 (b) and Figure 12 of [30]. Its north-south meandering is clearly captured by the local causal states.

C. Extreme Weather Events

The strong qualitative correspondence between local causal state structural segmentation and LCS gives validation that our method can capture meaningful structure in complex spatiotemporal systems. Our aim now is to use the structural segmentation to build extreme weather segmentation masks for climate data. Each event, e.g. hurricanes or atmospheric rivers (ARs), are identified as a unique set of structured behaviors that are captured by the local causal states in a structural segmentation.

For example, a structural segmentation of the water vapor field of the CAM5.1 global atmospheric model is shown on YouTube [32] and in Figure 1 (e), (f). While signatures of hurricanes and ARs are visually apparent, these events are not uniquely identifiable from the local causal states. However, hurricanes and ARs have characteristic structural signatures in other physical fields, including thermodynamic quantities such as temperature and pressure. So it is not surprising that they can not be uniquely identified from a structural segmentation of the water vapor field alone; this would be akin to describing hurricanes as just local concentrations of water vapor.

In addition to further optimization and scaling, we are currently working on implementing multi-variate local causal state reconstruction to incorporate additional physical fields. Using, for example, structural segmentation of vorticity, temperature, pressure, and water vapor fields we will be able to identify hurricanes as high rotation objects with a warm, low pressure core that locally concentrate water vapor. Similarly, the inclusion of water vapor transport will help identify ARs, as well as the use of larger lightcone templates that will better capture their large-scale geometry.

Though our structural segmentation requires information from multiple physical observables to identify extreme weather events, the generality of the local causal states will allow us to do this. An automated, objective identification of sets of local causal states across the various physical observables that uniquely corresponds to particular extreme weather events will be challenging but, we believe, achievable.

ACKNOWLEDGEMENTS

Adam Rupe and Jim Crutchfield would like to acknowledge Intel® for supporting the IPCC at UC Davis. Prabhat and Karthik Kashinath were supported by the Intel® Big Data Center. This research is based upon work supported by, or in part by, the U. S. Army Research Laboratory and the U. S. Army Research Office under contract W911NF- 13-1-0390, and used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

REFERENCES

[1] K. A. Emanuel, "The dependence of hurricane intensity on climate," *Nature*, vol. 326, no. 6112, p. 483, 1987.

[2] P. J. Webster, G. J. Holland, J. A. Curry, and H.-R. Chang, "Changes in tropical cyclone number, duration, and intensity in a warming environment," *Science*, vol. 309, no. 5742, pp. 1844–1846, 2005.

[3] M. Mudigonda, S. Kim, A. Mahesh, S. .Kahou, K. Kashinath, D. Williams, V. Michalski, T. OBrien, and Prabhat, "Segmenting and tracking extreme climate events using neural networks," in *Deep Learning for Physical Sciences (DLPS) Workshop, held with NIPS Conference*, 2017.

[4] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica, Prabhat, and M. Houston, "Exascale deep learning for climate analytics," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*, p. 51, IEEE Press, 2018.

[5] M. J. Chiyu, J. Huang, K. Kashinath, Prabhat, P. Marcus, and M. Niessner, "Spherical CNNs on unstructured grids," in *International Conference on Learning Representations*, 2019.

[6] T. Cohen, M. Weiler, B. Kicanaoglu, and M. Welling, "Gauge equivariant convolutional networks and the icosahedral CNN," in *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.), vol. 97 of *Proceedings of Machine Learning Research*, (Long Beach, California, USA), pp. 1321–1330, PMLR, 09–15 Jun 2019.

[7] Prabhat, O. Rübel, S. Byna, K. Wu, F. Li, M. Wehner, and W. Bethel, "TECA: A parallel toolkit for extreme climate analysis," *Procedia Computer Science*, vol. 9, pp. 866–876, 2012.

[8] C. A. Shields, J. J. Rutz, L.-Y. Leung, F. M. Ralph, M. Wehner, B. Kawzenuk, J. M. Lora, E. McClenny, T. Osborne, A. E. Payne, P. Ullrich, A. Gershunov, N. Goldenson, B. Guan, Y. Qian, A. M. Ramos, C. Sarangi, S. Sellars, I. Gorodetskaya, K. Kashinath, V. Kurlin, K. Mahoney, G. Muszynski, R. Pierce, A. C. Subramanian, R. Tome, D. Waliser, D. Walton, G. Wick, A. Wilson, D. Lavers, Prabhat, A. Collow, H. Krishnan, G. Magnusdottir, and P. Nguyen, "Atmospheric

river tracking method intercomparison project (ARTMIP): project goals and experimental design," *Geoscientific Model Development*, vol. 11, no. 6, pp. 2455–2474, 2018.

[9] J. H. Faghmous and V. Kumar, "A big data guide to understanding climate change: The case for theory-guided data science," *Big data*, vol. 2, no. 3, pp. 155–163, 2014.

[10] Prabhat, K. Kashinath, M. Mudigonda, K. Yang, J. Chen, A. Grenier, and B. Toms, "ClimateNet: bringing the power of deep learning to the climate community via open datasets and architectures." https://www.nersc.gov/research-and-development/data-analytics/big-data-center/climatenet/, 2018.

[11] C. Olah, A. Satyanarayan, I. Johnson, S. Carter, L. Schubert, K. Ye, and A. Mordvintsev, "The building blocks of interpretability," *Distill*, 2018. https://distill.pub/2018/building-blocks.

[12] P. Holmes, J. L. Lumley, G. Berkooz, and C. W. Rowley, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*. Cambridge university press, 2012.

[13] G. Haller, "Lagrangian coherent structures," *Ann. Rev. Fluid Mech.*, vol. 47, pp. 137–162, 2015.

[14] P. W. Anderson, "More is different," *Science*, vol. 177, no. 4047, pp. 393–396, 1972.

[15] K. Emanuel, "Tropical cyclones," *Annual review of earth and planetary sciences*, vol. 31, no. 1, pp. 75–104, 2003.

[16] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley, "A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition," *Journal of Nonlinear Science*, vol. 25, no. 6, pp. 1307–1346, 2015.

[17] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, p. 8502, 2015.

[18] N. Rubido, C. Grebogi, and M. S. Baptista, "Entropy-based generating Markov partitions for complex systems," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 3, p. 033611, 2018.

[19] H. Zenil, N. A. Kiani, A. A. Zea, and J. Tegnér, "Causal deconvolution by algorithmic generative models," *Nature Machine Intelligence*, vol. 1, no. 1, p. 58, 2019.

[20] S. Wolfram, "Computation theory of cellular automata," *Comm. Math. Phys.*, vol. 96, p. 15, 1984.

[21] P. Grassberger, "Toward a quantitative theory of self-generated complexity," *Intl. J. Theo. Phys.*, vol. 25, p. 907, 1986.

[22] C. R. Shalizi and J. P. Crutchfield, "Computational mechanics: Pattern and prediction, structure and simplicity," *J. Stat. Phys.*, vol. 104, pp. 817–879, 2001.

[23] J. P. Crutchfield, "Between order and chaos," *Nature Physics*, vol. 8, no. January, pp. 17–24, 2012.

[24] C. Shalizi, "Optimal nonlinear prediction of random fields on networks," *Discrete Mathematics & Theoretical Computer Science*, 2003.

[25] G. Goerg and C. Shalizi, "LICORS: Light cone reconstruction of states for non-parametric forecasting of spatio-temporal systems," *arXiv:1206.2398*, 2012.

[26] A. Rupe and J. P. Crutchfield, "Local causal states and discrete coherent structures," *Chaos*, vol. 28, no. 7, pp. 1–22, 2018.

[27] D. Ruelle and F. Takens, "On the nature of turbulence," *Comm. Math. Phys.*, vol. 20, pp. 167–192, 1971.

[28] A. Brandstater, J. Swift, H. L. Swinney, A. Wolf, J. D. Farmer, E. Jen, and J. P. Crutchfield, "Low-dimensional chaos in a hydrodynamic system," *Phys. Rev. Lett.*, vol. 51, p. 1442, 1983.

[29] A. Hadjighasem, M. Farazmand, D. Blazevski, G. Froyland, and G. Haller, "A critical comparison of Lagrangian methods for coherent structure detection," *Chaos*, vol. 27, no. 5, p. 053104, 2017.

[30] A. Hadjighasem and G. Haller, "Geodesic transport barriers in Jupiter's atmosphere: Video-based analysis," *Siam Review*, vol. 58, no. 1, pp. 69–89, 2016.

[31] G. Haller, A. Hadjighasem, M. Farazmand, and F. Huhn, "Defining coherent vortices objectively from the vorticity," *Journal of Fluid Mechanics*, vol. 795, pp. 136–173, 2016.

[32] "Project disco segmentation videos." https://www.youtube.com/channel/UCwKTJloOOFQHVHDwkpqIdYA, Accessed: 2019-04-10.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

DETECTING WAVEGUIDES FOR ATMOSPHERIC PLANETARY WAVES: CONNECTIONS TO EXTREME WEATHER EVENTS

Rachel H. White[1]

***Abstract*—Recent work has implicated atmospheric waveguides in the occurrence of extreme weather events, including the severe European heatwave of 2003 and Russian heatwave of 2010; such waveguides were also present during the recent European heatwave in June 2019. Previous work has been limited to the study of zonal mean waveguides during time periods of extreme events. In this paper an objective atmospheric waveguide detection algorithm is described, along with its application to the study of waveguide occurrence statistics in the ECMWF ERA-interim and NCEP-DOE R2 re-analysis datasets. Maps of climatological frequency of waveguide occurrence are presented for northern hemisphere summer. We show a connection between anomalously high quasi-stationary wave amplitude (associated with extreme events) and upstream frequency of waveguide occurrence for high ($k = 6 - 8$) wavenumbers.**

I. MOTIVATION

Extreme weather events, such as heatwaves, droughts and flooding, have a devastating impact on society, causing increased mortality and suffering, as well as economic losses. The death toll associated with the European heatwave of 2003 is estimated to exceed 70,000 [1], while the 2010 Russian heatwave killed ∼55,000 people, caused an annual crop yield drop of 25%, and resulted in ∼US$15 billion in economic losses [2]. Many surface temperature extremes, particularly over land, are associated with particular atmospheric wave dynamics, specifically high amplitude atmospheric 'Rossby' (or planetary) waves [3]. In particular, atmospheric waves with a zonal wavenumber of approximately 6-8 (the number of full waves in a latitude band around the globe) have been shown to be connected to extreme events [3], [4]. Some recent extreme events, including the severe European and Russian heatwaves, were associated with high amplitude planetary waves that became relatively stationary for an anomalously extensive period of time, known as quasi-stationary waves [5], [6].

Current research suggests that atmospheric waveguides may play a role in creating conditions conducive to quasi-stationary, amplified, planetary waves [7], [5], [8]. Petoukhov (2013) and Petoukhov (2016) [7], [5] present examples of four large-scale extreme events during which a waveguide was present, in contrast to one date where no waveguide was present and no such extreme events occurred. No analysis has previously been produced, however, on climatological statistics of waveguides calculated on wind fields with high temporal resolution, and thus it was not known how unusual the waveguides observed during the extreme events really were. Previous analysis has also focused on waveguides diagnosed from the zonal mean flow; however, Fragkoulidis et al. (2018) [3] show that many extreme events are associated with Rossby wave packets that do not extend around an entire longitudinal circle. To analyse waveguides in multiple years of daily data in zonally asymmetric flow requires analysis of thousands of K_S distributions; an objective waveguide detection algorithm is developed for this purpose. This paper details the detection algorithm, and presents some results on the climatological statistics of waveguide occurrence as a function of longitude, specifically average waveguide frequency and latitude. These data are presented for two different re-analysis datasets (gridded 'observations'). Lastly, evidence is presented connecting the occurrence of waveguides to the occurrence of anomalously high quasi-stationary wave activity.

II. WAVEGUIDE DETECTION ALGORITHM

Rossby waveguides are features of the atmospheric flow that restrict Rossby wave meridional propagation, effectively trapping Rossby waves in the latitude band of the waveguide [9], [10]. Rossby waveguides can

Corresponding author: R. White, rachel.white@cantab.net [1]Earth Sciences Departments, Barcelona Supercomputing Center

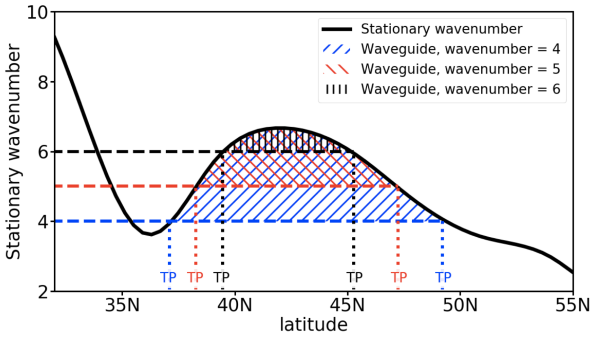


Fig. 1. Idealized distribution of stationary wavenumber K_S showing a waveguide present for planetary waves with zonal wavenumber $4 < k < 7$.

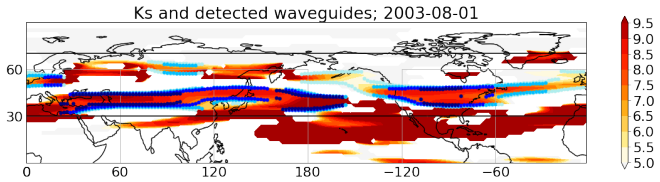


Fig. 2. Example of waveguide detection. Red shading shows K_S from ERA-interim data for 1st August 2003. Blue dots show the TPs where a waveguide is detected for $k = 5$ (lightest blue) to $k = 8$ (darkest blue).

be detected as a local maximum (in the meridional direction) of the barotropic stationary wavenumber K_S: $K_S = \sqrt{\frac{\beta - \partial^2 u/\partial y^2}{u}}$, where u is the zonal wind and β is the meridional gradient of the coriolis parameter. A waveguide requires two turning points (TPs), latitudes at which $K_S = k$, where k is the zonal wavenumber of the trapped wave; one poleward and one equatorward of the local maximum. Figure 1 shows a distribution of K_S that creates a waveguide for waves with zonal wavenumber $4 < k < 7$; the turnings points (TPs) are marked and waveguides are shaded for waves with zonal wavenumbers 4, 5 and 6.

The detection algorithm loops through each time and longitude, searching for waveguides for $k = 5, 6, 7, 8$, as defined by the following criteria:

- Two TPs between 30N and 70N
- At least 5 degrees latitude between the TPs
- Zonal wind > 0.5 within waveguide

If a waveguide is detected, the algorithm saves the following data: date, longitude and latitude of turning points. The algorithm is capable of detecting multiple simultaneous waveguides at a given longitude (at separate latitudes).

The theory of wave propagation based on the stationary wavenumber assumes that the stationary wavenumber is calculated on the background flow, i.e. the flow in which the wave is travelling, and does not include anomalies associated with the wave itself. To fulfil this requirement we process daily zonal wind data by taking a Fast Fourier Transform (FFT) in longitude and reconstructing the field retaining only wavenumbers 3 or lower. In addition, a 10 day low pass Lanczos filter smooths the data in time. The stationary wavenumber is calculated on this processed data, before the waveguide detection algorithm is applied.

Output from the waveguide detection algorithm is validated by manually checking the waveguides for ERA-interim data for dates chosen at random for various seasons and years. Figure 2 provides an example of one of these checks. The red shading shows K_S, whilst the blue points mark TPs for waveguides of $k = 5$ (lightest blue) to $k = 8$ (darkest blue). Waveguides are detected as expected (a local maximum in K_S in the meridional direction), with higher wavenumbers showing narrower waveguides, consistent with expectations.

III. WAVEGUIDE STATISTICS

The stationary wavenumber is calculated on daily data from both the ECMWF ERA-interim re-analysis [11] and the NCEP-DOE AMIP-II reanalysis (R2) [12], and waveguides are detected from 1980-2015.

The seasonal cycle of waveguide frequency for $k = 5, 6, 7$ and 8 is shown in figure 3. During the summer months a waveguide for wavenumber $k = 7$ exists in the zonal mean flow approximately 3% of the time (although this value is sensitive to the minimum waveguide width criterion and the resolution of the dataset). Thus the $k = 7$ waveguide observed by Petoukhov et al. (2016) [5] during the 2003 European heatwave, for example, was rare, but not as rare as the heatwave event itself (QSW activity over Europe exceeded the 99th percentile during August 2003 [6]). Figure 3 also shows the strong seasonal cycle in waveguide frequency, with waveguides much less frequent during the winter months. This is consistent with stronger values of K_S in summer waveguides than in winter [8]. As u appears in the dominator of the equation for K_S, the strongest jets do not necessarily produce the strongest waveguides: weaker jets with sharp gradients (i.e. narrow jets) are more likely to produce waveguides for high k. The frequency of detected waveguides in winter will also be reduced by the criterion that both turning points must occur between 30 and 70N; the upper level jet in winter is often centered around 30N,

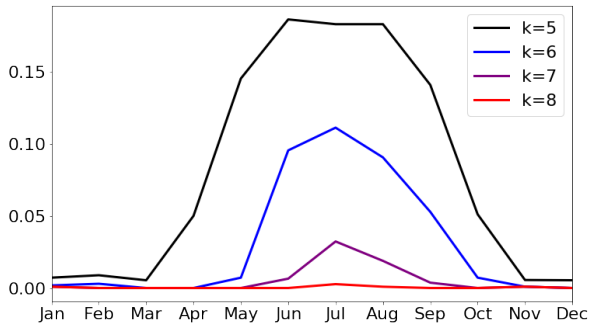


Fig. 3. Seasonal cycle of the frequency of at least one waveguide for zonal wavenumbers $k = 5, 6, 7, 8$ in zonal mean ERA-interim 3 degree resolution data.

and thus a southern TP may occur equatorwards of 30N. In this paper we focus on NH summer, June-August (JJA); for a focus on winter, changing the waveguide latitude criteria should be considered.

Figure 4 shows maps of the climatological waveguide frequency for $k = 6, 7, 8$ for the ERA-interim dataset at 3 degree resolution. Black contours show the climatological zonal wind at 250 mb. Waveguide frequency follows the jets, with a relatively strong longitudinal dependence.

Figure 5 shows the frequency of at least one waveguide as a function of longitude for $k = 5, 6, 7, 8$ for different re-analysis datasets and wavenumbers. Red and blue lines show the ERA-interim re-analysis dataset at 2 and 3 degree horizontal resolution respectively. The black lines show the NCEP-DOE R2 re-analysis dataset at 2.5 degree resolution. The shading shows +/- one standard deviation calculated on the seasonal values, assuming each season is independent.

The impact of resolution on the absolute values of waveguide frequency is clear, with higher resolution datasets showing a higher waveguide frequency for all k. The relative difference is larger for higher k. The presence of a waveguide requires a high K_S, which, as previously discussed, is achieved through a high second meridional gradient, $\partial^2 u/\partial y^2$, and a relatively low u. Reducing the horizontal resolution will result in smoothed meridional gradients, reducing the frequency of high k waveguides.

Other than absolute magnitude differences associated with the horizontal resolution, the different re-analysis datasets show qualitatively very similar waveguide frequency. It is clear that there are longitudinal regions where waveguides are much more frequent than others, giving support to this method of studying waveguides in zonally asymmetric flow, rather than in the zonal mean.

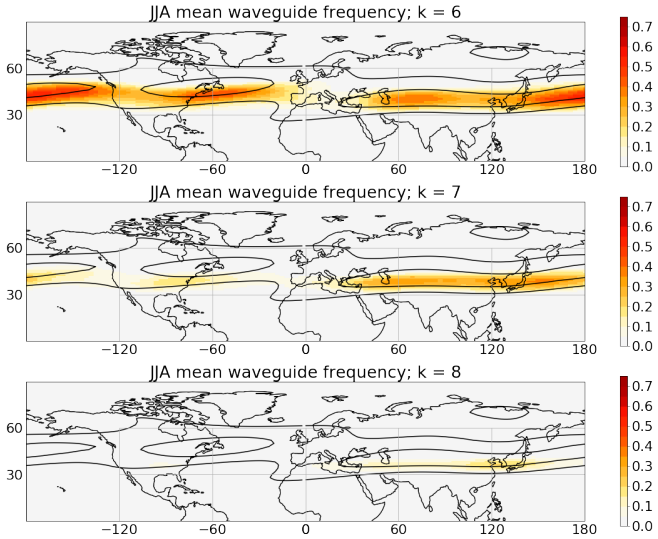


Fig. 4. Climatological JJA frequency of waveguides with $k = 6, 7, 8$ for ERA-I reanalysis at 3 degree resolution (colours). Black contours show the climatological zonal wind, with a contour interval of 10m/s.

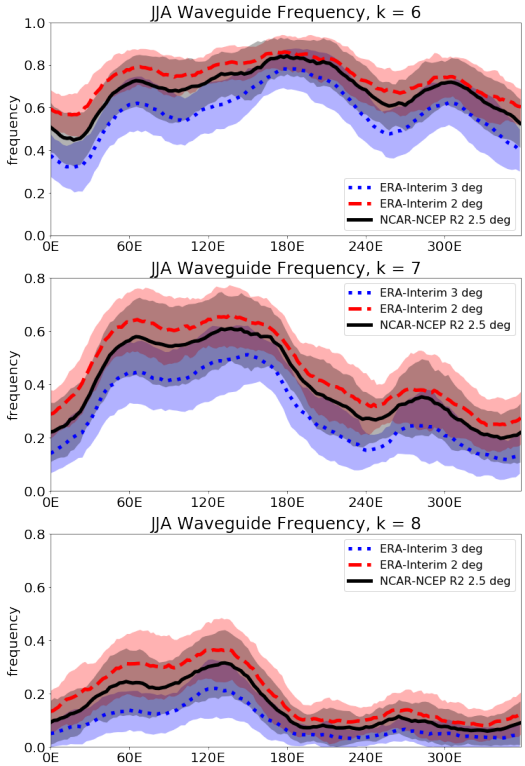


Fig. 5. Frequency of at least one waveguide occuring as a function of longitude for JJA. Red dashed (blue dotted) line shows ERA-interim re-analysis data at 2 (3) degree resolution; solid black line shows the NCEP-DOE R2 reanalysis at 2.5 degree resolution. Shading shows interannual variability: +/- one standard deviation of seasonal average values.

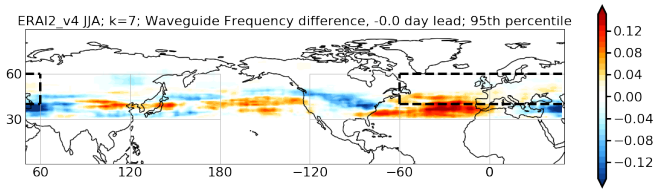


Fig. 6. Difference in $k = 7$ waveguide frequency between JJA days exceeding 95th percentile of QSW activity averaged over the black dashed box and days below 5 percentile (ERA-interim data).

IV. Waveguides and Quasi-Stationary Waves

Lastly, we show a connection between the existence of waveguides and the occurrence of high amplitude quasi-stationary waves in the ERA-interim data. Using the QSW dataset of Wolf et al. (2018) [6], downloaded from CEDA, we calculate the area-weighted mean QSW activity over a region in the North Atlantic/European sector (60W-60E, 40-60N). High QSW activity in this region is associated with JJA extreme hot surface temperatures over central Europe [6]. Anomalies from the seasonal cycle are calculated in this QSW activity, and then waveguide frequency maps are created for all JJA days when these QSW anomalies are above (below) the 95th (5th) percentile.

Figure 6 shows the difference in waveguide frequency between anomalously high (95th %ile) and low (5th %ile) QSW activity. The dashed black box shows the region over when the QSW activity anomalies are averaged. There is a clear region over the Atlantic where waveguides are more frequent when there is anomalously high QSW activity, although the increase in waveguide occurrence lies to the south of the region of anomalous QSW activity. This result is robust to changes in the resolution of the ERA-interim dataset, and to changes in the minimum waveguide width criterion.

Future work will focus on further exploring this association, including study of whether there is predictability of the quasi-stationary waves based on waveguide activity. We will also focus more closely on links between waveguide occurrence and extreme events in surface temperature. Additional work will explore the ability of climate models to reproduce waveguide statistics, taking into account the strong dependence on dataset resolution.

V. Conclusions

A waveguide detection algorithm is developed and described for objectively identifying waveguides for atmospheric planetary waves on zonally asymmetric data. Waveguides are detected in the Northern hemisphere from 1980-2015 in ERA-interim and NCEP-DOE R2 re-analysis data, with high consistency between these two datasets. The horizontal resolution of the dataset plays a strong role in the absolute frequency of waveguides of a particular k; however, the variations with longitude are similar across different datasets and resolutions. This waveguide is typically centered on approximately 40 degrees N. Waveguides over the Atlantic region are more frequent during periods of anomalously high quasi-stationary wave activity over the Atlantic/European sector; this suggests that waveguide presence may be connected with anomalously high QSW activity,

Acknowledgments

RHW received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie Grant Agreement 797961.

References

[1] J.-M. Robine, S. L. K. Cheung, S. Le Roy, H. Van Oyen, C. Griffiths, J.-P. Michel, and F. R. Herrmann, "Death toll exceeded 70,000 in europe during the summer of 2003," *C. R. Biol.*, vol. 331, pp. 171–178, Feb. 2008.

[2] D. Barriopedro, E. M. Fischer, J. Luterbacher, R. M. Trigo, and R. García-Herrera, "The hot summer of 2010: Redrawing the temperature record map of europe," *Science*, vol. 332, pp. 220–224, Apr. 2011.

[3] G. Fragkoulidis, V. Wirth, P. Bossmann, and A. H. Fink, "Linking northern hemisphere temperature extremes to rossby wave packets: Rossby wave packets and temperature extremes," *Q.J.R. Meteorol. Soc.*, vol. 144, pp. 553–566, Jan. 2018.

[4] K. Kornhuber, V. Petoukhov, S. Petri, S. Rahmstorf, and D. Coumou, "Evidence for wave resonance as a key mechanism for generating high-amplitude quasi-stationary waves in boreal summer," *Clim. Dyn.*, pp. 1–19, Nov. 2016.

[5] V. Petoukhov, S. Petri, S. Rahmstorf, D. Coumou, K. Kornhuber, and H. J. Schellnhuber, "Role of quasiresonant planetary wave dynamics in recent boreal spring-to-autumn extreme events," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, pp. 6862–6867, June 2016.

[6] G. Wolf, D. J. Brayshaw, N. P. Klingaman, and A. Czaja, "Quasi-stationary waves and their impact on european weather and extreme events," *Q.J.R. Meteorol. Soc.*, vol. 144, pp. 2431–2448, Oct. 2018.

[7] V. Petoukhov, S. Rahmstorf, S. Petri, and H. J. Schellnhuber, "Quasiresonant amplification of planetary waves and recent northern hemisphere weather extremes," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, pp. 5336–5341, Apr. 2013.

[8] B. Hoskins and T. Woollings, "Persistent extratropical regimes and climate extremes," *Curr Clim Change Rep*, vol. 1, pp. 115–124, Sept. 2015.

[9] B. J. Hoskins and D. J. Karoly, "The steady linear response of a spherical atmosphere to thermal and orographic forcing," *J. Atmos. Sci.*, vol. 38, pp. 1179–1196, June 1981.

[10] B. J. Hoskins and T. Ambrizzi, "Rossby wave propagation on a realistic longitudinally varying flow," *J. Atmos. Sci.*, vol. 50, pp. 1661–1671, June 1993.

[11] D. P. Dee, S. M. Uppala, A. J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M. A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A. C. M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A. J. Geer, L. Haimberger, S. B. Healy, H. Hersbach, E. V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A. P. McNally, B. M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, and F. Vitart, "The ERA-Interim reanalysis: configuration and performance of the data assimilation system," *Q.J.R. Meteorol. Soc.*, vol. 137, pp. 553–597, Apr. 2011.

[12] M. Kanamitsu, W. Ebisuzaki, J. Woollen, S.-K. Yang, J. J. Hnilo, M. Fiorino, and G. L. Potter, "NCEP–DOE AMIP-II reanalysis (r-2)," *Bull. Am. Meteorol. Soc.*, vol. 83, pp. 1631–1644, Nov. 2002.

INFORMATION EXCHANGE IN THE HIGH DIMENSIONAL IDEALIZED SYSTEMS AND IN CLIMATE INTER-MODEL COMPARISON

Praveen Kumar Pothapakula[1], Cristina Primo[1], Bodo Ahrens[1]

***Abstract*—We tested axiomatically proposed Transfer Entropy (TE) and the first principle based derived Information Flow (IF) on high dimensional systems which consist of common driving variables. Estimators like TE-linear, TE-karskov, TE-kernel, and IF-linear are tested for robust estimation. The TE and IF estimations not only detected reliable information exchange but also erroneous links due to common drivers. Furthermore, we tested Conditional Transfer Entropy (CTE) which removed the erroneous link in the idealized test cases. Applying the CTE to inter-model comparison, we found significant differences between the processes in coupled and uncoupled regional climate simulations on daily time scales over the Mediterranean region.**

I. MOTIVATION

Detecting and quantifying the interactions among sub-components in complex systems assists in understanding underlying dynamics. Methods from information theory are used to detect and quantify these interactions in various complex systems e.g., in the fields of neurosciences ([1]), climate sciences ([2]), earth system sciences ([3]).

Often in the climate science community, correlation analysis, empirical orthogonal functions, linear regressions, and cross-correlations are used in identifying interactions among linear systems. For non-linear interactions mutual information ([4]) is often used. However, these methods cannot distinguish between a drive and the response system. Methods based on information theory help to overcome these shortcomings and assist to distinguish between the drive and response systems. In this study we use two such methods, the axiomatically proposed Transfer Entropy (TE) [5] and Information Flow (IF) derived from the first principles of information theory [6]. The TE has been widely used in climate applications, e.g., [2] in the identification of primary drivers of recent climate variability and in the information exchange between the South Atlantic anomaly and global sea level for the last 300 years. While the TE is a very useful tool, its estimation is challenging. For an overview of practical estimations of TE and its assumptions refer to [7]. For the detection of linear interactions between the subsystems, the parametric estimation of TE is straight forward using a multivariate Gaussian model (from hereafter TE-linear). The non-parametric estimation of TE is a difficult task. Some of the common non-parametric estimation techniques of TE in the literature include the binning, kernel density and k-nearest neighbor. These estimators are highly sensitive to the parameter selection in their implementation. In this study, we applied TE-linear, TE-kernel and TE-karskov estimators [8]. The IF is another information theory based method proposed by [6]. For the non-linear implementation of IF, the time evolution of the marginal probabilities must be computed which in turn depends on the system dynamics. If the system dynamics is unknown IF becomes difficult to apply. [9] proposed a simple and concise IF formula for linear systems (from hereafter IF-linear) which is straight forward to apply without the knowledge of system dynamics. The IF-linear has been already applied in many climate applications, e.g., detecting the causal structure between CO_2 and global temperatures.

In the study by [8], various algorithms of the TE-estimation i.e., TE-linear, TE-binning, TE-kernel, TE-karskov, and IF-linear are tested with two dimensional systems along with their sensitivity on free tuning parameters and time series lengths. The study proposed a composite use of the TE-kernel and TE-karskov estimators along with parameter testing for consistent results for highly non-linear systems and in addition TE-linear, IF-linear for linear systems. The study also highlighted the possibility of erroneous information

Corresponding author: P. Praveen Kumar, pothapakula@iau.uni-frankfurt.de, [1] Institute for Atmospheric and Environmental Sciences, Goethe University Frankfurt am Main.

exchange detection due to hidden or common drivers. In this work, we extend our analysis to test the estimators from [8] for high dimensional idealized systems which consist of common drivers. Furthermore, we introduce another estimator known as Conditional Transfer Entropy (CTE) to condition on a third variable in an attempt to remove the influence of common drivers. Thereafter, we apply these methods to inter–compare our newly developed regional climate coupled model with a stand-alone regional climate model. Regional Climate Models (RCM) are used to dynamically downscale global climate models for high resolution climate information over a region of interest. In this study, we employ the RCM, Consortium for Small-scale Modeling Climate Limited -Area Modeling (COSMO-CLM) together referred to as CCLM, based on non-hydrostatic equations as an atmosphere component. For the coupled atmosphere–ocean model, CCLM is coupled to the regional ocean model Nucleus for European Modelling of the Ocean (NEMO) over the Mediterranean (NEMO-MED12) [10] over the European region (active coupling of the ocean over the Mediterranean sea only). From hereafter the coupled runs are referred to as CP and uncoupled runs are referred to as UN.

II. METHOD

In this section, we introduce the concepts of information flow and transfer entropy along with their estimation techniques.

A. Transfer Entropy

The TE measures the deviation between the transitional probabilities which represent the system dynamics,

$$\mathrm{TE}_{y->x} = \sum_{x,y} p(x_{n+1}, x_n^k, y_n^l) \log \frac{p(x_{n+1}|x_n^k, y_n^l)}{p(x_{n+1}|x_n^k)}, \tag{1}$$

where k and l are the embedding dimensions of the destination and source variables, respectively.

1) Estimation of TE-kernel: This estimator uses the box step kernel Θ with $\Theta(x > 0) = 0$ and $\Theta(x < 0) = 1$ for the estimation of relevant joint probability distributions (e.g., $\hat{p}(x, y)$, $\hat{p}(x)$ and $\hat{p}(y)$). For example, the joint probability distribution $\hat{p}(x, y)$ is calculated as

$$\hat{P}_r(x_n, y_n) = \frac{1}{N} \sum_{n'=1}^{N} \Theta(|(x_n - x_{n'}), (y_n - y_{n'})| - r),$$

where, r is the kernel width. The conditional probabilities can be similarly computed from their respective component joint probabilities. These probabilities are substituted in Eq. 1 to calculate TE. Kernel estimators are model-free (i.e., they do not assume parametric distribution). We used the same methodology to compute CTE.

2) Estimation of TE-k-nearest neighbor: This non-parametric estimator uses an adaptive binning strategy by relying on the average distances to the k-nearest neighbor data points for the calculation of TE. Transfer entropy from Y to X, $TE_{y->x}$ with embedding dimensions k=1 and l=1 is calculated as follows: for each point in the highest dimensional space i.e., $z_i = (x_{n+1}, x_n, y_n)$, its neighbors distance d=max$||z_i - zj||$ is calculated. The number of points that fall within the range d in all the marginal spaces is counted. Thereafter, the number of points in each of the marginal spaces are substituted in the equation below to calculate TE:

$$\mathrm{TE}_{y->x} = \Psi(K) + < \Psi(n_{x_n} + 1) - \Psi(n_{x_{n+1}}, n_{x_n}) - \Psi(n_{x_n}, n_{y_n}) >,$$

Where K is the number of nearest neighbors, Ψ denotes the digamma function, while the angle brackets indicate averaging over all the points, $n_{x_{n+1}}, n_{x_n}$ and n_{y_n} are the number of points that fall within the range d in the marginal spaces. This method enables for bias correction. From here after this method is referred as TE-karskov. For more details, refer to information-theoretic toolkit of [11].

3) Estimation of TE-linear : For a multivariate Gaussian model, the entropy is given as

$$H(x) = \frac{1}{2} \log((2\pi e)^d |\Omega_x|),$$

where, d is the number of dimensions, $|\Omega_x|$ is the determinant of the $d \times d$ covariance matrix $\Omega_x = \bar{x}\bar{x}^T$, and the overbar indicates averaging. Furthermore, the TE is estimated as the sums and differences of the joint entropies.

B. Liang and Kleeman Information Flow

Consider a dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(t; \mathbf{x}) + \mathbf{B}(t; \mathbf{x})\dot{\mathbf{w}},$$

where $\mathbf{x}$ and $\mathbf{F}$ are n-dimensional vector, $\mathbf{B}$ is an $n \times m$ matrix, and $\mathbf{w}$ is an m-vector of standard Wiener process ($\dot{\mathbf{w}}$ is a vector of white noise). The rate of information flow from x_2 to x_1 for the above dynamical system, when n=2 is given by

$$T_{2->1} = -E\Big[\frac{1}{\rho_1}\frac{\partial(F_1\rho_1)}{\partial x_1}\Big] + \frac{1}{2}E\Big[\frac{1}{\rho_1}\frac{\partial^2 g_{11}\rho_1}{\partial x_1^2}\Big]$$

If the system dynamics F_1 and g_{11} are independent of x_2, then $T_{2->1} = 0$, which remarkably appears in the classical formalism. For systems with the dynamics unknown, the estimation of entropy evolution is a challenge. Hence, [9] under the linear assumption proposed a simple easy-to-use formula know as maximum likelihood estimator of information flow. Given two series x_1 and x_2, for consistency with the formulae of TE mentioned above, we consider $x = x_1$ and $y = x_2$, the information flow maximum likelihood estimator or IF-linear from the system y to x is given by

$$T_{y->x} = \frac{C_{xx}C_{xy}C_{y,dx} - C_{xy}^2 C_{x,dx}}{C_{xx}^2 C_{yy} - C_{xx}C_{xy}^2},$$

where C_{xx}, C_{yy} and C_{xy} are the covariances of x and y while the subscript dx indicates time series derived from x which is formed as $\frac{x(n+k)-x(n)}{k.dt}$, with k some integer greater than or equal to 1. This easy-to-use formula bridges the gap between theory and real applications and has been successfully applied to real-world applications.

C. Conditional Transfer Entropy

The TE can also be conditioned on a third possible source. The conditioning assists to remove the influence of the third variable on the quantification of information exchange between the source and destination systems. The conditional transfer entropy is given as

$$\text{CTE}_{y->x} = I(y_n^l; x_{n+1} | x_n^k, z_n)$$

where I represents mutual information, z is the conditioning variable. Conditional transfer entropy is also known as causation entropy which is known to detect reliable information exchange in case of indirect connections and dominance of neighbors.

III. Results and Discussion

In this section, we apply the above-discussed methods to a high dimensional idealized system and in climate inter-model comparison.

A. Application to idealized systems

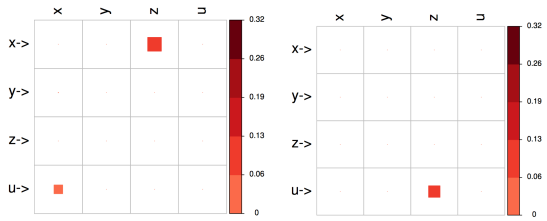
We considered a four-dimensional coupled linear system with the following governing equations

$$\begin{aligned} y_n &= \mathcal{N}(0,1) \\ u_n &= \mathcal{N}(0,1) \\ x_n &= 0.6y_{n-1} + 0.9u_{n-1} + 0.1\mathcal{N}(0,1) \\ z_n &= 0.5u_{n-2} + 0.1\mathcal{N}(0,1) \end{aligned} \tag{2}$$

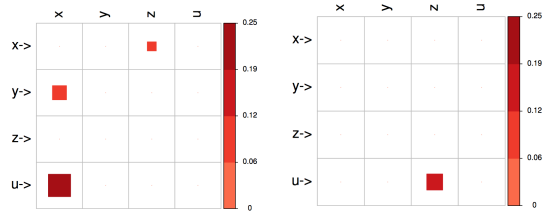
where $\mathcal{N}(0,1)$ is Gaussian noise with zero mean and unit variance. The values 0.6, 0.9 and 0.5 are the weight coefficients for y_{n-1}, u_{n-1} and u_{n-2} respectively. The system was initialized with $(x_0, y_0) = (0, 0)$. We integrated around 100000 iterations and considered the last 5000 steps for detecting and quantifying the information exchange. For the source and destination embedding dimensions for TE calculations, we chose $l = 1$ and $k = 1$. Figure 1 shows the information exchange among the variables in the coupled system with governing Eq. 2. The row variables in Fig. 1 represent the source and the column variables represent destination. The IF-linear measures a realistic information exchange from u to x at lag-1. While the information exchange between y to x is not recovered, unrealistic information exchange from x to z is seen at lag-1. At lag-2 realistic information exchange from u to z is measured. The TE-linear accurately captures the information exchange direction represented in Eq. 2 for both lag-1 and lag-2 except for the information exchange from x to z, which is unrealistic at lag-1. Similar behavior is seen for TE-karskov and TE-kernel (figure not shown) estimation. The free parameters are tuned for consistency for both TE-karskov and TE-kernel estimations. Moreover, only the significant values at 95 percentile confidence are shown in the figure. The erroneous information exchange from x to z at lag-1 is caused due to the common driver u. Hence the estimations from TE and IF-linear show that they are influenced by hidden/common drivers and thus can detect erroneous information exchange links. We also tested information exchange with different weight coefficients in Eq. 2. The results showed unreliable TE estimations at low weight coefficients (< 0.1). At high weight coefficients, the TE estimations showed realistic estimations at lag-1 and lag-2, except the erroneous link from x to z at lag-1. The IF-linear is unable to detect accurate estimations of information exchange for high as well as low weight coefficients at lag-1.

To remove the detection of erroneous information exchange links, we tested CTE estimations from TE-linear, TE-karskov, and TE-kernel. Figure 2 represents the information exchange among the variables in the coupled system with governing Eq. 2 at lag-1 (lag-2 not shown). The row variables in Fig. 2 represent the source and the column variables represent destination. The information exchange from the row to column variable was conditioned over all the other available variables at lag-1 and lag-2, and then the minimum significant information exchange value is shown. We expect that the condition removes the influence of common drivers. The CTE-linear, CTE-karskov in Fig. 2 and CTE-

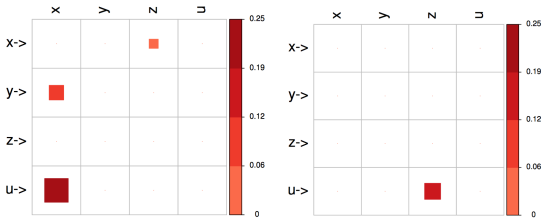
kernel (not shown) accurately captures the information exchange direction represented in Eq. 2 for lag-1. The erroneous link from x to z at lag-1 is removed due to the conditioning. Information exchange from u to z is also detected with CTE at lag-2 (figure not shown). Hence from this idealized system, we learn that CTE assists in removing the indirect influences. However, with very low values of weight coefficients, the CTE estimations showed unrealistic information exchange.



(a) IF-linear (lag 1) (b) IF-linear (lag 2)



(c) TE-linear (lag 1) (d) TE-linear (lag 2)

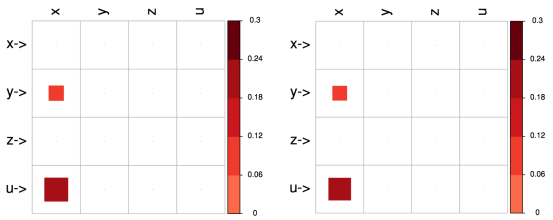


(e) TE-karskov (lag 1) (f) TE-karskov (lag 2)

Fig. 1: Information exchange in the coupled linear system (Eq. (2)) measured by (a)-(b)The IF-linear (nats/time) at lag-1 and lag-2 (c)-(e) with different variants of the TE measure (in nats).

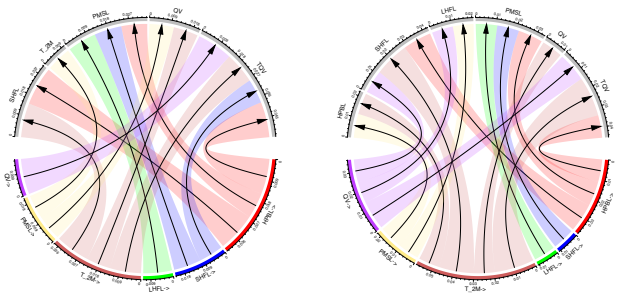
B. *Inter-comparsion of climate models*

We conducted coupled and uncoupled RCM model runs for the periods of 1979–2011 with a horizontal resolution of 0.44 degree ($\sim$ 50 km) over the European region. The initial and boundary conditions for CCLM were taken from the European Center for Medium -Range Weather Forecasts (ECWMF) ERA-Interim reanalysis data. The difference between the coupled and uncoupled run comes from the sea surface temperature values which were obtained from NEMO-



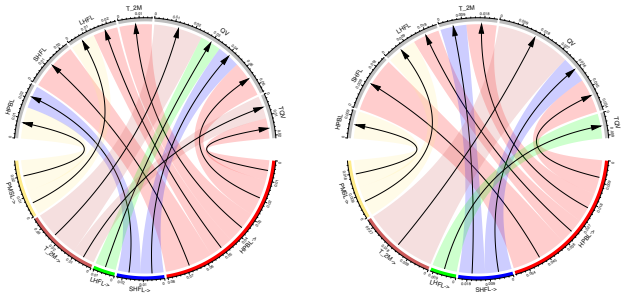
(a) CTE-linear (lag 1) (b) CTE-karskov (lag 1)

Fig. 2: Information exchange in the coupled linear system (Eq. (2)) measured by (a) The CTE-linear (nats/time) at lag-1 (b) the CTE-karskov measure (in nats) at lag-1



(a) CTE-linear (CP) (b) CTE-linear (UN)

Fig. 3: Information exchange measured by CTE-linear (nats) over Mediterranean Sea in (a) coupled model (b) uncoupled model (1980-2011).



(a) CTE-karskov (CP) (b) CTE-karskov (UN)

Fig. 4: Information exchange measured by CTE-karskov (nats) over Mediterranean Sea in (a) coupled model (b) uncoupled model (1980-2011).

MED12 for the former run and from ERA-Interim prescribed sea surface temperatures for the later over the Mediterranean sea. We aim to find if differences in the relation among the variables of interest exist between CP and UN climate simulations. For this purpose, we use CTE to remove the influence of indirect influences. We chose spatial mean of daily variables such as height of planetary boundary layer (HPBL), sensible heat flux (SHFL), latent heat flux (LHFL), 2m-temperature (T2M), near-surface specific humidity (QV), total water content (TQV) and surface pressure (PMSL) over the actively coupled Mediterranean (only ocean) region for both CP and UN simulations. We removed the trend and normalized the variables from 1980-2011 (time series length of 11680 days). Figure 3 represents the chord diagram for CTE-linear among all the variables for CP (Fig. 3a) and UN (Fig. 3b) respectively. The arrows indicate the direction of information exchange and the width represents the quantitative strength of information exchange. From Fig. 3, we see a significant difference in the relationships among variables between CP and UN simulations. This implies that processes in the CP and UN vary over the Mediterranean region on a daily scale. The CTE-karskov also shows that significant differences in the processes exist. For example, in the CP simulation the information exchange from HPBL to SHFL, LHFL, T2M, QV, and TQV exists, whereas in the UN, the information exchange from HPBL to TQV is missing. This suggests that the boundary layer in CP is coupled actively with the given variables than in the UN simulations. However, a detailed investigation of the differences in the processes is beyond the scope of the discussion in this article. Through this article, we aim to highlight the application of information theory metrics for inter-model comparison and process studies.

ACKNOWLEDGMENTS

Bodo Ahrens would like to acknowledge the support by the Senckenberg Biodiversity and Climate Research Centre (SBiK-F), Frankfurt am Main. The authors also acknowledge support by the German Federal Ministry of Education and Research (BMBF) under Grant MiKlip: FKZ01LP1518C and the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) in terms of the research group FOR 2416 Space-Time Dynamics of Extreme Floods (SPATE).

REFERENCES

[1] Vicente R, Wibral M, Lindner M, Pipa G (2011) Transfer entropy a model free measure of effective connectivity for the neurosciences. J. Comput. Neurosci 30:45–67. DOI 10.1007/s10827-010-0262-3

[2] Bhaskar A, Ramesh D.S, Vichare G, Koganti T, Gurubaran S (2017) Quantitative assessment of drivers of recent global temperature variability: an information theoretic approach. Clim Dyn 49:3877–3886.

[3] Ruddell B L and Kumar P (2009) Ecohydrologic process networks: 1. Identification. Water Resour. Res. URL https://doi.org/10.1029/2008WR007279

[4] Shannon C.E (1948) A Mathematical Theory of Communication. Bell Labs Technical Journal 27:379–423.

[5] Schreiber T (2000) Measuring information transfer. Phys. Rev 85:461–464. URL https://doi.org/10.1103/PhysRevLett.85.461

[6] Liang X S and Kleeman R (2005) Information transfer between dynamical system components. Phys. Rev. Lett 95:244101. DOI 10.1103/PhysRevLett.95.244101

[7] Runge J (2018) Causal network reconstruction from time series: From theoretical assumptions to practical estimation. Chaos 28,075310. URL https://doi.org/10.1063/1.5025050

[8] Pothapakula P K, Cristina P R and Ahrens B (2019) Quantification of information exchange in idealized and climate system applications. (submitted to Entropy)

[9] Liang X S (2014) Unraveling the cause–effect relation between time series. Phys. Rev. E 90:052150. DOI 10.1103/PhysRevE.90.052150

[10] Akhtar N, Brauch J and Ahrens B (2018) Climate modeling over the Mediterranean Sea: impact of resolution and ocean coupling. Clim Dyn 51:933948 DOI 10.1007/s00382-017-3570-8

[11] Lizier J T (2014) JIDT: an information–theoretic toolkit for studying the dynamics of complex systems.Front. Robot. AI. URL https://doi.org/10.3389/frobt.2014.00011

CLOUD CLASSIFICATION WITH UNSUPERVISED DEEP LEARNING

Takuya Kurihana[1], Ian Foster[1,2], Rebecca Willett[1,4], Sydney Jenkins[1,5], Kathryn Koenig[1,6], Ruby Werman[7], Ricardo Barros Lourenco[1], Casper Neo[1], Elisabeth Moyer[3]

***Abstract*—We present a framework for cloud characterization that leverages modern unsupervised deep learning technologies. While previous neural network-based cloud classification models have used supervised learning methods, unsupervised learning allows us to avoid restricting the model to artificial categories based on historical cloud classification schemes and enables the discovery of novel, more detailed classifications. Our framework learns cloud features directly from radiance data produced by NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) satellite instrument, deriving cloud characteristics from millions of images without relying on pre-defined cloud types during the training process. We present preliminary results showing that our method extracts physically relevant information from radiance data and produces meaningful cloud classes.**

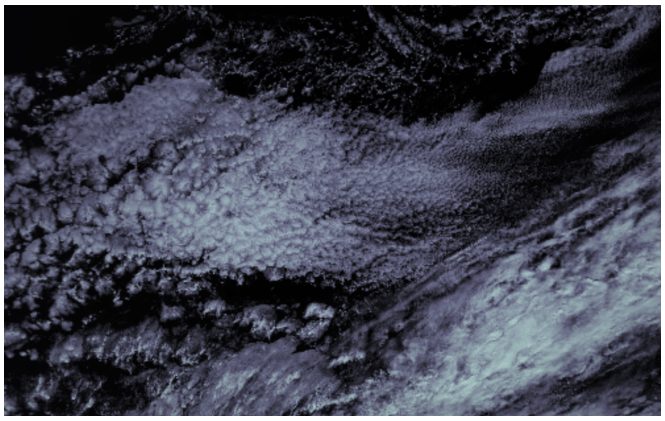


Fig. 1. MODIS Terra satellite visible imagery on December 1, 2015 over the East Pacific ocean, off the coast of California, capturing stratocumulus clouds at a variety of scales. Clouds at lower right are higher in altitude. We use this scene in all examples that follow.

I. MOTIVATION

Clouds play a dominant role in the Earth's radiation budget, both reflecting sunlight and trapping infrared radiation. Their responses are the principal source of uncertainty in numerical simulations of future climate, because even state-of-the-art climate models cannot accurately resolve cloud formation and evolution on scales from sub-kilometers to thousands of kilometers [1]. NASA satellite instruments have observed cloud behavior for several decades, providing us with a rich dataset that can potentially inform understanding of cloud dynamics and feedbacks, but these large datasets have not yet been fully employed, in part because computing power has only recently approached the necessary scale. Clouds are therefore a timely target for large scale computational analyses that can automate detection of cloud attributes and identify scientifically relevant cloud classes.

Cloud classification effectively reduces the dimensionality of information in satellite images, rendering them tractable to analysis. Attempts to use neural network methods for this purpose date to the 1990s, when Lee et al. [2] used human-labeled images to train shallow, fully connected networks to recognize stratocumulus, cumulus, and cirrus clouds. Similar efforts have continued sporadically up to the present (e.g., [3–5]), all involving supervised learning based on images labeled with historically defined classes. However, none have produced an operational tool for automated analysis of cloud images. Supervised learning generally falls short because historical classes are artificial and are well-defined only for "classic" examples that make up a small fraction of total data. Furthermore, those classes do not distinguish scales of features that in nature may vary by an order of magnitude or more. For example, in the MODIS example image of Figure 1, stratocumulus clouds show a wide range of cell sizes, and cloud textures and patterns vary in complex ways.

Unsupervised learning methods may be a more ap-

Corresponding author: I Foster, foster@uchicago.edu
[1]Department of Computer Science, University of Chicago
[2]Data Science and Learning Division, Argonne National Lab
[4]Department of Statistics, University of Chicago
[5]Department of Physics, University of Chicago
[6]Harris School of Public Policy, University of Chicago
[7]College of Letters & Science, University of California, Berkeley
[3]Department of the Geophysical Sciences, University of Chicago

propriate means of making use of the complex information in large multi-spectral satellite datasets. Such methods allow novel data-driven cloud types to emerge from imagery data, and in principle can track changes in cloud textures and patterns over time, by identifying both changing frequencies of individual cloud classes and evolving characteristics within a class.

A scientifically useful operational classification system would:

1) produce *physically reasonable* classes with scientifically relevant distinctions
2) capture information on cloud *spatial distributions*, i.e., be not reproducible using only mean properties over the target area
3) produce classes that in high-dimensional space are *cohesive* within each class but *separated* between classes
4) be *rotationally invariant*, i.e., insensitive to the orientation of an image
5) be *stable*, i.e., produce similar or identical classes when different subsets of the data are used.

We describe here the construction of a prototype data-driven workflow for cloud classification based on unsupervised deep learning. In the following, we introduce our model architecture and clustering procedure, apply it to images from the MODIS satellite instrument, and evaluate its ability to meet some of the key criteria listed above.

II. METHOD

A. Model Architecture

We leverage recent work in self-supervised learning, in which an encoder-decoder network is trained to recover an input image. We use a deep convolutional autoencoder [6] to obtain dimensionally reduced information from input data. Autoencoders have been widely used to retrieve dimensionally reduced information from high-dimensional input data. The resulting lower-dimensional latent representations incorporate important input features, simplifying the classification task. In the general framework of an autoencoder, the learning process minimizes the loss function L:

$$\min L(x) = \min ||x - F(x)||_p \ , \tag{1}$$

where x is the input image; $F(\cdot)$ is a function which maps the input image on the dimension-reduced representation and then reconstructs the image from the intermediate information, meaning $F(x)$ is the reconstructed input image; and $||\cdot||_p$ denotes the p-norm of the two images.

Our loss function combines four metrics: L1 and L2 loss, corresponding to $p = 1$ and $p = 2$ in Equation 1; the high frequency error norm after passing through the Sobel filter to detect edges of input clouds; and the multi-scale structure similarity index (MSSIM) [7], a multi-band version of SSIM [8], an index often used in computer vision to assess image similarity. We use the Adam optimizer [9], a combination of RMSprop and stochastic gradient descent with momentum, to optimize our loss function, with a learning rate of 10^{-4}.

We also include a convolution layer in our model in order to preserve spatial structure of the input image. The convolution operation implements a small-size filter to subset the entire image iteratively, with specified stride and width. The filter kernel operation extracts local features and parses the activation layer. The convolutional layer with activation function is described as

$$h^l = f\left(\sum_i \sum_j x^l_{(i+w-1)(j+s-1)} \otimes W^l_{ij} + b^l\right) \tag{2}$$

where h^l is the lth layer's latent representation; f is a nonlinear activation function; $x_{(i+w)(j+s)}$ is a $w \times s$ domain for the convolutional filter; W_{ij} is the weight at the ith column and jth row; $\otimes$ is the convolutional operation; and b denotes the bias. We set the filter size to 3×3 and use Leaky Rectified Linear Unit (Leaky ReLU) as the activation function $f(x) = \max(0.3x, x)$, as that performs better than common ReLU. Additionally, we build a residual connection every two convolutional layers to improve network performance, and add batch normalization after each residual connection. Between residual blocks, the size of an input image is scaled by a factor of two. In the encoder, the width and the stride are halved at each block, while the depth is doubled. In the decoder, these transformations are reversed, with the minor modification that we apply a transposed convolution kernel to map each input pixel to 3×3 pixels for up-sampling. The overall model architecture is illustrated in Figure 2.

We implement the convolutional autoencoder in the TensorFlow deep learning library [10] and use the Horovod framework [11] for data parallelization. Our encoder-decoder architecture stacks 20 convolutional layers and has 297 232 trainable parameters, and our latent representation has size 8 × 8 × 128. Training took 100 000 steps and 17 hours to converge the loss function on four NVIDIA K80 GPUs on the University of Chicago's Midway compute cluster. We chose a batch size of 32 in accordance with common deep

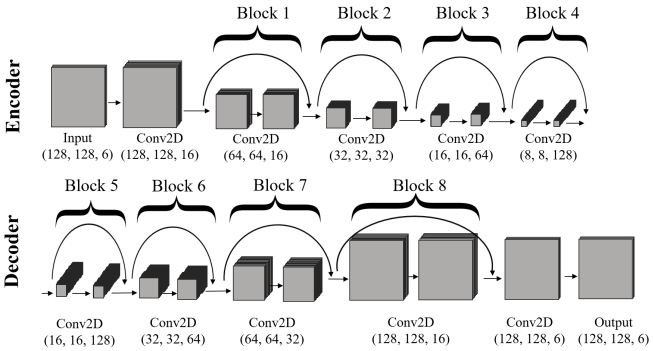


Fig. 2. Our autoencoder architecture. The encoder consists of four blocks, each with two convolutional layers activated by Leaky ReLU with residual connections; the decoder has the mirror structure of the encoder. The arrows represent the flow of input images, and each bracketed triple gives the height, width, and channel dimensions of the layer(s) above.

learning parameter settings and processed 789 GB of training images. We describe the input data and its preprocessing below.

B. Dataset

We test our workflow to perform cloud feature extraction on multi-spectral data from the MODIS instrument. Our input data is MODIS level 1B calibrated radiance imagery at 1 km resolution (MOD021KM; hereafter MOD02). This product has 36 spectral bands, from the visible to the thermal infrared; we work with bands 6, 7, 20, 28, 29, and 31, as these are the most important for the MODIS06 Level 2 algorithms that characterize cloud properties [12, 13]. Bands 6, 7, and 20 (1.6 , 2.1 , and 3.7 µm) are encoded in the algorithm to estimate cloud optical properties (e.g., optical thickness and effective radius), and the brightness temperatures at bands 28, 29, and 31 (7.3 , 8.5 , and 11 µm) are used in the separation of high and low clouds and the detection of cloud phase. Note that because we seek to discover aspects of these physics variables in our classification, we do not use derived properties such as brightness temperature, but instead input radiance data directly.

To enable efficient learning of cloud features, we define the unit of our input data, a "patch," as a small subset of a typical MODIS image: 128 km × 128 km × 6 selected bands, out of an image of 2030 km × 1354 km × 36 bands. To select input patches that contain clouds, we align the MOD02 data to its corresponding MODIS35 Level 2 cloud flag product ("MOD35"), and define a patch to be valid if more than 30% of the patch is comprised of cloud pixels as detected by MOD35. We then train the network using ∼1.01 million patches: about 1% of the full 19-year dataset from a single MODIS satellite instrument.

C. Clustering

We use hierarchical agglomerative clustering (HAC) to merge data points by minimizing cluster variance, thus building a tree structure during the merging process. We choose HAC because it exhibits greater stability with respect to initialization than does k-means clustering. Our linkage metric is Ward's method, formulated as following

$$\delta\text{dist}(X_A, X_B) = \frac{n_A n_B}{n_A + n_B} ||C_A - C_B||^2 \ , \qquad (3)$$

where the distance between two clusters X_A and X_B is evaluated as the squared distance between the centroids of merged clusters C_A and C_B weighted by the number of patches in these clusters n_A and n_B.

To choose the number of clusters for an analysis, we would ideally determine the number for which clustering results are stable (allowing the permutation of clustering categories). As an approximate measure of stability, we measure the similarity of clusters using the Adjusted Mutual Information score (AMI). We first obtain pseudo ground-truth labels by applying clustering to ∼320,000 patches, and then conduct tests with varying subsets of patches (chosen at random from the full dataset) and varying numbers of clusters, and compute the AMI score between the ground truth labels and subsets. The AMI score typically stabilizes at cluster numbers of about 10 and higher. Most demonstration analyses shown below use 12 clusters; a larger number is likely desirable for eventual science use.

Fig. 3 shows the complete pipeline from input data to resulting clusters.

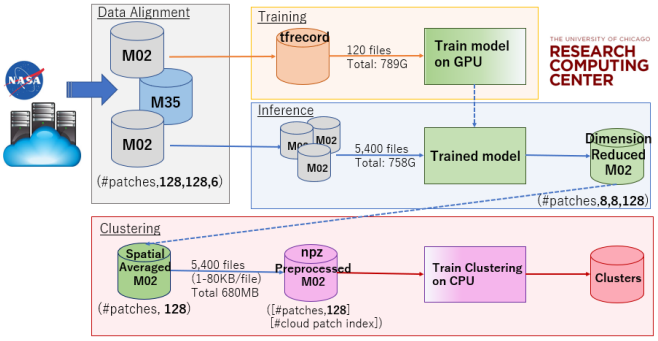


Fig. 3. Cartoon of the framework used in this work. M02 and M35 are the MOD02 and MOD35 satellite data products used as inputs. Orange and blue arrows show the paths taken by the training and test data, respectively; the red arrow depicts the clustering process.

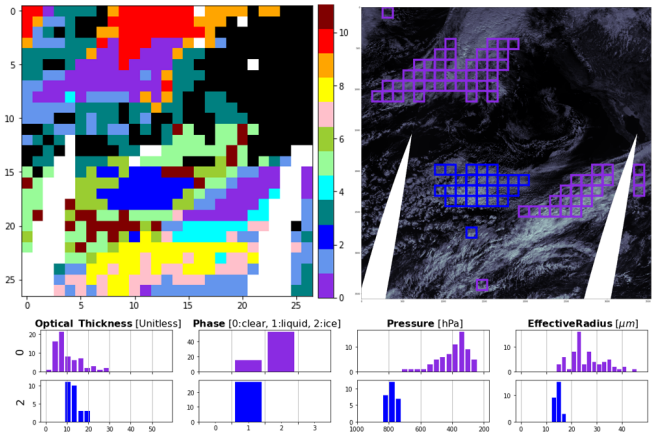


Fig. 4. Results of clustering results via autoencoder of MOD02 level 1B radiances in a representative MODIS swath image (December 1st, 2015, 11-44 N, 144-112 E; center of scene is shown in Fig. 1). Orbital coverage gaps leave missing data on sides of swath. Top left: labeled patches, classified into 12 clusters. Color bar shows cluster number in the 491 patches used; white indicates no data or invalid data; black indicates patches with <30% cloud pixels. Top right: raw visible image (band 1, which is not used as input to the autoencoder), with clusters #0 (violet) and #2 (blue) highlighted. Bottom: histograms of path-mean values of four derived cloud physics parameters in clusters #0 and #2: optical thickness, phase, cloud top pressure, and effective radius. Cluster #2 captures stratocumulus and Cluster #0 two instances of high-altitude cirrus.

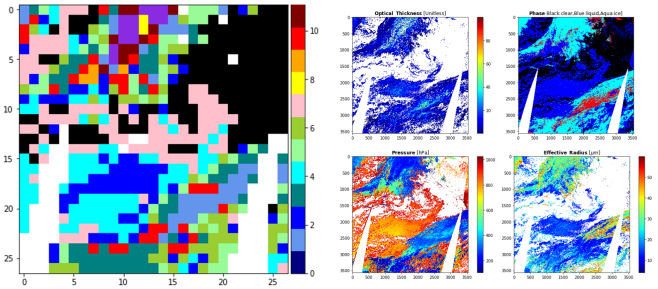


Fig. 5. Results of clustering based on patch-mean values of five MOD06 parameters (the 4 shown in Fig. 5 plus cloud water path) in the same swath as in Fig. 4. Left: labeled patches, classified into 12 clusters, with same figure conventions as in Fig. 5. Right: spatial distribution (heat maps) of values of four MOD06 cloud physics parameters. Clustering based on patch-mean parameter value produces less spatially coherent assigned classes than in Fig. 4 and does not capture important physical gradients, e.g. the sharp transition in effective radius at lower right.

III. Evaluation

We report on three initial evaluations of our framework's capabilities.

As a first, we evaluate the physical reasonableness of assigned cluster labels in our full workflow. That is, we ask whether clusters are associated with reasonable distributions of patch-mean values of physical variables. Fig. 4 shows results for the representative MODIS swath also shown in Figure. 1. Left panel shows the cluster labels assigned to each patch in the image; right panel shows the raw visible image (band 1) and highlights patches assigned to two selected clusters (#0 and #2); and bottom panels show the distributions for these patches of four derived cloud physics parameters. Clustering is clearly correlated with meaningful physical cloud attributes: cluster #2 (blue) is stratocumulus and cluster #0 is cirrus, likely convective outflow.

We then conduct a simple test of whether our clustering via autoencoder captures richer and more meaningful information on cloud distributions and properties than can be provided by the deterministic algorithms used to produce derived cloud parameters. To have scientific value, our framework must produce information beyond that encoded in MOD06 products. We apply agglomerative clustering directly to MOD06 physics parameters and evaluate how well patches are classified without the guidance of dimension-reduced MOD02 radiance information. Results suggest that clustering via autoencoder produces classes that are spatially more cohesive and that better capture important physical transitions. (Compare left panels of Figures 4 and 5.)

Finally, we examine the spatial distribution of the latent representation itself using t-Distributed Stochastic Neighbor Embedding (t-SNE). This nonlinear dimensionality reduction technique maps each point in a high-dimensional space to a two-dimensional point such that similar objects are placed near to each other and dissimilar objects far apart, with high probability [14]. Resulting patch clusters are cohesive and distinct, suggesting that agglomerative clustering within our latent representation meaningfully separates different patch types (Figure 6).

IV. Conclusions

We describe here a prototype application of unsupervised learning to the problem of automated classification of clouds in multi-spectral satellite imagery. Our convolutional autoencoder generates a latent representation that, when clustered, yields physically meaningful cloud classes that pass a number of requisite tests for a scientifically useful tool. Assigned classes appropriately produce spatially coherent classifications, and capture meaningful aspects of cloud physics without being reproducible from mean values of physics parameters alone. This work supports the possibility of using unsupervised data-driven frameworks for automated cloud classification and pattern discovery without requiring the prior hypothesis of ground-truth labeled data. While results here are preliminary, they suggest that similar

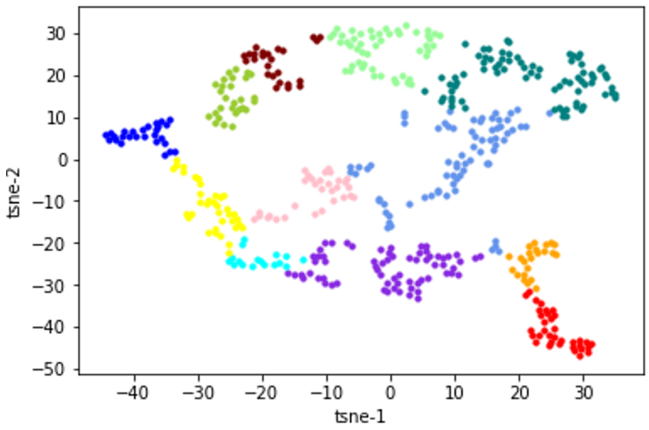


Fig. 6. A t-SNE visualization of the latent representations of the MOD02 patches of Figure 4, with cluster assignments represented by the same colors as in that figure. Patches in each cluster are projected near to each other in the t-SNE map.

frameworks can be used to analyze multi-year, global data to track short-term evolution during cloud lifecycles and long-term trends in the distribution of cloud features and characteristics.

More generally, unsupervised learning methods have broad potential applicability in the Earth sciences. In the satellite era, a primary challenge to environmental scientists is not gathering data but finding meaning in overwhelmingly large datasets. Unsupervised learning has the potential to reveal patterns directly learned from observation, which can then provide new insights and help diagnose drivers of system behavior.

ACKNOWLEDGMENTS

This work was supported by the Center for Robust Decision-making on Climate and Energy Policy (RDCEP), NSF award SES-1463644, and used computers at the U.Chicago Research Computing Center and the Argonne Leadership Computing Facility, a DOE Office of Science User Facility, contract DE-AC02-06CH11357.

REFERENCES

[1] M. Yoshiaki, K. Yoshiyuki, Y. Ryuj, Y. Tsuyoshi, Y. Hisashi, and T. Hirofumi, "Deep moist atmospheric convection in a subkilometer global simulation," *Geophysical Research Letters*, vol. 40, no. 18, pp. 4922–4926, 2013.

[2] L. R. Jonathan, C. Weger, K. S. Sailes, and M. W. Ronaldo, "A neural network approach to cloud classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 28, no. 5, pp. 846–855, 1990.

[3] T. Bin, A. S. Mukhtiar, R. A. S. Mahmood, H. V. H. Thomas, and L. R. Donald, "A study of cloud classification with neural networks using spectral and textural features," *IEEE Transactions on Neural Networks*, vol. 10, no. 1, pp. 846–855, 1999.

[4] W. Robert and L. H. Dennis, "Spatial variability of liquid water path in marine low cloud: The importance of mesoscale cellular convection," *Journal of Climate*, vol. 19, no. 9, pp. 1748–1764, 2005.

[5] Z. Jinglin, L. Pu, Z. Feng, and S. Qianqian, "CloudNet: Ground-based cloud classification with deep convolutional neural network," *Geophysical Research Letters*, vol. 45, no. 16, pp. 8665–8672, 2018.

[6] G. E. Hinton and S. Z. Richard, "Autoencoders, minimum description length and Helmholtz free energy," in *Advances in Neural Information Processing Systems 6*, pp. 3–10, Morgan-Kaufmann, 1994.

[7] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *37th Asilomar Conference on Signals, Systems & Computers*, vol. 2, pp. 1398–1402, Ieee, 2003.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[9] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2015.

[10] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, *et al.*, "Tensorflow: A system for large-scale machine learning," in *12th USENIX Symposium on Operating Systems Design and Implementation*, pp. 265–283, 2016.

[11] A. Sergeev and M. Del Balso, "Horovod: Fast and easy distributed deep learning in Tensorflow," *arXiv preprint arXiv:1802.05799*, 2018.

[12] S. Platnick, K. G. Meyer, M. D. King, B. Marchan, T. G. Arnold, Z. Zhang, P. A. Hubanks, R. E. Holz, P. Yang, W. L. Ridgway, and J. Riedi, "The MODIS cloud optical and microphysical products: Collection 6 updates and examples from Terra and Aqua," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 1, pp. 502–525, 2017.

[13] B. A. Baum, P. W. Menzel, R. A.Frey, D. C. Tobin, R. E. Holz, S. A. Ackerman, A. K. Heidinger, and P. Yang, "MODIS cloud-top property refinements for collection 6," *Journal of Applied Meteorology and Climatology*, vol. 51, no. 6, pp. 1145–1163, 2012.

[14] L. van der Maaten and G. Hinton, "Visualizing high-dimensional data using t-SNE," *Journal of Machine Learning Research*, no. 9, pp. 2579–2605, 2008.

DATA-DRIVEN TEMPORAL ATTRIBUTION DISCOVERY OF TEMPERATURE DYNAMICS BASED ON ATTENTION NETWORKS

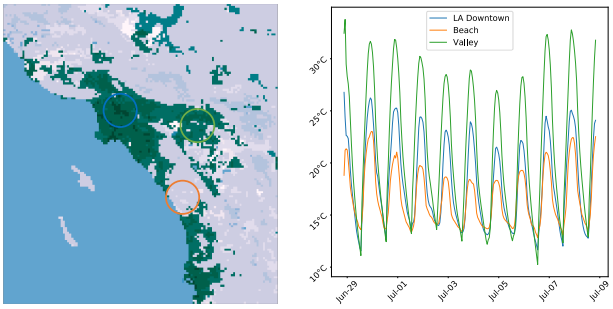
Sungyong Seo[1], Jiachen Zhang[2], George Ban-Weiss[2], Yan Liu[1]

***Abstract*—Our goal is to discover attributions leading temperature variations from climate observations that can be extended to understanding of urban heat island effect (UHIE), which is particularly problematic along with the rapid expansion of urban regions recently. We use a deep learning model formed with attention and convolutional networks to extract possible attributions and time lag mainly affecting temperature dynamics. Our work focuses on the Southern California area, specifically three meteorologically different regions to verify if different attributions are associated with the temperature observations. Furthermore, we interpret the results and provide the prediction error to support the necessity of the discovering module.**

I. MOTIVATION

The urban heat island effect (UHIE) is a phenomenon where urban areas have higher surface and atmospheric temperatures than surrounding suburban and rural regions. Several environmental processes drive the UHIE: (1) anthropogenic heat release from high solar absorbing materials (e.g., concrete roof or asphalt road); (2) lower evapotranspiration due to massive use of impervious surfaces; (3) geometry of urban canyons that traps air and alters wind flows [1]. These higher temperatures lead to unfavorable impacts such as increases in building cooling energy use [2] during summer, and have possibility to exacerbate urban air pollution [3], [4], human thermal comfort (e.g., high nocturnal temperature causes insomnia and dehydration) [5], and heat waves.

To mitigate the UHIE, many strategies have been suggested to decrease the locally concentrated heat and temperatures. For example, increasing vegetation fraction (e.g., planting trees, green roofs) could mitigate the higher urban temperature by increasing the amount of evaporation [6]. Painting rooftops white is another common strategy to reduce the heat by allowing high solar reflectance [7], [8], [9].



(a) (b)

Fig. 1: (a) Land type of Southern California. Greenish areas are urban areas and whitish areas are mostly grassland. Blue, orange, and green circle areas correspond to Downtown Los Angeles, beach, and valley regions, respectively. (b) Temperature variations in the different regions.

While previous studies examined the causal relationship between the UHIE and urban features (e.g., albedo, vegetation fraction, and impervious rate), these features are mostly static. Therefore, they do not take into consideration how time-varying observations (e.g., humidity, wind flows) affect the temperature dynamics, which is required to understand the UHIE. Since the temperature dynamics is not a simple function of static features but a result of complicated interactions of all climate-related factors, it is particularly important to discover important attributions from other meteorological observations to understand latent physical processes leading temperature variations. In addition, discovering the temporal attributions allows us to understand different meteorological characteristics of different regions. For example, while Downtown Los Angeles (Blue circle in Figure 1a) has much higher impervious rate and lower vegetation fraction compared to Valley district

Corresponding author: S Seo, sungyons@usc.edu [1]Department of Computer Science, University of Southern California [2]Department of Civil and Environmental Engineering, University of Southern California

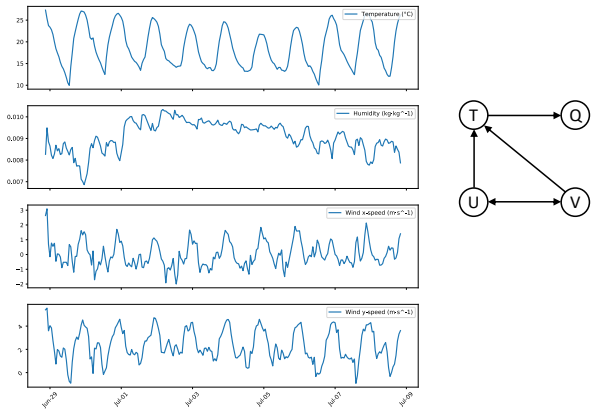


Fig. 2: (left) Examples of climate observations and (right) latent relationships of the time series (T: 2m temperature, Q: 2m specific humidity, U: Surface wind speed (x-component), V: Surface wind speed (y-component))

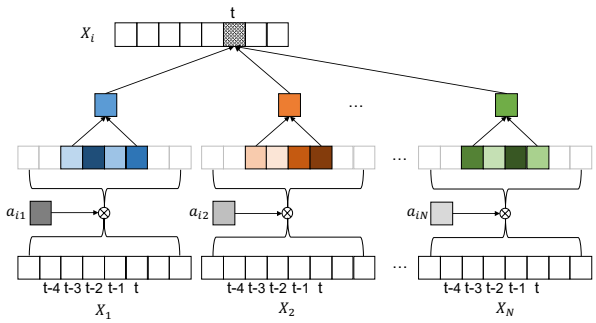


Fig. 3: The architecture of the proposed temporal attributions discovery model. The colored and gray-scale squares denote the CNN kernel weights (size 4) and attention weights, respectively. By extracting the highest weights in a_{ij} and the convolutional kernel, the attributions and time lag are obtained.

(Green circle in Figure 1a), the Downtown area shows much cooler diurnal temperature due to the wind flow from Pacific ocean (Figure 1b). Thus, in addition to the static features that contribute to the UHIE, time-varying meteorological features such as wind speed also play an important role in determining urban temperatures. Moreover, these time-varying features can influence the effectiveness of heat mitigation strategies that are designed to modify the aforementioned static features. For example, increasing vegetation fraction in the Downtown and the Valley district can lead to different temperature reductions due to the differences in their meteorological conditions.

In this work, we provide a framework providing potential temporal attributions in climate observations (temperature, humidity, wind flow, etc.) to figure out which observations likely lead temperature variations in different areas (Figure 2). We, then, demonstrate that how the attributions are different over the different regions and the results supporting the necessity of the customized mitigation strategies for different meteorological environments.

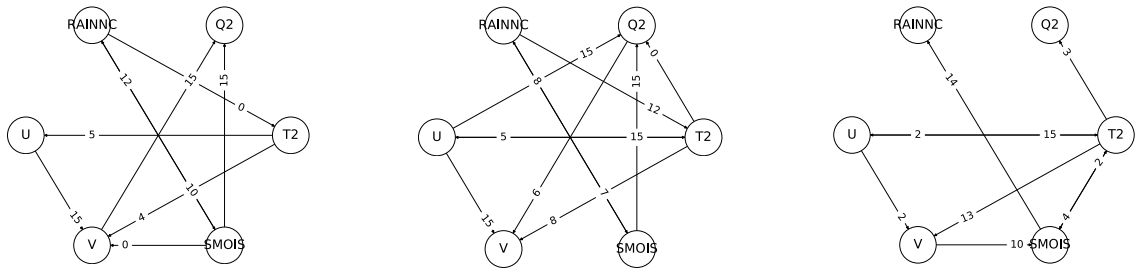
II. METHOD

While most machine learning methods aim for curve fitting based on the correlations between input and output ($p(y|x)$), attribution (or important features) discovery is another branch which deep neural networks are targeting recently. Considering significant attributions is particularly important for climate modeling because a predictive model based on correlations only has difficulty in extracting robust relationships, and thus, it can be interpreted as incorrect *cause and effect* relations, which potentially lead ineffective strategies or solutions. In this work, we modify temporal causal discovery framework proposed in [10] for modeling climate observations. The model mainly consists of two parts: attention and convolutional networks.

a) Attention neural network: Attention mechanism is one of the examples showing the representational power of deep learning. Specifically, the attention method quantifies a *relation* between two objects. For example, in language translation [11], the objects correspond to the source and target words and attention weights (or scores) are computed by the similarity of two words. In other words, if two words are strongly associated with each other, the attention score will be higher than that of two independent words. This method can be used to find a set of two objects that are mutually dependent on each other. Note that the attention weight does not provide any further information to distinguish whether two objects has causality or correlation. Potential causality can be discovered with *temporal precedence* assumption. Since the observations recorded at t, $x(t)$, cannot affect past observations, $x(t-d)$ where $d > 0$, if a model provides a significant attention score between the two observations measured in different time, the preceding one likely leads the following observation.

$$\hat{X}_i(t) = f(\{a_{ij}X_j(\tau < t)|j \in \mathcal{N} \text{ and } j \neq i\}) \quad (1)$$

where $\mathcal{N}$ is a set of given time series and $f(\cdot)$ can be neural networks to predict X_i, which is a target series. a_{ij} is the attention coefficient between two series X_i and X_j, and it provides how much the j-th observation, X_j, is important to predict the target X_i.



(a) Downtown Los Angeles (b) Beach (c) Valley

Fig. 4: Extracted attributions over 3 different regions. The nodes denote different time series and the arrows demonstrate the possible cause and effect. The edge numbers are time lag of a source observation leading a target observation.

b) Convolutional neural network: While Equation 1 is able to discover which series is more important for predicting the target series, it is neither very efficient when the length of time series is very long nor able to discover a time lag between the series. First, fully connected networks (multilayer perceptron, MLP) are hard to handle input sequence whose length is varying. Second, if recurrent neural networks (RNN) are considered, vanishing gradients problem is also a bottleneck to handle long-term dependency. Instead, convolutional neural networks (CNN) can be an alternative to discover important attributions between fixed length past observations and a target.

$$\hat{X}_i(t) = f(\{a_{ij}X_j(t-d \leq \tau < t) | j \in \mathcal{N} \text{ and } j \neq i\}) \quad (2)$$

where d is a sliding kernel size. CNN has many advantages for time series prediction; (1) parallel computation which allows faster training than RNN (2) easily increase the receptive field with dilation [12] (3) discovery the number of time steps between a cause and its effect by interpreting the kernel weights. The model we used in this work is similar to one proposed in [10], however, we exclude the target series in the input set to dig out important attributions among different time series, which might be occluded by the self-causality. All parameters involved in this model are trained through the minimization of the time series prediction error, $||X_i(t) - \hat{X}_i(t)||$ over $t \in T$ and $i \in \mathcal{N}$ where T is the total length of given series.

III. EXPERIMENTAL SETTINGS

For dataset, we use the simulated meteorological data on the Southern California area (Figure 1a) from June/28/2012 to July/14/2012 and the data are recorded hourly. Among a number of observations we use *Temperature* (T2), *Humidity* (Q2), *Wind speed (x, y direction)* (U and V), *Precipitation* (RAINNC), *Soil moisture* (SMOIS) as source and target time series. As Figure 1b shows, different meteorological environments lead totally different physical processes and thus, the dynamics of the observations and attributions will be different as well. To verify if the proposed model is able to provide the different attributions, we split the data into 3 different regions: (1) Downtown Los Angeles (2) Beach, and (3) Valley (See Figure 1a where each region is located.).

To train the model, we first choose one time series (target series) from the input set and use other series to predict the target series. The size of the convolutional kernel is 4 and we considered one hidden layer to increase the length of the receptive field. Once the training is done, we sort the attention scores to discover the important attributions of the target series and find where the maximum kernel weight is for the time lag.

TABLE I: Attributions on temperature variations.

Region	Attributions
Downtown	Precipitation (0)
Beach	Wind x-speed (15), Precipitation (12)
Valley	Wind x-speed (15), Soil moisture (2)

IV. RESULTS AND DISCUSSION

Figure 4 shows the discovered attributions among the given time series. While there are some common relations, the attribution graphs on the different environments are fairly different as we expected previously. Here we focus on the attributions leading temperature variations that can affect the UHIE and the result is summarized in Table I. Interestingly, the temperature in Downtown is not strongly dependent on wind flows and

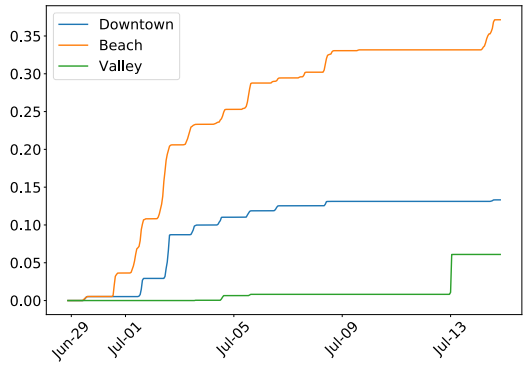


Fig. 5: Precipitation over the different regions.

TABLE II: Prediction error (MSE)

Region	with Attention	without Attention
Downtown	0.058940	0.063173
Beach	0.032778	0.037150
Valley	0.083281	0.093732

it is due to an urban canyon effect where the trapped temperature is not easily dissipated through the wind flow. Instead, the temperature dynamics is instantly susceptible to the external cooling factor, precipitation. Since there is little rainfall during the summer season, it implies that there is no distinct attribution from the input series directly changing T2 in Downtown. On the other hand, temperature in beach is relatively easy to be varied by the wind from Pacific ocean (lateral direction). Unlike the other two regions, Valley temperature observations are susceptible on soil moisture and it can be attributed to the precipitation of Valley which is much less than that of the other regions.

To verify if the module for discovering temporal attributions is effectively helpful for time series prediction, we provide the mean square error (MSE) of the model with or without the attention networks. Table II shows that the attention module is able to not only extract important attributions but also reduce the regression error.

V. CONCLUSION AND FUTURE WORK

Attribution discovery of the temperature dynamics is particularly important for understanding the UHIE. In this work, we propose a framework based on attention and convolutional neural networks to figure out the possible causes affecting temperature variations on different environments. Experimental results support our assumption and they help us to interpret climate observations with the data-driven model. For the future work, it is required to justify if the proposed attributions are actual causes under a more robust causal discovery frameworks and furthermore, rigorous interpretation should be studied.

REFERENCES

[1] P. Vahmani, F. Sun, A. Hall, and G. Ban-Weiss, "Investigating the climate impacts of urbanization and the potential for cool roofs to counter future climate change in southern california," *Environmental Research Letters*, vol. 11, no. 12, p. 124027, 2016.

[2] M. Kolokotroni, I. Giannitsaris, and R. Watkins, "The effect of the london urban heat island on building summer cooling demand and night ventilation strategies," *Solar Energy*, vol. 80, no. 4, pp. 383–392, 2006.

[3] W. Tao, J. Liu, G. A. Ban-Weiss, L. Zhang, J. Zhang, K. Yi, and S. Tao, "Potential impacts of urban land expansion on asian airborne pollutant outflows," *Journal of Geophysical Research: Atmospheres*, vol. 122, no. 14, pp. 7646–7663, 2017.

[4] W. Tao, J. Liu, G. Ban-Weiss, D. Hauglustaine, L. Zhang, Q. Zhang, Y. Cheng, Y. Yu, and S. Tao, "Effects of urban land expansion on the regional meteorology and air quality of eastern china.," *Atmospheric Chemistry & Physics*, vol. 15, no. 15, 2015.

[5] M. A. Palecki, S. A. Changnon, and K. E. Kunkel, "The nature and impacts of the july 1999 heat wave in the midwestern united states: learning from the lessons of 1995," *Bulletin of the American Meteorological Society*, vol. 82, no. 7, pp. 1353–1368, 2001.

[6] D. Li, E. Bou-Zeid, and M. Oppenheimer, "The effectiveness of cool and green roofs as urban heat island mitigation strategies," *Environmental Research Letters*, vol. 9, no. 5, p. 055002, 2014.

[7] M. Zinzi and S. Agnoli, "Cool and green roofs. an energy and comfort comparison between passive cooling and mitigation urban heat island techniques for residential buildings in the mediterranean region," *Energy and Buildings*, vol. 55, pp. 66–76, 2012.

[8] A. Mohegh, P. Rosado, L. Jin, D. Millstein, R. Levinson, and G. Ban-Weiss, "Modeling the climate impacts of deploying solar reflective cool pavements in california cities," *Journal of Geophysical Research: Atmospheres*, vol. 122, no. 13, pp. 6798–6817, 2017.

[9] J. Zhang, A. Mohegh, Y. Li, R. Levinson, and G. Ban-Weiss, "Systematic comparison of the influence of cool wall versus cool roof adoption on urban climate in the los angeles basin," *Environmental science & technology*, vol. 52, no. 19, pp. 11188–11197, 2018.

[10] M. Nauta, D. Bucur, and C. Seifert, "Causal discovery with attention-based convolutional neural networks," *Machine Learning and Knowledge Extraction*, vol. 1, no. 1, pp. 312–340, 2019.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[12] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," *arXiv preprint arXiv:1511.07122*, 2015.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

EMULATING NUMERIC HYDROCLIMATE MODELS WITH PHYSICS-INFORMED CGANS

Ashray Manepalli[1], Adrian Albert[1,2,*], Alan Rhoades[2], Daniel Feldman[2], Prabhat[2]

***Abstract*—Process-based numerical simulation, including for climate modeling applications, is compute- and resource-intensive, requiring extensive customization and hand-engineering for encoding governing equations and other domain knowledge. On the other hand, modern deep learning employs a much simplified and efficient computational workflow, and has been showing impressive results across myriad applications in computational sciences. In this work, we investigate the potential of deep generative learning models, specifically conditional Generative Adversarial Networks (cGANs), to simulate the output of a physics-based model of the spatial distribution of the water content of mountain snowpack, or snow water equivalent (SWE). We show preliminary results indicating that the cGANs model is able to learn mappings between meteorological forcing (e.g., minimum and maximum temperature, wind speed, net radiation, and precipitation) and SWE output. Moreover, informing the model with simple domain-inspired physical constraints results in higher model accuracy, and lower training time. Thus Physics-Informed cGANs provide a means for fast and accurate SWE modeling that can have significant impact in a variety of applications (e.g., hydropower forecasting, agriculture, and water supply management).**

I. MOTIVATION

In many climate modeling applications, direct observation on large scales of the environmental variables of interest is challenging, requiring instead the use of computationally-expensive numerical simulation models [1]. Such numerical weather and climate models are based on a series of coupled partial differential equations (PDEs) that aim to represent the dynamics, thermodynamics, radiative, and mass-flux processes within the major components of the Earth system including the atmosphere, cryosphere, land-surface, and ocean. These PDEs are often representative of the forefront of scientific understanding - utilizing fundamental physics, hydrology, and climatology theory - but are computationally-expensive to solve, requiring the use of high-performance computing (HPC) environments and highly-specialized expertise to set up and operate [2].

[1] terrafuse, inc. [2] Berkeley National Lab
*Correspondence to: toni@terrafuse.ai

In addition, such process-based models of realistic systems often employ parametrizations to resolve subgrid processes that are generally poorly understood, making it hard to decipher model sensitivity and bias, especially when all components of the Earth system are coupled [2]. One such variable is the aforementioned SWE, requiring the use of a chain of expensive numerical models for simulation. Empirical SWE data is typically highly sparse or even completely unavailable due to its difficulty in acquisition from mountainous regions, as well as the high expense associated with maintaining measurements at adequete temporal and spatial resolutions [3]. Moreover, SWE has many important use cases across sectors of high societal impact, e.g., water supply, hydropower, and agriculture [4].

Challenges in snowpack modeling. We focus on two current shortcomings of process-based models. First, the forcing uncertainty in key meteorological variables, including precipitation amount and phase, air temperature, and humidity, is shown to be comparable to or larger than snowpack model structural uncertainty [5]. Second, snow models heavily rely on temperature dependent thresholds to determine the phasing of incident precipitation and the magnitude and duration of the cold content of snow, or the interplay between snow density, depth, and temperature prior to melt. Therefore, a key outstanding need in the community would be to test how biases in precipitation intensity, duration and frequency and phase drive divergence in the snowpack accumulation season and how biases in surface energy and mass flux drive early spring melt [6].

II. METHODS, DATA, AND MODELS

Data. For all experiments presented here we have used a reanalysis dataset developed by Livneh[8] (L15) for the California Sierra Nevada mountain range. The L15 data was originally obtained by combining hydrologic simulation runs of the Variable Infiltration Capacity (VIC) model bounded by spatially interpolated in-situ meteorological station measurements. This dataset contains meteorological data and simulated SWE, used to

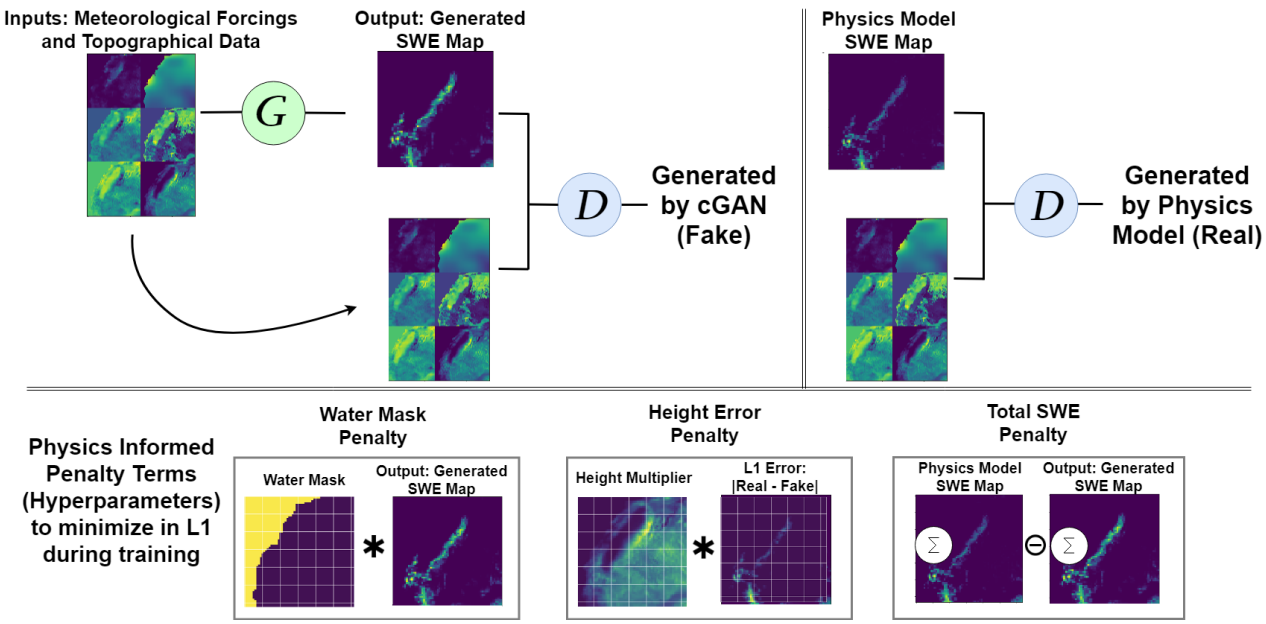


Fig. 1. Architecture and diagram of the conditional GAN used, a heavily modified variant of Pix2Pix [7]

train, assess, and constrain cGAN model. All data channels are resized and normalized from 321x321 (4km grid size) to 64x64 (17km grid size). Input channels were concatenated along a third axis, similar to the RGB format for images)

Physics-informed conditional GANs. We formulate our emulation problem as an image-to-image translation task. The goal is to transform an image from domain X, gridded meteorological variables, to domain Y, SWE grids. The pipeline of training a GAN emulator of SWE is illustrated in Figure 1. In our setting, training samples from the two domains X and Y are assumed paired, $\{(x^i, y^i)\}_{i=1}^N$ as in [7]. Here we denote by x samples from domain X and by y samples from domain Y.

We have incorporated certain domain knowledge into our model via additional penalty terms into the optimization loss function, as follows:

- Areas of higher elevation typically have larger amounts of snow (and therefore SWE), and we add penalties to large errors in such areas accordingly;
- As a significant portion of the data we study covers water areas such as the Pacific Ocean, where no snowpack can exist, we penalize the model harshly for placing SWE values in these areas;
- We penalize the difference in total SWE between cGAN solutions and physics model output, to ensure that total stored water mass is properly estimated.

Training details. As in [9], we modified the standard GAN training scheme by first training the generator purely on L_1 loss term to estimate the conditional mean (for the first 5 epochs), and later adding an adversarial loss term to teach the generator finer details. We have also observed that this slight modification enables faster convergence to better solutions (with lower overall loss values). All deep learning training and inference was performed on a single consumer Graphical Processing Unit (GPU), the NVIDIA GTX 1080ti.

III. EVALUATION

Having trained the cGAN model as described above on a training set of 8 years of data (input/output pairs as described above at daily resolution), we have first tested its performance on a holdout sample of two years of data. This is a standard regression setting, for which we compute typical performance metrics. Even in this much simplified setting where we don't explicitly model time, the model achieves a mean absolute percentage error (MAPE) of 9.54%, indicating that it has learned a reasonably accurate mapping from meteorological and topographical data to simulated SWE.

In Figure 3 we show a comparison between cGAN and physics-based model output over 2-week periods at the end (June/August), start (November), and peak (April), respectively, of the SWE season (left, middle, and right panels in the figure, respectively). These are key periods of interest to mountain snowpack researchers and water resource managers, as they are check-in points in the lifecycle of mountain snowpack dynamics. We show the histograms of normalized pixel

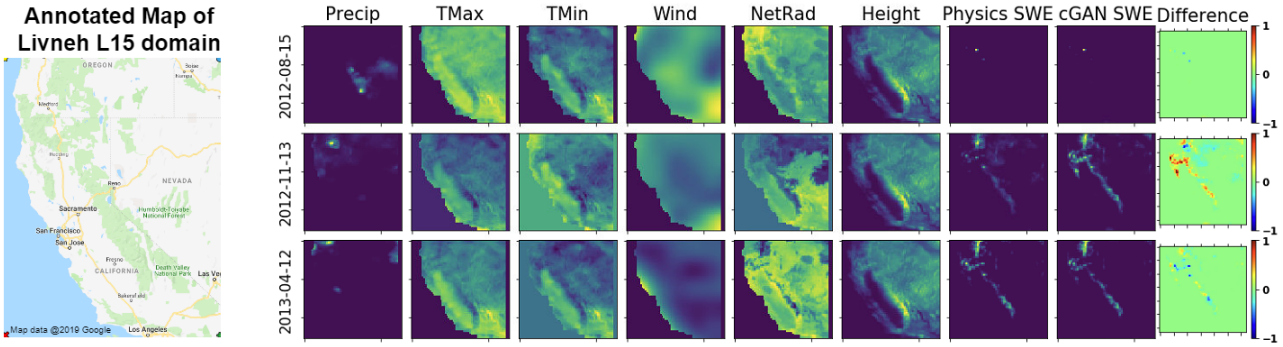


Fig. 2. Three samples (per row) of model inputs (meteorological forcings) on the first 6 columns, physics model output (column 7). cGAN output (column 8) and difference between physics model and cGAN (column 9). Rotated to match Sierra Nevada range (left).

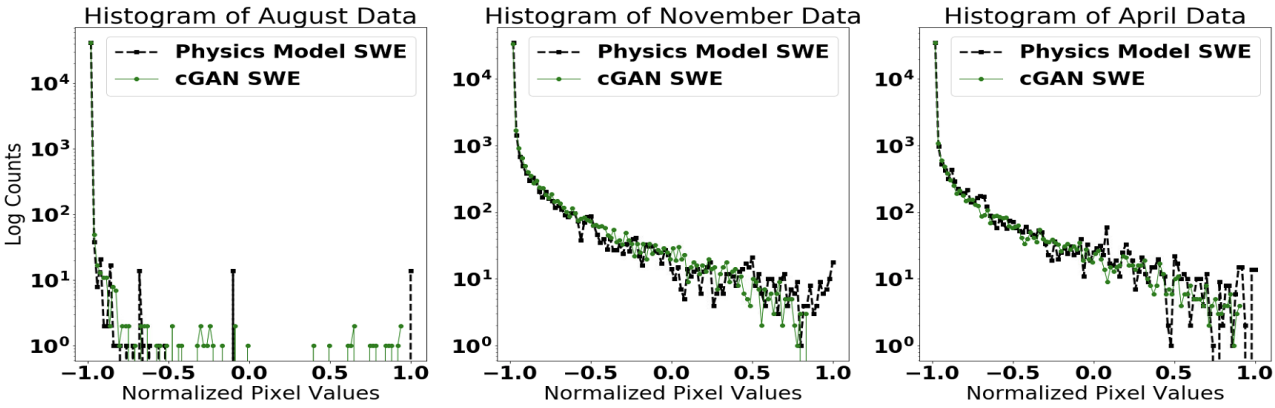


Fig. 3. Histograms of normalized pixel values comparing cGAN (green) and physics-model (black) across key snowpack seasons.

values of cGAN output (green line) and physics-based model output (black line). Note that the cGAN model is able to accurately recover the distributions of values of the physics-based model.

Next, we calculate and plot the power spectral density function (PSD) for both the cGAN and physics-based model output. This metric incorporates information across all spatial frequency scales, and is defined as:

$$PSD = 10 \log_{10} |\mathcal{F}(\rho(SWE, SWE))|^2, \qquad (1)$$

where $\mathcal{F}$ denotes the Fourier transform and ρ the two-point-correlation function, defined as usual. In Figure 4 we show the PSD profile comparisons for the key SWE seasons. Here too we observe strong performance - the spectral properties of the outputs the cGAN and the physics model are very similar, indicating that the cGAN performs well not just at 'memorizing' averages, but is also capable of recreating high frequency details.

Lastly, we have validated our hypothesis that inference time with a trained cGAN is extremely fast, taking less than 10 seconds to generate over 1000 simulated SWE grids on a GTX 1080ti, a consumer GPU. This suggests a speedup factor of around 1000x compared to just the raw runtime of a VIC model used to generate the SWE grids, which by our estimates takes $\sim$ 100 core-hours to simulate 100 years of SWE output. Even when taking the $/sim30$ minutes of training time into account, we still observe a massive speedup. When focusing on a particular geography at a lower resolution, this speedup will allow for a much faster iteration to analyze how changes in meteorological forcing lead to changes in SWE, a topic of future research.

IV. INCORPORATING PHYSICS: ABLATION STUDY

To understand the effect of the physics-informed constraints and other inputs to our model, we performed an ablation study. We found that the inclusion of one input channel in particular, Net Radiation (NetRad) - a measure of the difference between incoming and outgoing atmospheric radiative energy - increased performance over all measured metrics. This matches up well intuitions from atmospheric science, where satellite radiometer data is used for the estimation of snowpack variables. We also find that our physics informed penalties improve performance on all metrics and convergence rates. This falls in line with physical

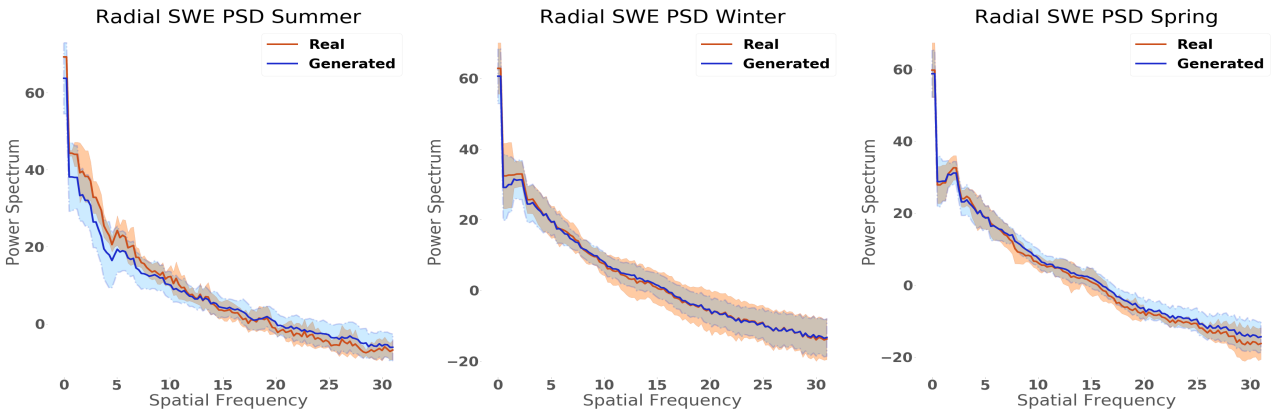


Fig. 4. Power Spectral Density of cGAN and Physics Model over different hydrologic seasons: respectively the end (Summer: July/August), start (Fall/Winter: November/December), and peak (Spring: April/May) of SWE season.

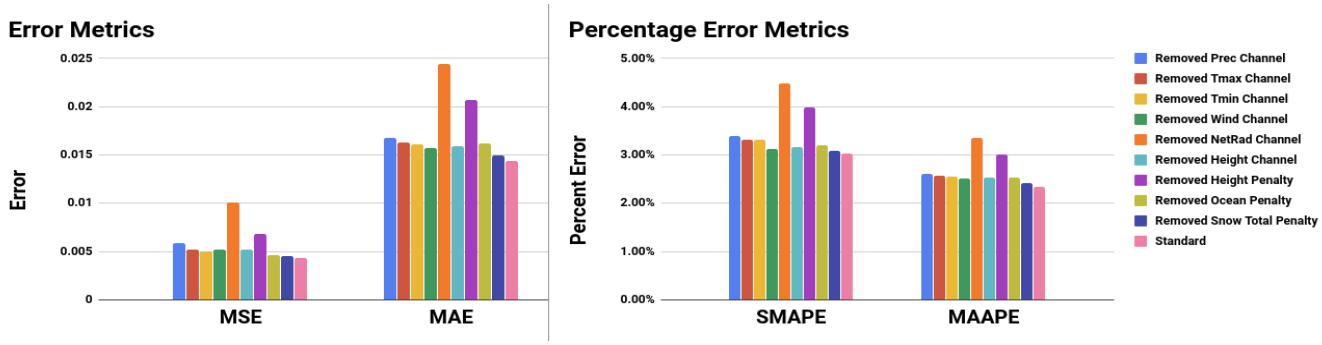


Fig. 5. Error metrics computed during ablation study. Each bar label represents the penalty or channel that was removed for the corresponding experiment. We observe that the inclusion of the Net Radiation channel and Height penalty both improve model quality, as the models excluding them resulted in significant increases in error.

intuition - penalizing physically incoherent solutions results in better performance.

We found that including 'physics penalties' improved performance across metrics. The inclusion of a penalty for the GAN assigning snow on known water areas greatly improved convergence during training. The inclusion of the height-based penalty of SWE errors at high altitudes made the model far stronger at generating sparse gridded SWE outputs and at recreating the tails of distributions as seen in Figure 3. The penalty on total SWE error did result in a slight increase in mean error, but forced the model to generate solutions with stronger physical coherence than as indicated by MAPE alone.

Figure 5 contains representative metrics logged for each parameter combination in the ablation study. Metrics computed are Mean Squared Error (MSE), Mean Absolute Error (MAE), Symmetric Mean Absolute Percent Error (SMAPE), and Mean Arctangent Absolute Percent Error (MAAPE). The traditional Mean Absolute Percent Error (MAPE) was also computed, but we found it not useful as a metric for model performance in SWE due to the frequently sparse maps - 'actual' data points are frequently zero, resulting in large error amplification. We recommend SMAPE and MAAPE as alternative metrics to correct this very problem.

V. FUTURE WORK.

Future work includes investigation of further "soft" and "hard" constraints such as additional loss terms/penalties and fundamental architectural changes can improve the training, coherence, and metrics of the model. Furthermore, our work so far has modeled SWE as an $i.i.d.$ process with no temporal or stochastic elements. Further work can be done to incorporate temporal modeling techniques to better model this fundamentally uncertain spatio-temporal process.

cGANs and Deep Learning are not the only ways to estimate hydro-meterological variables like SWE. Other numerical and machine learning models have their own strengths, such as numerical stability and explicit likelihood calculations. cGANs and Deep Learning have great advantages in speed and hardware acceleration, but future work can also be done to investigate and incorporate strengths from other fundamentally different models.

REFERENCES

[1] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[2] T. Schneider, S. Lan, A. Stuart, and J. Teixeira, "Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations," *Geophysical Research Letters*, 2017.

[3] R. Essery, N. Rutter, J. Pomeroy, R. Baxter, M. St"ahli, D. Gustafsson, A. Barr, P. Bartlett, and K. Elder, "Snowmip2: An evaluation of forest snow process simulations," *Bull. Amer. Meteor. Soc*, vol. 90, pp. 1120–1136, 2009.

[4] A. M. Rhoades, A. D. Jones, and P. A. Ullrich, *Assessing Mountains as Natural Reservoirs with a Multi-Metric Framework*. Earth's Future (In Revision), 2018.

[5] M. S. Raleigh, J. D. Lundquist, and M. P. Clark, "Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework, hydrol," *Earth Syst. Sci*, vol. 19, pp. 3153–3179, 2015.

[6] K. S. Jennings, T. S. Winchell, B. Livneh, and N. P. Molotch, *Spatial variation of the rain-snow temperature threshold across the Northern Hemisphere*. Nat. Commun, 2018.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," *ArXiv e-prints*, Nov. 2016.

[8] B. Livneh, T. J. Bohn, D. W. Pierce, F. Munoz-Arriola, B. Nijssen, R. Vose, D. R. Cayan, and L. Brekke, "A spatially comprehensive, hydrometeorological data set for mexico, the u.s., and southern canada 19502013," *Scientific Data*, vol. 2, no. 150042, 2015.

[9] D. Hall, J. Stewart, and C. Tierney, "Adversarial networks for satellite simulation and dataset translations," *Proceedings of the 8^{th} Workshop on Climate Informatics, Boulder, CO, 2018*, 2018.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

STUDY OF THE IMPACT OF CLIMATE CHANGE ON PRECIPITATION IN PARIS AREA USING METHOD BASED ON ITERATIVE MULTISCALE DYNAMIC TIME WARPING (IMS-DTW)

Mohamed Djallel Dilmi[1], Laurent Barthès[1], Cécile Mallet[1], Aymeric Chazottes[1]

Abstract— **Studying the impact of climate change on precipitation is constrained by finding a way to evaluate the evolution of precipitation variability over time. Classical approaches (feature-based) have shown their limitations for this issue due to the intermittent and irregular nature of precipitation. In this study, we present a novel variant of the Dynamic time warping method quantifying the dissimilarity between two rainfall time series based on shapes comparisons, for clustering annual time series recorded at daily scale. This shape-based approach considers the whole information (variability, trends and intermittency…). We further labeled each cluster using a feature-based approach. While testing the proposed approach on the time series of Paris-Montsouris, we found that the precipitation variability increased over the years in Paris area.**

I. MOTIVATION

Climate change is a reality widely recognized today [1], [2]. It corresponds to a lasting change in parameter statistics such as those of the distribution of temperatures or precipitations over a period of several decades to several thousand years. These changes may be due to processes intrinsic to the Earth, external influences such as solar radiation or, more recently, human activities. The anthropogenic climate change responsible for global warming is the result of greenhouse gas emissions generated by human activities, which alter the composition of the planet's atmosphere. Due to the direct correlation between greenhouse gas emissions (particularly CO2) and temperature, the impact of climate change on temperature has been studied extensively. Nevertheless, its impact on precipitation remains unclear; it is sometimes assumed that precipitation variability does not change in a warming climate [3], [4], or that mean precipitation and its variability change at the same rate [5], or even that the precipitation variability increases in warmer climate [6], [7]. It remains difficult to observe and predict any impact of climate change due to the intermittent and irregular nature of precipitation [8], [9]. These different works show the difficulty to conclude about the evolution of rainfall variability in the context of climate change.

Only rain gauges provide continuous observation of precipitation throughout the last century. They provide daily time series of uninterrupted rainfall with valuable information on the history of the regional climate and thus allow to study the evolution of precipitation over time. Since precipitation is an intermittent phenomenon that takes the form of precipitating events, the daily scale, close to the duration of the events, is well suited for the study of precipitation behavior over the past 150 years. In climatology, it is customary to split the time series into hydrological years (from September to August) and compare them when trying to detect an evolution.

Using a feature-based approach (e.g., mean rain rate, maximum values of rain rates …) to characterize the variability of rainfall time series and describe its evolution over years is sub-optimal [10]

Corresponding author: M. D. DILMI djallel.dilmi@latmos.ipsl.fr
[1] LATMOS/CNRS/UVSQ/Université Paris-Saclay, 11 boulevard d'Alembert, 78280 Guyancourt, France

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

because there is more information (about how the variability evolves) that is available if the entire annual time series is used than if only some extracted features are used. Dilmi et al. [11] proposed a shape-based approach for rainfall time series comparison that considers the whole information (variability, tends and intermittency…) of time series while quantifying the dissimilarity between them, the details of the method are presented in part II. In part III, we applied this approach to cluster annual time series observed at daily scale, which allowed us to track the evolution of precipitation over the years. The interpretation of the raw outputs of this approach is performed by applying a feature-based approach.

In this paper, we describe a multi-step approach that combines the shape-based clustering algorithm and the feature-based approach to overcome the individual limitations of both approaches, in order to better investigate and describe the evolution of precipitation variability over the years.

II. METHOD

Given a long time series of daily-scale rain rates RR [mm.h^{-1}] that was split into a set of **N** hydrological annual time series, **we first compare all pairs (i,j) of the NxN annual time series.** The comparison method we used is the iterative multiscale dynamic time warping (IMs-DTW) method [11]. Described as a time-normalized distance between two rainfall time series, it is a variant of the Dynamic time warping [12][13] which searches for an optimal match (called alignment) between the two time series. It allows to stretch and compress some sub-sections of the time axis at different time scale while respecting some constraints in order to minimize the dissimilarity between the compared time series. Then, it assigns a dissimilarity score based on the found alignment to the compared pair of annual time series. Finally, all dissimilarity scores are ranked in a matrix called dissimilarity matrix [NxN].

The second step is using the dissimilarity scores to perform clustering of annual time series. The clustering method we used is K-medoids [14][15] which is a variant of K-means where the cluster medoid is defined to be the closest annual time series to the set of annual time series in the cluster. The clustering method is as follows:

1) Randomly initialize K medoids $m_k \; with \; k = 1 \dots K$ (K annual time series), one for each cluster C_k.
2) For every annual time series in the dataset, find the nearest cluster's medoid based on the dissimilarity score calculated above, and then assign the time series to the corresponding cluster.
3) Find the central medoid m_k for new clusters $C_k \;\; k = 1 \dots K$ which minimize the formula : $m_k = \arg\min_{j \in C_k} (\sum_{i \in C_k} dissim(i,j)^2)$ with dissim(i,j) representing the dissimilarity score between the two annual time series i an j.
4) Repeating steps 2 and 3 until there are no more changes.

The third step consists of labeling each cluster separately based on a detailed analysis of its medoid and the characteristics it shares with the cluster. Extracted precipitation indices were used in this step (ex. Maximum and standard-deviation of rain rates over the year, precipitation amount, precipitation duration, consecutive wet days, consecutive dray days …) [1], as well as other environmental features that can be easily interpreted [1] and historical rain bibliography [7][16].

Finally, after labeling each cluster, **we analyze the evolution of their frequencies over time.** For this we split the time axis into several successive time intervals large enough and from each cluster calculated the frequency of its presence, and sought to identify trends. In general, a period of 30 years is suggested for intervals by the world Meteorological Organization (WOM).

III. EVALUATION

We chose to apply the approach to study the area of Paris, specifically on the station of Paris-Montsouris in France, which provides an unbroken time series of daily rain gauge measurements of precipitation since 1873, validated by the World Meteorological Organization (WMO).

Following the approach proposed above, we carried out an analysis of the time series measured between Sept.1st, 1873 and Aug. 31st, 2019 at the daily scale, split into 146 annual time series.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

The comparison of all annual time series by pairs using the IMs-DTW provides a set of alignments. Figure 1 illustrates, by way of example, the alignment obtained between the two annual time series representing years 1890 and 1900. The alignment structure ensures the association of the peaks and the rainy periods of the two years, we note that the associations of the matching present relatively small offsets with up to two weeks. This observation is true for all the alignments proposed by the IMS-DTW between the 146 annual times series.

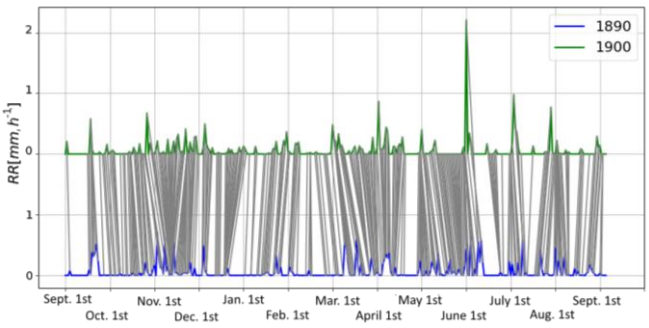


Figure 1: The alignment that links the two annual rainfall time series representing the two years 1890 and 1900 measured at the station Paris-Montsouris

The small distortions of time axis (less than 2 weeks) between the annual time series implicitly guarantee the seasonal comparison of different years (summer compared to summer and winter to winter). Sometimes, a season can be shifted by a few weeks, and the IMS-DTW not only allows to correct but also to precisely evaluate this shift.

The dissimilarity scores are ranked in the dissimilarity matrix D[146x146] presented in Figure 2. The higher the value, the more dissumilar two years are.

All observations reported on "climate change in Paris" published by Météo-France [14] are verified and visible on the dissimilarity matrix D. Among the annual time series that are standing out on the dissimilarity matrix, we can mention:

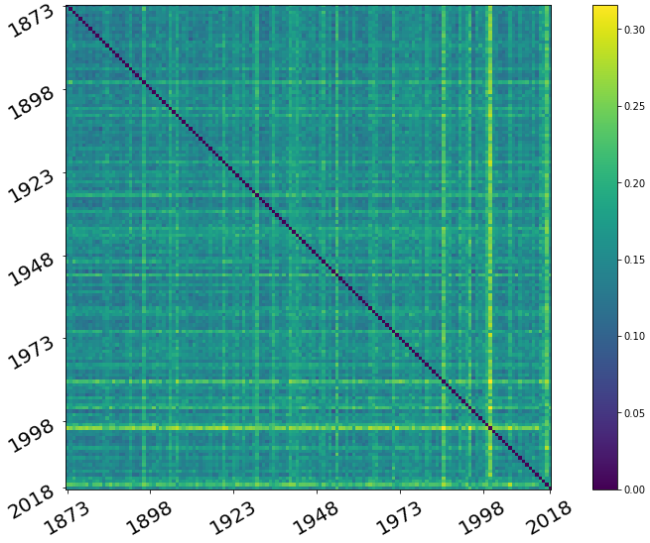


Figure 2: the dissimilarity matrix D (146x146) between the 146 annual time series measured in Paris-Montsouris between 1873 and 2019

1) The annual time series 1906 (#34) known in the Parisian history for the flood of the Seine River and listed as an exceptional year, presents important dissimilarities from other years.
2) There is also a band that marks the four years corresponding to the period from September 1st, 1940, to August 31st, 1944 (# 68 to 71). This could be explained by disturbances in data collection during the war period.
3) The annual time series 1954 (# 82), also listed as an exceptional year for the flood of the Seine River.
4) The annual time series 2000 (# 128) has strong dissimilarities (the maximums) compared to the other annual time series of the study.

The time series clustering into four clusters ($K = 4$) of annual time series provides four central medoids. Their analysis, associated with that of the corresponding cluster, allows the following ascertainments:

- The central medoid of the first cluster (#C1) is 1953 (#81), this cluster groups 34 annual time series. All are characterized by particularly low precipitation intensity, variability and water accumulation; as well as longer periods without rain.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

These characteristics correspond to the label **"drought years"**.

- The central medoid of the fourth cluster (C4) is 1884 (#12), this cluster (#C4) groups 28 annual time series, all are characterized by extreme variability and violent thunderstorms in summer that recorded exceptional values for rain rates and variability. These characteristics correspond to the label **"remarkable years with extreme variability".** All the remarkable years reported in MétéoFrance's report [14] belong to this cluster.
- The two clusters C2 (1995 (#123) as medoid) and C3 (1931 (#59) as medoid) were difficult to labeling because the two of them group normal years.

Considering the appearance frequency of each cluster in seven successive 26 years' time intervals between 1873 and 2019 allows studying the possible evolution of precipitation (see table 1).

Temporal Window		Cluster C1	C2	C3	C4
1873	1899	29	19	16	07
1892	1918	21	19	19	14
1911	1937	18	15	29	11
1930	1956	18	11	24	25
1949	1975	21	23	81	21
1968	1994	18	23	11	21
1987	2018	09	28	16	36
Gradient		↘	↘↗	↗↘	↗

Table 1. Temporal evolution of frequencies for each cluster [% by column]. The temporal window is 26 years length with an overlap of 7 years.

With a coarse time window, we see that the presence of the first cluster C1 (**drought years**) decreases over time while that of C4 (remarkable years with extreme variability) increases or stagnates at the same time which corroborates the assumption that the precipitation variability increases over time (# warmer climate) between 1873 and 2019.

IV. LIMITATION AND CURRENT WORK

The proposed method shows that there exists a temporal evolution of precipitation when studiying Paris-Montsouris time series. However limitations of the current work include the fact that we have not yet demonstrated the robustness of the results in table 1, i.e. we have not yet tested whether other rainfall time series measured in Paris area could lead to similar results nor demonstrated that results show a real impact of climate change rather than just representing a temporal correlation. We are currently working on both of these topics, and are furthermore exploring the correlation between these clusters and environnemental parameters (such as temperature). We also plan to extend the study area and to use other rainfall time series. The proposed unsupervised classification of annual time series based on a new dissimilarity measure seems particularly well adapted to the intermittent nature of precipitation time-series. The K medoid approach was used so as to facilitate the interpretation step, which allowed to label each year in function of theire structure of annual precipitation. Finally, the analysis of the frequency of these labels of years from 1873 to the current day makes it possible to describe the annual evolution of precipitation observed at a daily scale.

References

[1] Zhiying Li, Xiao Li, Yue Wang, Steven M. Quiring, Impact of climate change on precipitation patterns in Houston, Texas, USA, Anthropocene, 25, 2019, ISSN 2213-3054, doi:10.1016/j.ancene.2019.100193.

[2] L'Agence Parisienne du Climat et Météo-France, (2016). *Le changement climatique en France au XXeme siècle*, juillet 2015 - ISBN : 978-2-9548167-3-9.

[3] Hawkins, E. & Sutton, R. The potential to narrow uncertainty in projections of regional precipitation change. *Clim. Dyn.* **37**, 407–418 (2011).

[4] Thompson, D. W. J., Barnes, E. A., Deser, C., Foust, W. E. & Phillips, A. S. Quantifying the role of internal climate variability in future climate trends. *J. Clim.* **28**, 6443–6456 (2015).

[5] Gellens, D. & Roulin, E. Streamflow response of Belgian catchments to IPCC climate change scenarios. *J. Hydrol.* **210**, 242–258 (1998).

[6] Website : http://wikhydro.developpement-durable.gouv.fr/index.php/Changement_climatique_-

%C3%A9volution_des_pr%C3%A9cipitations (last check april, 23rd 2019.

[7] Pendergrass, A. G., Knutti, R. , Lehner, F. , Deser , C. and Sanderson, B. M. , "Precipitation variability increases in a warmer climate", *Scientific reports* (2017).

[8] Verrier S., Mallet C., Barthès L., Multiscaling properties of rain in the time domain, taking into account rain support biases, Journal of Geophysical Research: Atmospheres, American Geophysical Union, 2011, 116, pp.D20119

[9] Akrour N., Chazottes A., Verrier S., Mallet C., Barthès L., Simulation of yearly rainfall time series at microscale resolution with actual properties: Intermittency, scale invariance, and rainfall distribution Water Resources Research, American Geophysical Union, 2015, 51 (9), pp.7417-7435.

[10] Dilmi. M. D., Méthodes de classification des séries temporelles de précipitations : application à un réseau de pluviomètres, Ph.D. thesis (2019), Sorbonne Universités.

[11] Dilmi, M. D., Barthès, L., Mallet, C., Chazottes, A. , "Iterative multiscale dynamic time wapring : a tool for rainfall time series comparison". *International Journal of Data Science and analytics*, June 2019.

[12] Sakoe, H. and Chiba, S., "A similarity evaluation of speech patterns by dynamic programming", *Dig. Nat. Meeting, Inst. Electron. Comm. Eng.* Japan, p. 136. 1970.

[13] Salvador S., Chan., P. "FastDTW: Toward accurate dynamic time warping in linear time and space". *Intelligent Data Analysis*, 11(5), 561-580 (2007).

[14] Kaufman, L., and Rousseeuw, P. J., "Clustering by means of medoids*", in ln : Dodge Y & editor, ed., , North Holland/ Elsevier*, p 405-416, 1987.

[15] Harikumar, S. and PV, S, "K-medoid Clustring for Heterogeneous DataSets", *Procedia Computer Science*, Vol. 70 , PP 226-237, 2015.

[16] Moisselin, J.M., Schneider, M., Canellas, C. and Mestre, O., "Le changement climatique en France au XXeme siècle : Étude des longues séries homogénéisées de données de température et de precipitations", *La Météorologie*, 38, pp. 45-56, 2002.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

GRAPH-GUIDED REGULARIZATION FOR IMPROVED SEASONAL FORECASTING

Abby Stevens[1], Rebecca Willett[1,1b], Antonios Mamalakis[2], Efi Foufoula-Georgiou[2,3], James Randerson[3], Padhraic Smyth[4], Stephen Wright[5] & Alejandro Tejedor[6]

***Abstract*—Understanding the factors that determine regional climate variability and change is a challenge with important implications for the economy, security, and environmental sustainability of many regions around the globe. Unprecedented quantities of high-resolution climate data provide an enormous opportunity to explore this question systematically and exhaustively. Simple, off-the-shelf machine learning and statistical analysis methods can yield misleading results when applied directly to such data. Standard model selection methods are fragile in the face of complex dependence structures in the climate system. This abstract describes a regression scheme that explicitly accounts for spatiotemporally correlated features via a regularization approach based on an underlying correlation graph. Using large ensemble climate outputs to estimate the strength of correlations among features, we form a graph with edge weights corresponding to pairwise correlations. This graph is used to define a graph total variation regularizer that promotes similar weights for highly correlated features. We apply our scheme to predicting winter precipitation totals in the southwestern US using sea surface temperatures (SST) over the entire Pacific basin at multiple time lags, and demonstrate that our method provides strong predictive performance.**

I. MOTIVATION

The growing quantities of high-resolution Earth observations and climate model output [1] provide an opportunity to discover previously unknown teleconnections (long-range connections among climate modes) with strong predictive potential to improve seasonal-to-subseasonal forecasting. However, many statistical prediction schemes which aim to exploit established climate teleconnections between large-scale modes of variability (e.g., the El Niño-Southern Oscillation, the Pacific North America pattern, the Madden-Julian Oscillation, etc.) and regional hydroclimate fail to capture the highly complex and nonstationary nature of the climate system. On the other hand, dynamical models show limited predictive skill at lead times longer than two weeks [2], due to imperfect physical conceptualizations and inaccurate initial conditions.

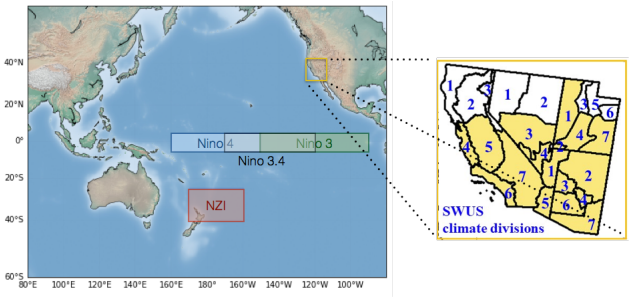


Fig. 1. The Pacific basin with known predictive regions highlighted and the specific climate regions over the expanded SWUS.

Recently, a new teleconnection between sub-tropical sea surface temperatures off the coast of New Zealand (NZI) and regional precipitation in the Southwestern US (SWUS) (Figure 1) was discovered, exhibiting stronger and earlier predictive potential than any other known mode of variability, including the El Niño Southern Oscillation (ENSO), which has long been used for SWUS seasonal precipitation forecasting [3]. A natural question to ask is whether there are additional, undiscovered teleconnections that, once identified, could improve seasonal forecasting.

Rather than relying on ad hoc methods for discovering important teleconnections, we seek to cast the forecasting problem as a regression problem in which modes (influential climate patterns) are not specified in advance but rather are allowed to emerge from the data as sources of predictability. Such a method must account for small sample sizes and high dimensionality, strong spatiotemporal dependencies among the predictors, and the need for interpretability in the climate

[1]Department of Statistics, University of Chicago, [1b] Department of Computer Science, University of Chicago, [2]Department of Civil and Environmental Engineering, University of California, Irvine, [3]Department of Earth System Science, University of California, Irvine, [4]Department of Computer Science, University of California, Irvine, [5]Computer Sciences Department, University of Wisconsin, Madison, [6]Max Planck Institute for the Physics of Complex Systems, Dresden Germany

sciences. Regularized regression has shown promise in accomplishing this task ([4], [5], [6], [7]).

However, a key challenge is that the covariates or features of such models (i.e., SSTs over space and time) are highly correlated, violating key assumptions underlying many modern high-dimensional regression methods such as the Lasso. Simultaneously, we seek to leverage data generated via simulation of physics-based climate models in addition to observational data. Treating simulation data as additional samples of observational data fails to account for model errors and sampling bias associated with simulations [8]. This abstract describes an approach in which we perform regularization over a graph corresponding to the correlation structure underlying the data. This approach (a) is provably robust to strong correlations among features or covariates and (b) leverages climate model simulations by using them to set parameters of the regularizer.

II. Problem Formulation

The SWUS region is highly vulnerable to drought, which has severe economic and ecological implications. Accurate and early prediction in this region is therefore of particular interest. The largest amount of yearly precipitation occurs during winter (November-March), and there is high inter-annual variability. The SWUS is divided into 16 climate divisions of interest [3], and we are able to extract winter averages from each region for 1940-2015 [1].

Given known teleconnections that are the current state-of-the-art for seasonal forecasting [3], we use as our predictors sea-surface temperature (SST) at various time lags across the Pacific basin. Specifically, we consider mean SST on a $10^\circ \times 10^\circ$ grid in July, August, September, and October. The data are from the 20th Century Reanalysis project [2]. We consider 226 locations over the Pacific at 4 different time lags, for a total of $p = 904$ features and $n = 75$ years of observations.

For a given year i and climate division r, we seek to solve the regression problem

$$y_r^{(i)} = \sum_{j=1}^{p} X_j^{(i)} \beta_j + \epsilon^{(i)}$$

where $X_j^{(i)}$ is the summer SST measurement at (location, time lag) j preceding $y_r^{(i)}$, the winter precipitation observation for region r. We also assume $\epsilon^{(i)} \sim N(0, \sigma^2)$. Although precipitation is non-negative and tends to be skewed in distribution, we find experimentally that we can reasonably approximate monthly winter totals over the SWUS regions with Gaussian noise. X and y are both centered and normalized to have unit variance and mean zero. We know that the columns of X are highly correlated in both space and time. Our goal is to estimate coefficients β that yield low prediction error for the seasonal forecasting problem and are physically interpretable from a climate science perspective.

[1] https://www.ncdc.noaa.gov/cag/time-series/us
[2] https://www.esrl.noaa.gov/psd/data/20thC_Rean/

III. Methods

A. Graph total variation

For a response $y \in \mathbb{R}^n$ and covariates $X \in \mathbb{R}^{n \times p}$, $p \gg n$, we seek to estimate β^* such that $y = X\beta^* + \epsilon$ where β^* is well-aligned with the correlation structure of X. For a zero-centered X, let $\Sigma := E(X^T X)$ be the covariance matrix of X and $\hat{\Sigma}$ be an estimate of Σ. Let $\hat{s}_{j,k} := \text{sign}(\hat{\Sigma}_{j,k})$. Our estimator, which we call *graph total variation* (GTV) [9], is given by

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta} \|y - X\beta\|_2^2 \\ &+ \lambda_{TV} \sum_{j,k} |\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k}\beta_k| \\ &+ \lambda_1 \|\beta\|_1, \end{aligned} \tag{1}$$

where λ_1 and λ_{TV} are regularization parameters chosen through cross validation. We can interpret this estimator from a graph perspective by defining a *covariance graph* based on $\hat{\Sigma}$. Let $G = (V, E, W)$ be an undirected weighted graph with vertices $V = \{1, 2, \ldots, p\}$, edges $E := \{(j,k) : |\hat{\Sigma}_{j,k}| > 0, j \neq k\}$, and weight matrix W with $W_{j,k} = w_{j,k} = |\hat{\Sigma}_{j,k}|^{1/2}$. Let $\Gamma \in \mathbb{R}^{|E| \times p}$ be the *weighted edge incidence matrix*, of G, where each row ℓ represents a pair of connected vertices (j_ℓ, k_ℓ):

$$\begin{aligned} \Gamma_{\ell, j_\ell} &= |\hat{\Sigma}_{j_\ell, k_\ell}|^{1/2} \\ \Gamma_{\ell, k_\ell} &= -\text{sign}(\hat{\Sigma}_{j_\ell, k_\ell}) |\hat{\Sigma}_{j_\ell, k_\ell}|^{1/2} \end{aligned} \tag{2}$$

We can thus simplify the total variation term in (1) as

$$|\hat{\Sigma}_{j,k}|^{1/2} |\beta_j - \hat{s}_{j,k}\beta_k| = \|\Gamma\beta\|_1$$

It is worth highlighting the differences between GTV and similar well-studied structured estimators, such as the fused Lasso [10] and the generalized Lasso [11]. The theoretical guarantees of these estimators assume that X satisfies the restricted eigenvalue condition [12], which is often violated when the columns of X are

highly correlated. GTV, on the other hand, performs well in the presence of strong correlations.

GTV promotes estimates of β that contain sparse clusters of coefficients corresponding to highly correlated variables. That is, the stronger the correlation between X_j and X_k, the more similar $\hat{\beta}_j$ and $\hat{\beta}_k$. In the case of perfect correlation (i.e. $X_j = X_k$), Lasso will assign weight to either X_j or X_k, while GTV will distribute the weight evenly across X_j and X_k. This results in more interpretable model selection, which is of interest in climate and other application areas. This estimator adaptively selects clusters of features aligned with an estimated underlying graphical structure.

B. Covariance matrix estimation

The GTV regularization term depends on a reliable estimate of Σ. It is well documented ([13], [14]), that, in high-dimensional settings where $p \gg n$, the sample covariance $\hat{\Sigma}_S = \frac{1}{n} XX^T$ is not a consistent or accurate estimate of Σ. In some application areas, including climate science, there exists side information that is not based on X with which we can estimate Σ.

Climate models are physical mathematical models that simulate how energy and matter interact. In climate and related domains, there is hope that leveraging climate models alongside observations can help reduce uncertainties in predictive schemes [15]. One way to do this is data augmentation, which treats the simulation data as independent and identically distributed draws from the distribution of the observed data and combines all data to feed to the model, but this can be problematic in light of model biases and sensitivities to initialization [8].

We propose treating these climate simulations as side information we can use to estimate Σ. We use simulations from the CESM Large Ensemble Project, known as LENS[1]. LENS is a 40-member ensemble of Community Earth System Model V1 (CESM1) simulations, each of which is subject to the same radiative forcing scenario but with slightly perturbed initial temperature conditions. We linearly interpolate the LENS simulations of summer SST onto the same spatial grid as our observations.

Letting $X_L \in \mathbb{R}^{40n \times p}$ be the centered matrix of stacked features from all LENS ensemble members, we let $\hat{\Sigma}_L$ be the sample covariance matrix of X_L. We are assuming that both X and X_L are draws from a distribution with the same covariance matrix, Σ, and since X_L has a much higher sample size than X, we believe $\hat{\Sigma}_L$ is a better estimate of Σ than $\hat{\Sigma}_S$.

[1] http://www.cesm.ucar.edu/projects/community-projects/LENS/

C. Multitask GTV

We know that precipitation patterns across the entire SWUS are tied to similar summer atmospheric events, but possibly to a different degree for each region. Because of this, we seek to simultaneously solve m regression problems, a technique known as *multitask learning*. We assume that there is an unknown subset of covariates that are relevant for prediction, and this subset is preserved across the m regions. Let $Y = [y^{(1)}, y^{(2)}, \dots y^{(m)}] \in \mathbb{R}^{n \times m}$ be the matrix of the m response vectors corresponding to each climate region shown in Figure 1 and $B = [\beta^{(1)}, \beta^{(2)}, \dots \beta^{(m)}] \in \mathbb{R}^{p \times m}$ be the matrix of the corresponding m coefficient vectors. We wish to solve the following objective function, which we refer to as *MultiGTV*:

$$\hat{B} = \arg\min_B \|Y - XB\|_F^2 + \lambda_1 \sum_{r=1}^{m} \left(\|\Gamma \beta^{(r)}\|_1 + \|\beta^{(r)}\|_1 \right) + \lambda_2 \sum_{j=1}^{p} \|B_{j,:}\|_2 \quad (3)$$

where Γ is the edge-incidence matrix from (2) and $B_{j,:}$ is the j^{th} row of B. The first regularization term encourages coefficient estimates for each region that are sparse and well-aligned with the covariance of X, while the second promotes similarity in the support of the coefficient vectors across regions. We use a variant of an ADMM algorithm (Alternating Direction Method of Multipliers) [16] to solve this objective function.

IV. EVALUATION

For all experiments, the models are trained on the first 50 years of data and the remaining 25 years are held out for testing. Regularization parameters are chosen through 5-fold cross validation on the training data, and all reported errors are computed on the test data.

We compare the performance of GTV and MultiGTV with other well-known structured regression methods. Because GTV estimates coefficients for a single response and MultiGTV estimates many responses simultaneously, in order to compare performance in a meaningful way we report errors as follows. For a given region, we compute the mean squared error (MSE) on the corresponding test data. Then, we compute and report the area-weighted means and standard errors (SE) of the MSE across all regions in the SWUS.

We benchmark our methods against ordinary least squares (OLS) and Lasso (along with the multitask version of Lasso) in order to set a baseline. Then, we consider the following graph-based methods, which all solve a form of the GTV objective function with the edge-incidence matrix Γ defined by a variety of graphs:

1) Fused Lasso: The graph includes only immediate spatial and temporal neighbors
2) GTV with $\hat{\Sigma}_S$: The graph includes edges and weights based on the covariance of the SST observations, $\hat{\Sigma}_S$
3) GTV with $\hat{\Sigma}_L$: The graph includes edges and weights according to the covariance of the LENS SST simulations, $\hat{\Sigma}_L$

Each of these graphs can be used in the MultiGTV method as well. The results of all methods under consideration are shown in Table I.

Method	**MSE**	**SE**
OLS	1.018	0.038
Lasso	1.014	0.037
Fused Lasso	1.069	0.062
GTV with $\hat{\Sigma}_S$	0.989	0.052
GTV with $\hat{\Sigma}_L$	**0.949**	0.041
Multitask Lasso	0.964	0.029
Multitask Fused Lasso	0.929	0.036
MultiGTV with $\hat{\Sigma}_S$	0.921	0.035
MultiGTV with $\hat{\Sigma}_L$	**0.919**	0.027

TABLE I
AREA-WEIGHTED MEAN AND STANDARD ERRORS OF THE MSE FOR EACH OF THE 16 CLIMATE REGIONS UNDER CONSIDERATION. THE TOP SECTION INCLUDES METHODS THAT ESTIMATE COEFFICIENTS FOR EACH REGION SEPARATELY AND THE BOTTOM INCLUDES METHODS THAT ESTIMATE ALL REGIONS SIMULTANEOUSLY. THE BEST PERFORMING METHOD IN EACH SECTION IS IN BOLD.

We see that GTV outperforms the other methods in both the multitask and regular cases. Additionally, we see that estimating the covariance graph using the LENS data ($\hat{\Sigma}_L$) provides stronger predictive performance than using just the observations ($\hat{\Sigma}_S$). The multitask methods outperform their single-response counterpart, but it is worth noting that MultiGTV with both $\hat{\Sigma}_S$ and $\hat{\Sigma}_L$ result in nearly identical predictive performances, suggesting that the improvement in performance when using the LENS simulations seen in the ordinary GTV setting is not as influential in the multitask setting.

V. DISCUSSION

In this abstract we argue that the seasonal forecasting problem is improved by the use of graph-based regularization methods that explicitly account for spatial and temporal correlations among the features. We also present a novel method of leveraging large ensemble climate models to estimate the covariance graph for use in GTV, which results in the highest predictive performance of the methods considered. The intuition behind this discovery is that there are long-range teleconnections among climate variables that extend beyond nearest-neighbors approaches like fused Lasso, and accounting for these relationships is important for the forecasting problem.

This work lays the methodological foundation for a data-driven approach to seasonal forecasting that is grounded in physics. There are many potential next steps for this research. First, we note that the method as presented does not account for the nonlinearity of the system dynamics. We acknowledge that external forcing like anthropogenic climate change and/or natural multidecadal oscillations in the Pacific can affect the predictive skill of the algorithm. One of the next steps of this research is to account for changes in the weights of the predictors as a function of time.

Next steps also include a rigorous investigation of the locations and time lags at which Pacific SSTs are selected by our model to be predictive of SWUS precipitation. This analysis will address the robustness of the selected variables to help determine whether or not they correspond to true teleconnections from both a statistical and physical standpoint.

Finally, we hope to further validate our methods on different seasonal forecasting problems. Our methods are flexible and can easily adapt to different climate variables and prediction settings.

ACKNOWLEDGMENTS

Thanks to the following grants for funding this work: NSF DMS-1930049, NSF OAC-1934637, NSF DMS-18393366, ECCS-1839441 and DOE DE-AC02-06CH11357.

REFERENCES

[1] J. T. Overpeck, G. A. Meehl, S. Bony, and D. R. Easterling, "Climate data challenges in the 21st century," *Science*, vol. 331, pp. 700–702, 2011.
[2] N. A. of Sciences, "Strategies for subseasonal to seasonal forecasts," *The National Academies Press*, 2016.
[3] A. Mamalakis, J.-Y. Yu, J. T. Randerson, A. AghaKouchak, and E. Foufoula-Georgiou, "A new interhemispheric teleconnection increases predictability of winter precipitation in southwestern us," *Nature Communications*, vol. 9, 2018.
[4] A. R. Goncalves, A. Banerjee, V. Sivakumar, , and S. Chatterjee, "Structured estimation in high dimensions: applications in climate," in *Large-scale machine learning in the earth sciences*, ch. 2, pp. 13–32, CRC Press, 2017.

[5] S. Chatterjee, K. Steinhaeuser, A. Banerjee, S. Chatterjee, and A. Ganguly, "Sparse group lasso: Consistency and climate applications," *Proceedings of the 2012 SIAM International Conference on Data Mining*, 2012.

[6] S. He, X. Li, V. Sivakumar, and A. Banerjee, "Interpretable predictive modeling for climate variables with weighted lasso," *AAAI Conference on Artificial Intelligence*, 2019.

[7] T. DelSole and A. Banerjee, "Statistical seasonal prediction based on regularized regression," *Journal of Climate*, vol. 30, no. 4, 2017.

[8] K. A. Mckinnon, A. Poppick, E. Dunn-Sigouin, and C. Deser, "An observational large ensemble to compare observed and modeled temperature trend uncertainty due to internal variability," *Journal of Climate*, vol. 52, no. 19, 2017.

[9] Y. Li, B. Mark, G. Rasutti, and R. Willett, "Graph-based regularization for regression problems with highly-correlated designs," *arXiv preprint arXiv:1410.5093*, 2018.

[10] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, pp. 91–108, 2005.

[11] R. J. Tibshirani and J. Taylor, "The solution path of the generalized lasso," *The Annals of Statistics*, vol. 39, no. 3, pp. 1335–1371, 2011.

[12] G. Raskutti, M. J. Wainwright, and B. Yu, "Restricted eigenvalue properties for correlated gaussian designs," *Journal of Machine Learning Research*, vol. 11, pp. 2241–2259, 2010.

[13] T. T. Cai, Z. Ren, and H. H. Zhou, "Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation," *Electronic Journal of Statistics*, vol. 10, no. 1, pp. 1–59, 2014.

[14] A. Maurya, "A well-conditioned and sparse estimation of covariance and inverse covariance matrices using a joint penalty," *Journal of Machine Learning Research*, vol. 17, pp. 1–28, 2016.

[15] T. Schneider, S. Lan, A. Stuart, and J. Teixeira, "Earth system modeling 2.0: a blueprint for models that learn from observations and targeted high-resolution simulations," *Geophysical Research Letters*, vol. 44, pp. 12,396–12,417, 2017.

[16] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, 2011.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

TOWARDS PHYSICS-INFORMED DEEP LEARNING FOR SPATIOTEMPORAL MODELING OF TURBULENT FLOWS

Rui Wang[1], Adrian Albert[2], Karthik Kashinath[2], Mustafa Mustafa[2], Rose Yu[1]

***Abstract*—While deep learning (DL) has shown tremendous success in a wide range of domains, it remains a grand challenge to incorporate physical principles in a systematic manner to the design, training and inference of such models. Physics informed deep learning aims to infuse the principles governing the dynamics of the physical systems into data-driven DL models. However, existing studies are either limited to linear dynamics or spatial modeling of the physical systems.**

In this paper, we study the challenging task of spatiotemporal modeling of velocity fields for a nonlinear turbulent flow. We create benchmarks for the task and conduct comprehensive evaluations for various state-of-the-art physics informed DL methods. Our results show that the Convolutional-Deconvolutional Neural Network constrained by the Continuity Equation consistently performs the best for varying forecasting horizons. We also provide valuable technical insights for different physics informed DL techniques.

I. INTRODUCTION

Modeling the dynamics of physical processes that evolve over space and time and over multiple scales is a fundamental task that is wide-ranging in computational science. The current paradigm in atmospheric computational fluid dynamics (CFD) is physics-driven simulation: known physical laws encoded in systems of coupled partial differential equations (PDEs) are solved over space and time via numerical differentiation and integration schemes. Often, models of realistic systems employ empirical parametrization schemes and closures to model multi-scale physics that cannot be fully resolved. Simulating a single scenario of the spatio-temporal evolution of a realistic system over a realistic domain is extremely compute-intensive with current numerical approaches.

Corresponding author: *wang.rui4@husky.neu.edu* [1]Northeastern University, Boston, MA 02215 [2] Lawrence Berkeley National Laboratory, Berkeley, CA 94720

In recent years, machine learning has made important strides in all areas of computational and physical sciences. The main promises of modern deep learning for scientific applications are to automate, accelerate, and streamline highly compute-intensive and customized modeling workflows by employing a unified and scalable computational approach [1]. Previous work (e.g., [2], [3], [4]) has shown the effectiveness of machine learning models at representing complex spatiotemporal processes. Incorporating prior knowledge about physical or statistical properties of the system (e.g., as soft constraints, or penalties, into the optimization loss function) has shown to lead to more accurate solutions with faster convergence, while needing less training data. However, existing work is still limited. For example, [4] only considers a system with linear dynamics and [2] focuses on spatial modeling without the temporal aspect.

In this work, we study the spatiotemporal modeling of a highly nonlinear turbulent flow and investigate the performance of several state of the art physics-informed deep learning approaches. We benchmark these methods on the task of forecasting velocity fields at different future time horizons, given historic data of different lengths. The physical system we investigate in this work is Rayleigh-Benard convection, which is an idealized model for turbulent atmospheric convection. The system consists of a fluid bounded by two horizontal planar surfaces, where the lower surface is at a higher temperature than the upper surface. The adverse temperature gradient causes an unstable vertical profile of density, which results in convective motions for sufficiently large temperature-gradients. The governing equations for this physical system are the continuity, momentum and energy equations:

$$\nabla \cdot \mathbf{w} = 0$$
$$\frac{\partial \mathbf{w}}{\partial t} + (\mathbf{w} \cdot \nabla)\mathbf{w} = -\frac{1}{\rho_0}\nabla p + \nu \mathbf{w} + [1 - \alpha(T - T_0)]\chi$$
$$\frac{\partial T}{\partial t} + (\mathbf{w} \cdot \nabla)T = \kappa T,$$

where $\mathbf{w} = (u, v)$, p and T are velocity, pressure and temperature respectively, k is the coefficient of heat conductivity, ρ_0 is density at temperature T_0, α is the coefficient of thermal expansion, ν is the kinematic viscosity. Figure 1 shows a snapshot of the u and v velocities.

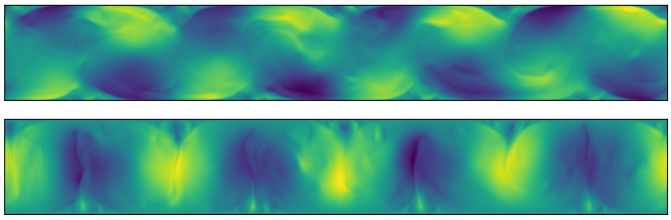


Fig. 1: Velocity field (u and v) at $t = 1000$ eddy turnovers

II. METHODS

We investigate various state-of-the-art physics-informed spatio-temporal DL models to learn and predict the time evolution of the turbulent nonlinear dynamics of the flow.

A. Purely Data Driven Models

We first attempt using three purely data-driven spatio-temporal models: Convolutional Neural Network (CNN), Convolutional-Deconvolutional Neural Network (CDNN) [5] and Convolutional-LSTMs (ConvLSTMs) [6]. We add skip connections to CDNN and residual connections to CNN to improve prediction performance. To produce multiple time-step forecasts, the predicted values are fed back in the models in an autoregressive manner.

B. Constrained CDNN

As described earlier, a goal is to incorporate known physics in the training process. One way is to constrain the training process with penalty terms from the governing equations. We test the influence of different terms used in [3]: divergence $\boldsymbol{\nabla} \cdot \mathbf{w}^2$, magnitude $\|\mathbf{w}\|^2$, and smoothness $\|\boldsymbol{\nabla} \cdot \mathbf{w}\|^2$. Based on the model performances on validation set, we only include smoothness(continuity equation) as a penalty term in the loss function of CDNN, $Loss = MSE(\hat{\mathbf{w}}, \mathbf{w}) + \lambda \|\boldsymbol{\nabla} \cdot \mathbf{w}\|^2$.

C. Deep Hidden Physics Model (DHPM)

Another way is directly using neural networks representing the solutions and making sure the output of the network satisfies the momentum and continuity equations during the training process. Our model is similar to [3]. We represent the velocity field $\mathbf{w} = (u, v)$ with two 4-layer neural networks with 200 neurons per hidden layer and use two 3-layer neural networks with 100 neurons per hidden layer to represent pressure p and a source term f that encapsulates the influence of temperature and viscosity. During the training, we make sure the outputs of these neural networks satisfy the continuity and momentum equations. Thus, the model can be formulated as

$$
\begin{aligned}
&\textit{Minimize} \\
&\|u - u^*\| + \|v - v^*\| + \\
&\|u_x + u_y\| + \|v_x + v_y\| + \\
&\|u_t + uu_x + vu_y - Pr(u_{xx} + u_{yy}) + p_x - f_1\| + \\
&\|v_t + uv_x + vv_y - Pr(v_{xx} + v_{yy}) + p_y - f_2\|
\end{aligned}
$$

$$
\begin{cases}
u = \textit{NN}(x, y, t) \\
v = \textit{NN}(x, y, t) \\
p = \textit{NN}(x, y, t, u, v) \\
f = \textit{NN}(x, y, t, u, v, u_x, v_x, u_y, v_y)
\end{cases}
$$

where u^* and v^* are true velocities and subscripts are partial derivatives.

D. PDE-CDNN

Finally, [4] successfully incorporated the solution of the energy equation into a DL model to predict sea surface temperature. We refer this model as PDE-CDNN. We modify this model to our problem by ignoring the two terms that contain acceleration and pressure in the momentum equation and linearizing the $(\mathbf{w} \cdot \nabla)u$ term, then for the u velocity, we have $\frac{\partial u}{\partial t} + (\mathbf{w} \cdot \nabla)u = \nu u$, which is similar to the energy (temperature) equation used in [4]. Thus, we can model u and v with two separate PDE-CDNN models but the inputs to these two parts are the same stacked u and v from previous time steps, and both parts are trained autoregressively.

III. EXPERIMENTS

A. Data

The dataset for our experiments is a 2-D turbulent flow simulated using a the Lattice Boltzmann Method. The numerical scheme is described in [7] and [8]. The values of the control-parameters are $Pr = 0.71$ (for air) and $Ra = 2.5 \times 10^8$. The numerical method was run for 23,630 eddy turnover times, and we use eddy turnovers from 1000 to 1200, where both total kinetic energy and Nusselt number have stabilized. The spatial resolution

for each snapshot is 1792 by 256 pixels. Figure 1 is the visualization of a velocity field at $t = 1000$ eddy turnovers in our dataset.

B. Setup

We divide each snapshot into 112 sub-images of size 64 by 64 pixels. We make the hypothesis that the data in a single sub-image contains enough information to forecast the future of the sub-images. We collected 10,080 sequences of sub-images of length 90 (in time). All images are also converted into coordinates (x, y, t) for training DHPM. We used a 60%-20%-20% training-validation-test split.

We trained all models using Adam optimizer and employed a learning rate decay of 0.9. Each model is evaluated with root mean square error (RMSE) metric and hyper-parameters are tuned using a validation set based on averages RMSEs of six steps ahead prediction. All models except DHPM are trained by the back-propagated through six steps ahead prediction errors. We computed the moving average of errors on validation set and used it as an early stopping criteria. All results are averaged over three runs.

IV. Results

The results of these experiments are summarized in Table I, including the average RMSEs of six velocity fields ahead predictions, time cost for one epoch, the best input length and the number of parameters for each model with tuned hyper-parameters. Notice that CDNN is the best architecture for modeling turbulence flow and we can obtain even lower prediction errors if we contrain its training with continuity equation as a penalty term. Compared with models, training ConvLSTMs is much more time-consuming. Also, even though last row suggests that DHPM, CNN and ConvLSTMs have significantly smaller number of parameters, we found that increasing their model complexity does not improve performance.

Figure 2 shows the relationship between input length and average RMSEs of velocity field predictions. Note that prediction errors of all models become stable after about 20 input steps, which suggests increasing the size of the history beyond the previous 20 timesteps does not improve predictions. Figure 3 shows growth of RMSE with prediction step. We see that Constrained CDNN consistently outperforms other methods. Figure 4 visualizes the relationship between the input length and forecasting horizons for Constrained-CDNN, which gives a sense of how much historical information is needed for predicting future velocity fields.

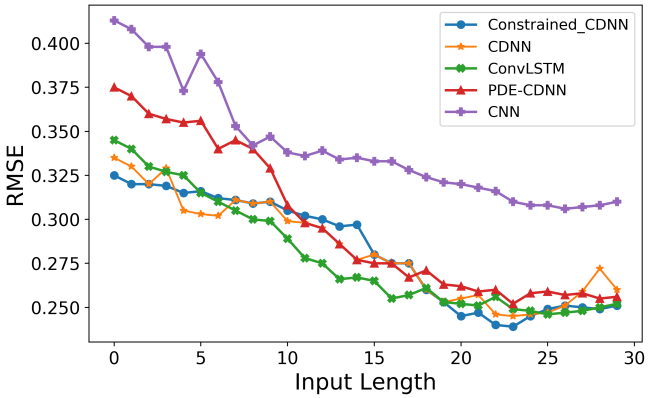


Fig. 2: Length of past input velocity fields vs. Average RMSEs of six step ahead velocity fields predictions.

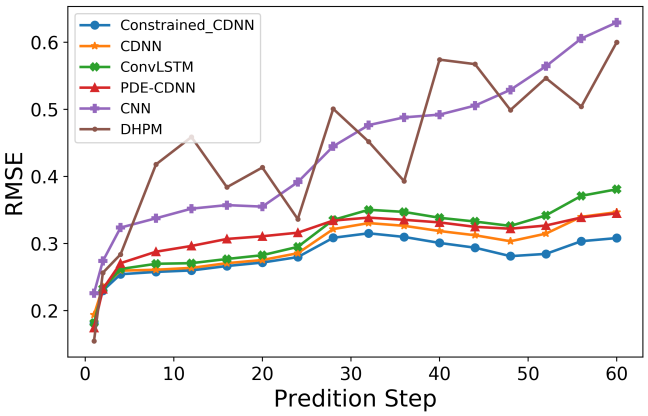


Fig. 3: Forecasting horizons vs. RMSEs.

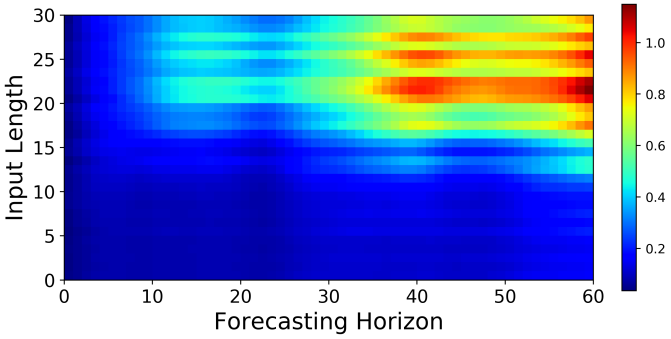


Fig. 4: Heat Map of RMSEs for Constrained-CDNN: Input length vs. Forecasting horizons

Figure 5 shows predictions of u velocity from the different models. PDE-CDNN does not perform as well as expected. We hypothesize that this is due to the fact that the most critical governing equation of this flow is the momentum equation, which is more complex than the energy equation, and that the simplifying assumptions that we use are not valid for this flow. We see that PDE-CDNN, CNN and DHPM perform poorly.

Models	DHPM	CNN	PDE-CDNN	ConvLSTMs	CDNN	Constrained CDNN
Average RMSE	0.278	0.302	0.252	0.246	0.245	**0.239**
Average Time for one epoch (s)	70.8	63.2	70.6	659.7	35.9	37.5
Best input length	NA	27	24	26	24	24
# Parameters(10^5)	10.5	15.8	499	23.8	249	249

TABLE I: Average RMSE of six velocity fields ahead predictions, time cost for one epoch, the best input length and the number of parameters of the best model for each architecture.

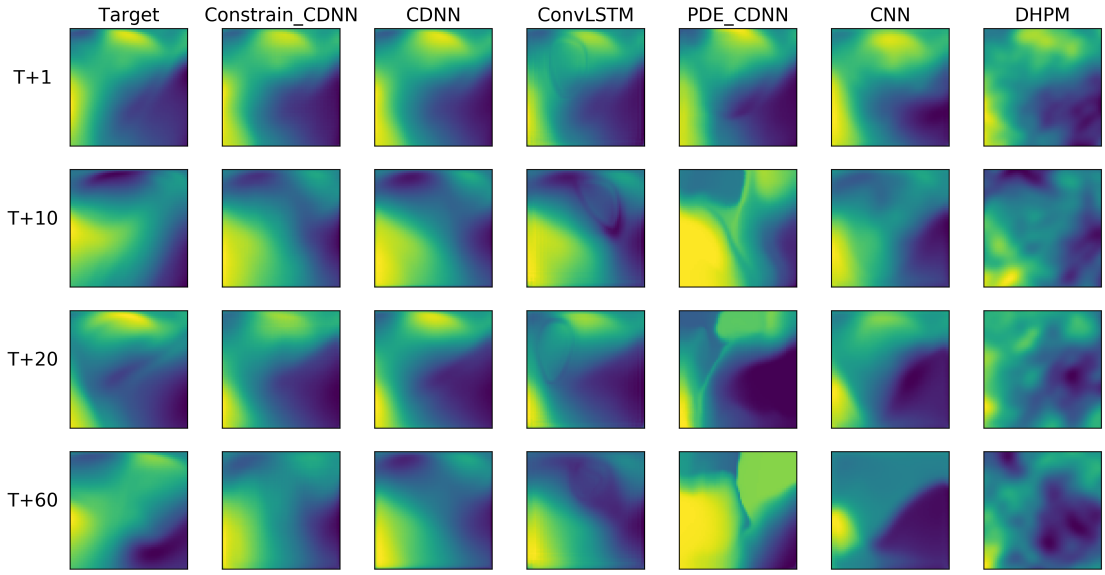


Fig. 5: Visualization of the predictions of velocity fields along x direction at $T+1$, $T+10$, $T+20$, $T+60$, Time T corresponds to the last input velocity field.

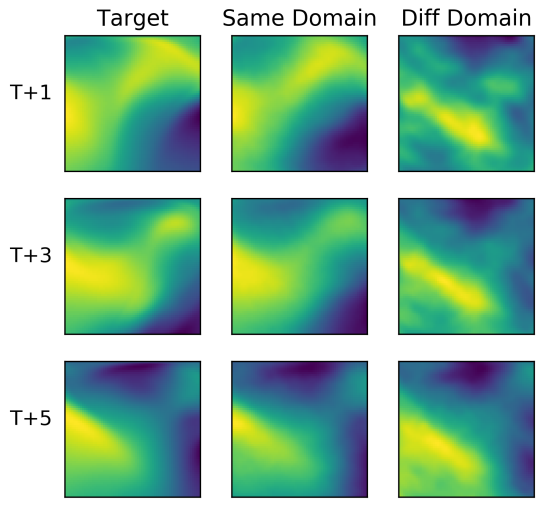


Fig. 6: The first column is the target, the second column is the predictions by the DHPM model trained on a sampled subset from the same domain as the target, and the third column is the predictions obtained by the model trained on a different domain.

Additionally, an interesting finding is that DHPM has poor generalization ability. In [3], we suppose that the reason why their results for fluid flows are good is because they sample the training set and test set from the same domain. The model performs poorly on a different spatial or temporal domain. In our case, the first column in Figure 6 is the target, the second column is the predictions by the model trained on a sampled subset from the same domain as the target, and the third column is the predictions obtained by the model trained on a different domain. It is evident that the model performs poorly when tested out of the domain of training. Dropout and regularization techniques do not improve its performance.

V. CONCLUSIONS AND FUTURE WORK

We compare and contrast several physics-informed DL approaches to the challenging task of predicting the spatio-temporal evolution of a turbulent flow. We found that the constrained CDNN with the continuity equation outperforms other methods. We also found that our simplifying assumptions for the warping scheme of the PDE-CDNN were too limiting. Our next step is to develop approaches that incorporate the momentum and energy equations without these limiting assumptions.

REFERENCES

[1] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," Nature, vol. 566, no. 7743, pp. 195–204, 2019.

[2] J.-L. Wu, K. Kashinath, A. Albert, D. Chirila, Prabhat, and H. Xiao, "Enforcing Statistical Constraints in Generative Adversarial Networks for Modeling Chaotic Dynamical Systems," arXiv e-prints, p. arXiv:1905.06841, May 2019.

[3] M. Raissi, "Deep hidden physics models: Deep learning of nonlinear partial differential equations," Journal of Machine Learning Research, 2018.

[4] P. G. Emmanuel de Bezenac, Arthur Pajot, "Deep learning for physical processes: Incorporating prior scientific knowledge," 2018. https://arxiv.org/pdf/1511.05440.pdf.

[5] C. C. Michael Mathieu and Y. LeCun, "Deep multi-scale video prediction beyond mean square error," 2016. https://arxiv.org/pdf/1511.05440.pdf.

[6] H. W. D.-Y. Y. Xingjian Shi, Zhourong Chen, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," 2015.

[7] P. L. L.-S. S. L. J. Wang, D. Wang, "Lattice boltzmann simulations of thermal convective flows in two dimensions," Computers and Math- ematics with Applications, vol. 65, no. 2, p. 262286, 2013.

[8] D. B. Chirila, Towards lattice boltzmann models for climate sciences: The gelb programming language with applications. PhD thesis, 2018. https://elib.suub.uni-bremen.de/peid/D00106468.html.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

CLIMATENET: BRINGING THE POWER OF DEEP LEARNING TO WEATHER AND CLIMATE SCIENCES VIA OPEN DATASETS AND ARCHITECTURES

Karthik Kashinath[1], Mayur Mudigonda[2], Kevin Yang[2], Jiayi Chen[2], Annette Greiner[1], Prabhat[1]

***Abstract*—Pattern recognition tasks such as classification, object detection and segmentation have remained challenging problems in the weather and climate sciences. While there exist many empirical heuristics for detecting weather patterns and extreme events, the disparities between the output of these different methods even for a single event are large and often difficult to reconcile. Given the success of Deep Learning (DL) in tackling similar problems in computer vision, we advocate a DL-based approach. However, DL works best in the context of supervised learning, when labeled datasets are readily available. Reliable, labeled training data is scarce in climate science. 'ClimateNet' is an effort to solve this problem by creating open, community-sourced expert-labeled datasets that capture information pertaining to class or pattern labels, bounding boxes and segmentation masks. In this paper we present the motivation, design and status of the ClimateNet dataset and associated model architecture.**

I. INTRODUCTION

Climate change is arguably one of the most pressing challenges facing humanity in the 21st century. While summary quantities such as increasing global mean temperature or sea-level rise are important and useful metrics to assess the global extent of climate change, the adverse impacts of our changing climate are most tangible at the local level via extreme weather events such as hurricanes and atmospheric rivers (ARs). Consequently, many nations and governments are considering adaptation and mitigation strategies to improve their resilience to changing extreme weather events and changing local climates. For instance, the state of California receives over 50% of its rainfall through ARs, and water resource managers are interested in understanding if and how AR intensities and tracks will shift in the future; potentially resulting in devastating floods or droughts or both. In the state of Florida, homeowners are interested in understanding if hurricanes will become more intense in the future, and/or make landfall more often. This has direct impact on society and the environment, as well as the economy via home prices and the insurance industry. Hurricanes have caused the US economy over $200B worth of damage in 2017; and a range of stakeholders are interested in a more careful characterization of the change in number, frequency and intensity of such devastating weather phenomena in the coming decades.

In order to address these important questions, climate scientists routinely configure and run high-resolution, high-fidelity simulations under a variety of climate change scenarios. Each simulation produces tens of TBs of output; which requires fast, precise and reliable automated analysis. Thus far, climate scientists have relied upon multi-variate threshold conditions for prescribing extreme weather patterns [1]. However, even within a single class of extreme weather events, there often is no consensus on the precise and accurate definition of that event. For example, the ARTMIP project has shown that across the dozen different AR detection methods and algorithms, results can differ by an order of magnitude [2]. Such methodological disparity is unacceptable for societal and environmental planning.

Since the beginning of this decade, DL has been applied successfully to solve challenging pattern recognition problems in computer vision, speech recognition, robotics and control systems [3], [4]. A key requirement for the success of supervised DL is the availability of plentiful high-quality labelled data. Further, the computer vision community has established that DL is effective at learning relevant features for solving pattern recognition tasks without requiring application-specific tuning. Recent work has demonstrated that DL can be used for solving pattern classification, localization and segmentation problems for climate datasets [5], [6], [7], [8]. [8] demonstrated that model predictions of ARs

Corresponding author: *kkashinath@lbl.gov* [1]Lawrence Berkeley National Laboratory, Berkeley, CA 94720 [2]U C Berkeley

and TCs segmentation masks sometimes exceeded the quality and realism of heuristic-based training data, a rather promising outcome for a first attempt at DL-based segmentation of climate data. The success of the above applications, however, was limited by the lack of plentiful high-quality reliable labeled data.

Given (i) the shortcomings of existing heuristics of detecting weather and climate patterns; (ii) the power of DL in recognizing complex patterns *without* requiring engineered features; and (iii) the scarcity of reliable labeled data; we have developed ClimateNet – a community-sourced labeling strategy to prepare a vast and reliable database of weather and climate pattern labels to push the frontier of DL methods for variety of important and urgent pattern recognition tasks in the weather and climate sciences.

II. CLIMATENET

Figure 1 presents the workflow for augmenting existing weather and climate datasets with ground truth label information. The unified DL workflow, as relevant to weather and climate sciences, can be split into two pieces: the Training phase and the Inference phase. First, the ClimateContours tool is used by human experts to produce a 'hand'-labeled database of weather patterns. Second, a unified DL model is trained to learn a common architecture representing various weather patterns. Finally, the trained model is applied to disparate datasets to obtain predictions.

Recent work in the DL community has shown that training on a multitude of tasks for the same underlying datasets can help with generalization of representations [9]. ClimateNet attempts to provide a similar framework by addressing a multitude of pattern recognition problems in weather and climate data simultaneously.

The goals of ClimateNet are twofold: (i) to create a vast and reliable database of high-quality expert-labeled datasets across a variety of weather and climate phenomena and dataset types, (ii) create and release trained deep neural network architectures, which can be further adapted and customized by the climate science community.

A. Data Ingest

ClimateNet currently takes as input raw climate simulation data from the CAM5.1 climate model [10]. This dataset contains 16 channels corresponding to physical quantities of interest to weather and climate scientists, for example, wind velocities, temperatures, pressures, and humidities at different vertical levels. All of these channels contain information relevant to the dynamics of weather and climate phenomena, but not all variables are needed to detect an event. Based on the experience and wealth of knowledge accumulated by meteorologists and climate scientists, and for relative ease of use, we provide a subset of these variables to the user to aid them in creating labels for TCs and ARs through an online interactive web-based tool called *ClimateContours*.

We are currently developing a framework for users to be able to directly ingest their own datasets and choose the variables of relevance to their specific needs. This will rapidly expand the scope of this tool for diverse datasets and applications, yet preserving the ease of use and variety of labeling options and meta data associated with labels.

B. ClimateContours

ClimateContours adapts the *LabelMe* [11] tool developed at MIT for crowdsourcing the 'hand'-labeling task that is all-important for training DL models. Figure 2 shows a screenshot of the online web-based labeling tool. On the left is integrated water vapor fields with examples of the boundaries of four TCs and three ARs obtained using the Toolkit for Extreme Climate Analysis (TECA), a software that implements expert-specified heuristics to generate label information [1]. These provide initial guesses to guide the user and expedite the labeling process. Users can modify these guesses and add/delete polygons for boundaries of events. Users are presented multiple physical variables for a given timestep (snapshot): vorticity, pressure, integrated water vapor (IWV), wind velocities (U and V at several vertical levels), and integrated vapor transport (IVT), which they toggle between easily to better inform themselves about the characteristics of weather patterns that exist in that snapshot. The right side of figure 2 shows integrated water vapor and wind velocity vectors (on top) and integrated vapor transport, sea level pressure and vorticity contours (on bottom).

A task presented to the user is to draw bounding polygons around all tropical cyclones (TCs) and atmospheric rivers (ARs) that exist in any given snapshot. The user is also given an option to rate their confidence level for each event, in order to obtain information about the reliability of any label. This option helps synthesize useful statistics of labels for probabilistic studies and uncertainty quantification. All of this information is stored in a xml file that is used to create the annotated dataset through post-processing. After numerous discussions with the climate community, we

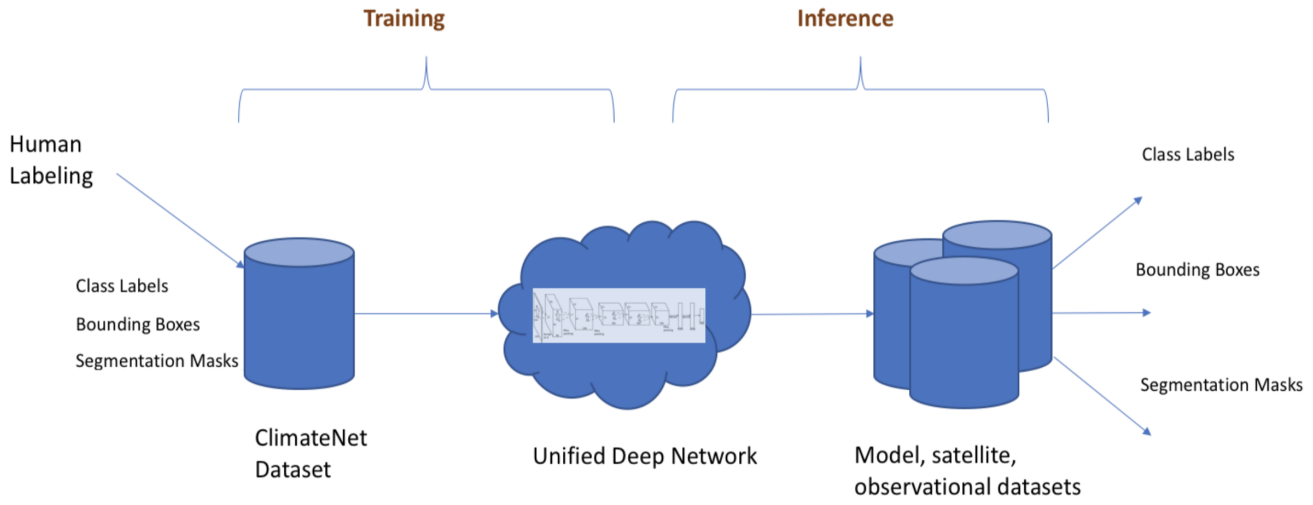


Fig. 1: ClimateNet schematic: Training and Inference phases

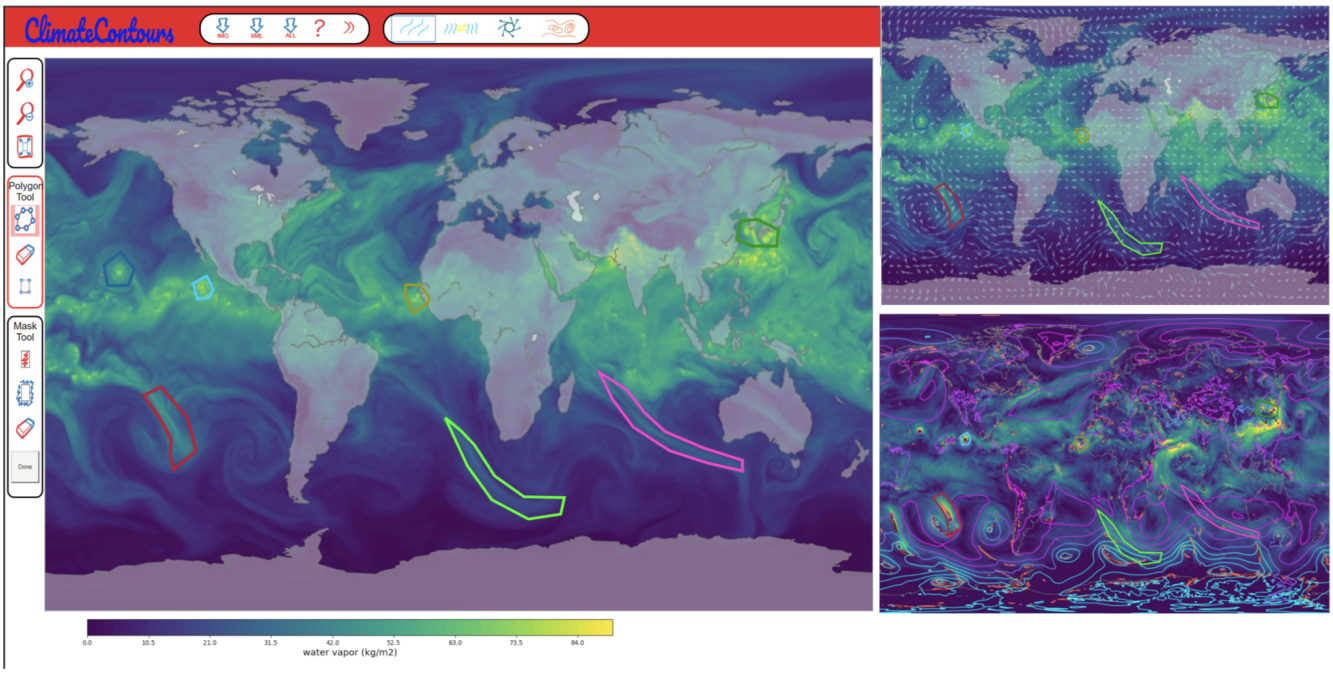


Fig. 2: ClimateContours: online labeling tool. https://climatecontours.nersc.gov/LabelMeAnnotationTool/tool.html

have decided to share the label information separately from the raw data files generated by simulation output.

The ClimateContours tool has been used by a dozen alpha-testers at this point, and we have accumulated a total of approximately 500 labeled images. We plan to conduct labeling campaigns at various universities, national labs and research institutes in the coming months, followed by broader deployment at the AGU and AMS annual meetings.

Further, we plan to expand the number of classes of events to include extra-tropical cyclones, weather fronts, meso-scale convective systems, storms of different types, atmospheric blocks etc., depending on the resolution and reliability of ingested datasets and the specific applications.

C. *Gold standards, noisy labels and quality control*

As part of our quality control we have launched a campaign amongst experts in climate science and meteorology to create a 'gold-standard' dataset, which will be used to evaluate user accuracy and reliability for quality control of labels, and to weed out 'adversaries'. We also intend to utilize user confidence of label quality to improve performance of the models. Further, we

will explore weakly supervised learning methods using noisy labels, an active area of ML research.

D. Model Training and Inference

We are developing a reference implementation of the segmentation architecture developed in [8] in TensorFlow. Once sufficient number of images have been acquired in the ClimateNet dataset, we intend to train the network, and make the network weights available for download on github. We anticipate users leveraging the trained network for 'out-of-the-box' segmentation applications on their datasets, as well as transfer learning applications for datasets with other events (weather fronts, extra-tropical cyclones, etc) and modalities (i.e. observational products).

III. CONCLUSIONS

Our goal is to contribute an end-to-end learning system, including a curated and annotated database, that can segment multiple event classes. We are developing capabilities for enabling users to load their own datasets, which will help expand and diversify our database. We have begun combining forces with EnviroNet, a similar project for a much broader set of applications and problems in the environmental sciences. We believe that easy access to curated datasets and a trained network architecture will be critical in addressing the pattern recognition problems described above and in lowering the barrier of entry for weather and climate scientists who are interested in incorporating DL into their existing workflows. We will report on results from this project at the upcoming meetings of the AGU and AMS.

REFERENCES

[1] M. Prabhat, S. Byna, V. Vishwanath, E. Dart, M. Wehner, and W. Collins, "Teca: Petascale pattern recognition for climate science," vol. 9257, 09 2015.

[2] C. A. Shields, J. J. Rutz, L.-Y. Leung, F. M. Ralph, M. Wehner, B. Kawzenuk, J. M. Lora, E. McClenny, T. Osborne, A. E. Payne, P. Ullrich, A. Gershunov, N. Goldenson, B. Guan, Y. Qian, A. M. Ramos, C. Sarangi, S. Sellars, I. Gorodetskaya, K. Kashinath, V. Kurlin, K. Mahoney, G. Muszynski, R. Pierce, A. C. Subramanian, R. Tome, D. Waliser, D. Walton, G. Wick, A. Wilson, D. Lavers, Prabhat, A. Collow, H. Krishnan, G. Magnusdottir, and P. Nguyen, "Atmospheric river tracking method intercomparison project (ARTMIP): project goals and experimental design," Geoscientific Model Development, vol. 11, no. 6, pp. 2455–2474, 2018.

[3] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," nature, vol. 521, no. 7553, p. 436, 2015.

[4] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," The Journal of Machine Learning Research, vol. 17, no. 1, pp. 1334–1373, 2016.

[5] Y. Liu, E. Racah, J. Correa, A. Khosrowshahi, D. Lavers, K. Kunkel, M. Wehner, W. Collins, et al., "Application of deep convolutional neural networks for detecting extreme weather in climate datasets," arXiv preprint arXiv:1605.01156, 2016.

[6] S. Hong, S. Kim, M. Joh, and S.-k. Song, "Globenet: Convolutional neural networks for typhoon eye tracking from remote sensing imagery," arXiv preprint arXiv:1708.03417, 2017.

[7] E. Racah, C. Beckham, T. Maharaj, S. Ebrahimi Kahou, M. Prabhat, and C. Pal, "Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events," in Advances in Neural Information Processing Systems 30 (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), pp. 3402–3413, 2017.

[8] T. Kurth, S. Treichler, J. Romero, M. Mudigonda, N. Luehr, E. Phillips, A. Mahesh, M. Matheson, J. Deslippe, M. Fatica, et al., "Exascale deep learning for climate analytics," in Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis, p. 51, IEEE Press, 2018.

[9] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," arXiv preprint arXiv:1511.06114, 2015.

[10] M. F. Wehner, K. Reed, F. Li, Prabhat, J. Bacmeister, C.-T. Chen, C. Paciorek, P. Gleckler, K. Sperber, W. D. Collins, A. Gettelman, and C. Jablonowski, "The effect of horizontal resolution on simulation quality in the community atmospheric model, cam5.1.," Journal of Modeling the Earth System, vol. 06, pp. 980–997, 2014.

[11] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," International journal of computer vision, vol. 77, no. 1-3, pp. 157–173, 2008.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

MACHINE LEARNING PARAMETERIZATIONS FOR OZONE: CLIMATE MODEL TRANSFERABILITY

Peer Nowack[1,2,3], Qing Yee Ellie Ong[4], Peter Braesicke[5], Joanna D. Haigh[1,2], Luke Abraham[6], John A. Pyle[6], and Apostolos Voulgarakis[2]

***Abstract*—Many climate modeling studies have demonstrated the importance of two-way interactions between ozone and atmospheric dynamics. However, atmospheric chemistry models needed for calculating changes in ozone are computationally expensive. Nowack et al. [1] highlighted the potential of machine learning-based ozone parameterizations in constant climate forcing simulations, with ozone being predicted as a function of the atmospheric temperature state. Here we investigate the role of additional time-lagged temperature information under preindustrial forcing conditions. In particular, we test if the use of Long Short-Term Memory (LSTM) neural networks can significantly improve the predictive skill of the parameterization. We then introduce a novel workflow to transfer the regression model to the new UK Earth System Model (UKESM). For this, we show for the first time how machine learning parameterizations could be transferred between climate models, a pivotal step to making any such parameterization widely applicable in climate science. Our results imply that ozone parameterizations could have much-extended scope as they are not bound to individual climate models but, once trained, could be used in a number of different models. We hope to stimulate similar transferability tests regarding machine learning parameterizations developed for other Earth system model components such as ocean eddy modeling, convection, clouds, or carbon cycle schemes.**

I. MOTIVATION

While being a greenhouse gas and air pollutant, ozone is also the only absorber of harmful solar UV-B radiation which would otherwise make life on Earth impossible [2]. Ozone's distribution in the atmosphere is constantly affected by anthropogenic and natural factors, from changes in the stratospheric circulation [3, 4] to chemical reactions [5–8]. Its importance for global radiative transfer in turn induces feedback effects on the Earth system by modulating temperature, dynamics and the biosphere [9–14]. A number of previous studies used machine learning methods to understand and to model factors influencing ozone, e.g. to forecast air quality [15, 16], to model ozone dry deposition [17] and iodide surface emissions [18], or to infer differences among chemistry models [19].

Here we explore the potential to use machine learning-based ozone parameterizations in constant forcing climate model simulations. Specifically, we focus on preindustrial simulations, which are core experiments in climate modeling intercomparsions [20, 21] and which are typically run for centennial to millennial time-scales. Atmospheric chemistry schemes add substantially to the overall high computational costs of such simulations, mainly because they require repeated numerical approximations to the transport equations of dozens of chemical tracers as well as to the large system of coupled chemical rate equations [22]. In preindustrial simulations, ozone's impact on climate is primarily determined by two-way interactions between the variability in ozone and climate. This is particularly true for the stratosphere, where changes in ozone have been found to modulate the Quasi-Biennial Oscillation [QBO, 23] or the polar vortices [24, 25].

Our paper is motivated by results presented in Nowack et al. [hereafter N2018, 1] who showed that, in certain simulations, the global ozone distribution could be well predicted by temperature-based machine learning regression functions. Here, we first revisit those results and test if further time-lagged temperature information improves the parameterization. We then investigate for the first time how such a parameterization could be transferred among climate models.

II. METHOD

As in N2018, we fit regression models that predict daily-mean ozone distributions at timestep t based on the previous day's temperature distribution. This time resolution is sufficient to capture the large-scale behaviour of the relatively slowly-moving stratosphere, where interactions between ozone and climate are par-

Corresponding author: Peer Nowack, p.nowack@imperial.ac.uk
[1]Grantham Institute, Imperial College London, UK. [2]Department of Physics, Imperial College London, UK. [3]Data Science Institute, Imperial College London, UK. [4]Department of Physics, University of Oxford, UK. [5]IMK-ASF, Karlsruhe Institute of Technology, Germany. [6]National Centre for Atmospheric Science and Department of Chemistry, University of Cambridge, UK.

ticularly important [e.g. 23]. More formally, we attempt to predict ozone mixing ratios $\mathbf{Y}$ in every model grid cell k of a global climate model as a function f of the global temperature state $\mathbf{X}$

$$Y_k^{(t)} = f(\mathbf{X}^{(t-1)}) \quad (1)$$

Below we further test the importance of lagged temperature information ($\mathbf{X}^{(t-2)}$, ..., $\mathbf{X}^{(t-\tau_{\max})}$).

Climate model data. The primary preindustrial climate model temperature and ozone data was produced using the HadGEM3-AO model from the UK Met Office [26], coupled to the atmospheric chemistry scheme UKCA [4, 27]. In these preindustrial simulations, atmospheric CO_2 is held at 285 ppmv. We use a continuous 50-year long time slice of daily mean ozone ($\mathbf{Y}_{\text{train}}$) and temperature ($\mathbf{X}_{\text{train}}$) data for training and cross-validation. Predictions (Figure 1) were then made on an independent 13-year long test set ($\mathbf{X}_{\text{test}}$, $\mathbf{Y}_{\text{test}}$). For the second part of this paper where we develop the model transferability workflow, we used 20-year long preindustrial temperature and ozone datasets ($\mathbf{X}_{\text{UKESM}}$, $\mathbf{Y}_{\text{UKESM}}$) produced by the new United Kingdom Earth System Model [UKESM, 28] for the Coupled Model Intercomparison Project Phase 6 [CMIP6, 21]. All UKESM data was bi-linearly interpolated to the HadGEM3-AO horizontal grid and we select identical vertical levels for the two models. All temperature data was pre-processed by removing each grid cell's training data mean $\mu_{\text{train},k}$ and by scaling to approximately unit variance

$$\mathrm{X}^{\text{norm}}_{\text{train,k}} = \frac{\mathrm{X}_{\text{train,k}} - \mu_{\text{train,k}}}{\sigma_{\text{train},k}} \quad (2)$$

$$\mathrm{X}^{\text{norm}}_{\text{test,k}} = \frac{\mathrm{X}_{\text{test,k}} - \mu_{\text{train,k}}}{\sigma_{\text{train},k}} \quad (3)$$

$$\mathrm{X}^{\text{norm}}_{\text{UKESM,k}} = \frac{\mathrm{X}_{\text{UKESM,k}} - \mu_{\text{train,k}}}{\sigma_{\text{train},k}} \quad (4)$$

Regression models and cross-validation. Following N2018, we use Ridge regression as the baseline approach for the temperature-ozone mapping. Ridge regression is a linear least squares regression augmented by L2-regularization to address the bias-variance tradeoff [29]. The cost function

$$J_k = \sum_{t=1}^{N} \left(Y_k^{(t)} - \sum_{j=1}^{p} c_{kj} X_{\text{pca},j}^{(t-1)} \right)^2 + \alpha \sum_{j=1}^{p} c_{kj}^2 \quad (5)$$

is minimized for each model grid cell k over N timesteps. We applied principal component analysis [PCA, 30] to $\mathbf{X}$ to speed up the training procedure. We here retain the first p=1000 components (equalling >95% represented variance; for extensive numerical tests see N2018 Supplementary Figure S1a) as a compromise between model complexity and model performance. Smaller (larger) values of α put weaker (stronger) constraints on the size of the coefficients, thus favoring overfitting (high bias). We use a standard time series cross validation method to find the best value for α, in which the time-ordered training data is split into five subsets of equal size. Preceding subsets are then sequentially used as training data for each subsequent subset (i.e. set 1 for 2, set 1+2 for 3 etc.). α is found according to the average generalization error.

We compare the performance of the Ridge regressions to Long Short-Term Memory (LSTM) neural networks, which can process time-lagged information highly effectively [31, 32]. We tested a range of bias and recurrent regularization parameter values in a non-stateful setting, varied the number of timesteps accounted for in the memory unit (up to 10 days), and also used different network architectures (one vs. two LSTM layers with up to 100 neurons per layer). All network architectures were fitted with ReLU activation functions for the LSTM layers and linear activations directed towards the output layer. In each case, we trained the network for 750 epochs using varying batch sizes (>256). The number of epochs was chosen according to their respective error learning curves as to avoid overfitting. To assert stability, we scaled all inputs/outputs to within (0,1)/(-1,1) range. Only the best settings after cross-validation on a 40-to-10 year training data split are discussed below. For all data pre-processing and regression tasks, we used Python's scikit-learn, keras and tensorflow packages [33, 34].

III. RESULTS

HadGEM3-AO test data results. Figure 1 shows four examples of ozone time series in different areas of the stratosphere[1]: in the tropical upper stratosphere (Figure 1a), where ozone concentrations are mainly determined by local photochemical reactions [5, 35], the tropical lower stratosphere where the longer time-scales of the Brewer-Dobson circulation and QBO pose the primary control mechanisms [Figure 1b; 3, 36], the mid-latitude

[1]There are on the order of 420,000 grid cells in HadGEM3-AO and more than 2,300,000 grid cells in UKESM. In the following, mainly to keep the computational expense of training the regression models in bay and to describe our approach intuitively, we focus our discussion on individual climate model grid cells which were found to be characteristic of the general method's performance in different atmospheric regimes. For global performance metrics using larger sets of grid cells see the extensive numerical results in the Supplementary Material of N2018.

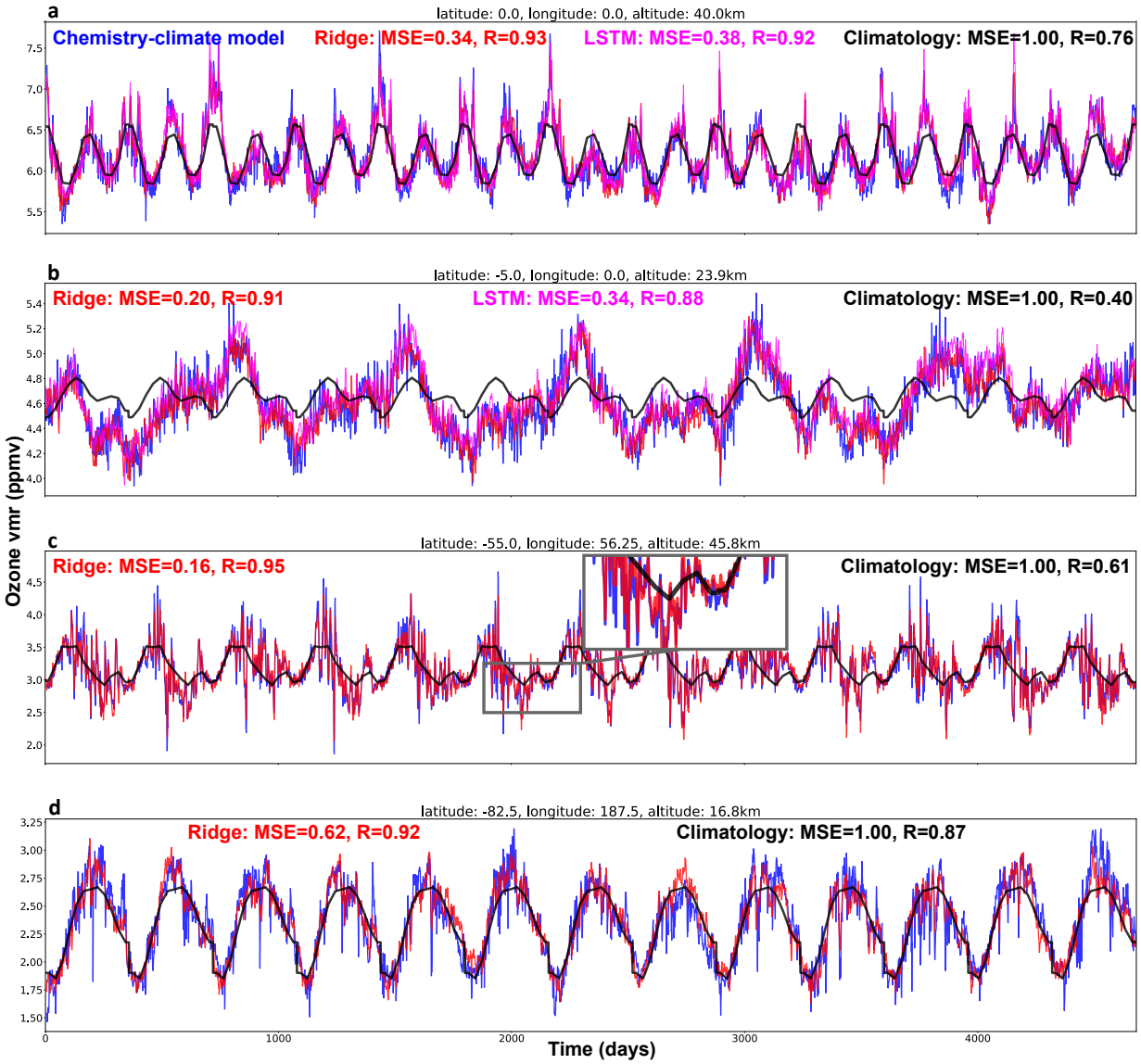


Fig. 1. Ozone time series for four climate model grid cells characteristic of different atmospheric regimes. (Black) Preindustrial climatology. (Blue) Ozone as simulated by the fully interactive chemistry-climate model. (Red) Ridge regression predictions. In the top two panels, we additionally show the results of LSTM predictions (magenta). Grid coordinates are as labeled. R is the Pearson correlation coefficient between the chemistry-climate model time series and the predictions. *MSE* is the ratio of mean squared errors of machine learning predictions relative to the chemistry-climate model data, divided by the same error for the climatologies (values <1 imply improvement).

upper stratosphere where seasonal wave-breaking is important [Figure 1c; 3] and the Southern Hemisphere lower stratosphere where the periodic break-down of the polar vortex is a key feature [Figure 1d; 8]. For each grid cell, we show the actual chemistry-climate model data (blue lines), the corresponding predictions using Ridge regressions (red) and seasonal ozone climatologies (black). For the first two grid cells we also show LSTM predictions (magenta), which here keep a temperature memory of up to five days. Intuitively, LSTMs could be useful as not only the present state of the atmosphere is important for ozone's distribution but also its history, e.g. in the lower stratosphere where ozone's lifetime is much longer than one day. Ozone climatologies (black) are frequently used in climate model simulations and here represent 50-year monthly-mean averages of $\mathbf{Y}_{\text{train}}$, which were subsequently linearly interpolated to daily time resolution. The climatology serves as a benchmark for a classic treatment of ozone in preindustrial simulations without interactive atmospheric chemistry. Due to the limited predictability of short-term ozone fluctuations one day in advance, see e.g. zoomed-in area in Figure 1c, we are interested in ozone predictions that broadly represent the state of ozone at any given time relative to this benchmark (see also discussion in N2018). We evaluate the predictive

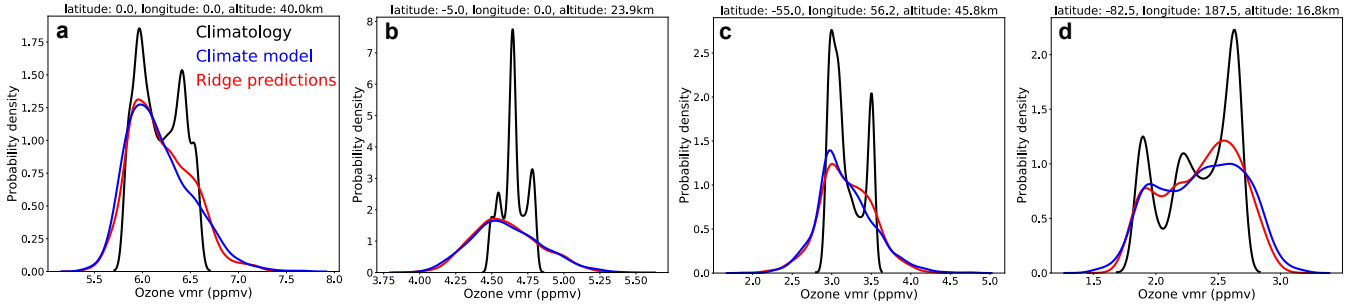


Fig. 2. Probability (kernel) density estimates for ozone volume mixing ratios in each of the four grid cells also shown in Figure 1.

skill relative to these climatologies through Pearson correlation coefficients (R) and mean squared error (MSE) ratios (details in caption of Figure 1), when matched against the actual chemistry-climate model data. These statistical results are shown inside each panel of Figure 1.

The Ridge regressions provide a good approximation to the true chemistry-climate model results, as is also evident from their respective ozone mixing ratio density estimates (Figure 2). Both MSE and R are improved in every grid cell relative to a model-consistent preindustrial climatology; in particular in the tropical mid-stratosphere where the QBO dominates and ozone variability is not captured well by an annual climatology (Figure 1b). Interestingly, the LSTM regressions do not provide a significant improvement over the Ridge regression approach despite their greater functional complexity and longer memory of the temperature history. Further performance gains might be achieved by more extensive parameter and network architecture tuning. However, the general result underlines previous findings in N2018, where other non-linear algorithms were also not found to perform better than Ridge regression. To further validate this result, we also carried out 5-day-lagged window Ridge regressions (not shown). We thus conclude that the single time-step Ridge regression is a solid method on this particular learning task and turn to the question of model transferability.

Model transferability. There are two main obstacles to transferring the parameterization from one climate model to another: differences in (a) the temperature fields and (b) the ozone fields. For example, even a constant background temperature difference, as will always occur among climate models [37], can seriously interfere with the regression task as the inputs may, for example, constantly take on extreme values relative to the original model's temperature distribution. This in turn renders the machine learning parameterization unable to make realistic ozone predictions.

To mitigate such effects, we found that the following intuitive re-calibration procedure led to good results: after standardization of the UKESM temperature data according to equation (4), the resulting $\mathbf{X}^{\text{norm}}_{\text{UKESM,k}}$ will still have average values in each grid cell significantly different from nil, simply due to the aforementioned time-average discrepancies in the background temperature state. We thus re-calibrated the temperature data to an approximately zero mean by subtracting the average offset over the first n number of years ($\overline{\mathbf{X}}^{\text{norm}}_{\text{UKESM,k}}$)

$$\mathbf{X}^{\text{adjusted}}_{\text{UKESM}} = \mathbf{X}^{\text{norm}}_{\text{UKESM}} - \overline{\mathbf{X}}^{\text{norm}}_{\text{UKESM}} \tag{6}$$

Empirically, we found that calibrations using the first n=5 or 10 years yielded almost identical results. In addition, we applied a re-calibration procedure to the predicted ozone field. We separate the ozone mixing ratios predicted by the regression (trained on HadGEM3-AO data) when provided with the re-calibrated temperature input from the UKESM model ($\mathbf{X}^{\text{adjusted}}_{\text{UKESM}}$) into a climatological plus a variability term for each cell

$$Y_{\text{HadGEM-consistent,k}} = Y_{\text{HadGEM-clim,k}} + Y_{\text{variability,k}} \tag{7}$$

We then use again n years of UKESM data to approximate a corresponding climatological term $Y_{\text{UKESM-clim}}$ which we use to replace the HadGEM climatology

$$Y_{\text{UKESM-consistent,k}} = Y_{\text{UKESM-clim,k}} + Y_{\text{variability,k}} \tag{8}$$

Figure 3 shows the results of the corresponding regressions (red) for three grid cells over the last ten years of the UKESM dataset, located in the tropical lower and mid-stratosphere as well as in the Northern Hemisphere polar stratospheric region. In addition, we show ozone climatologies from HadGEM3-AO (gray) and calculated from the first 5 years of the UKESM data (black). As before, we use the MSE ratio and correlation coefficients to compare the predictions to the actual UKESM interactive chemistry climate model results (blue). The quantitative results are given directly in each panel of Figure 3. The transferred Ridge regression performs typically far better than the climatologies

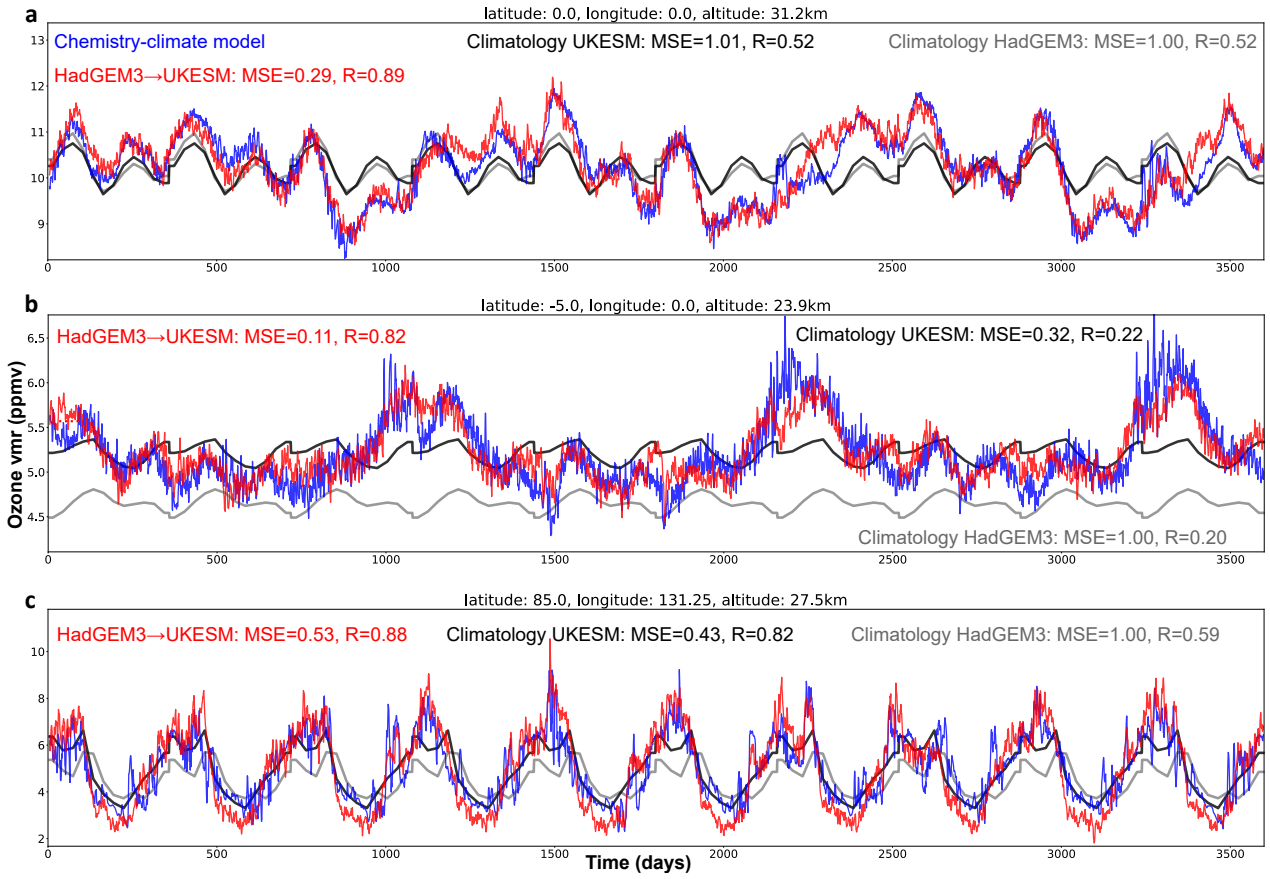


Fig. 3. As Figure 1, but for ten years of UKESM data using the re-calibrated Ridge regressions for predictions (red). To calculate the MSE ratio, we choose the errors when using the HadGEM3 climatology averaged over 50 years as the baseline (as in Figure 1).

both in terms of MSE and R. In terms of the MSE ratio, the HadGEM3-AO climatology performs particularly poorly in regions with major differences in background ozone levels between the two climate models (Figure 3b). Finally, note that a better than climatological performance is not guaranteed in this setting. For example in Figure 3c, where a UKESM-consistent climatology already achieves R-values greater than 0.8, the MSE is worse for the transferred regression, mainly due to constantly underestimated annual minimum values using the Ridge regression. Such a feature may even occur if we just used another atmospheric chemistry scheme. Its annually repeating consistency could also reflect changes in the underlying transport time-scales (i.e. dynamics) between the two models.

IV. DISCUSSION

We have presented two important extensions to temperature-based machine learning parameterizations:

1) Using LSTMs, which here took into account past temperature state information of up to 5-10 days, we tested for improved predictive skill. However, the results are comparable to those obtained with computationally cheaper Ridge regressions.

2) An intuitive re-calibration procedure to learn the ozone parameterization from data produced by one climate model (HadGEM3-AO) to then predict ozone values for another climate model (UKESM). The re-calibration is required for the temperature inputs, whereas the ozone correction can be chosen more flexibly, e.g. to adjust the field to a climatology of the modeler's choice. For example, in some cases there might be no ground truth ozone state (e.g. in models without interactive chemistry option) so that one might either choose ozone values consistent with HadGEM3-AO, or another background climatological field. Our method could therefore in principle be used to remove some persistent ozone model biases [see e.g. 38]. In conclusion, the method outlined here could be used to learn regression functions representing ozone variability from long existing chemistry-climate model datasets. Following re-calibration, the ozone parameterization could then be used in long climate model simulations such as under constant preindustrial forcing.

ACKNOWLEDGMENTS

Peer Nowack is supported through an Imperial College Research Fellowship. Qing Yee Ellie Ong's work on the transferability of the ozone parameterization was funded by the Laidlaw Scholars Undergraduate Research and Leadership Programme.

REFERENCES

[1] P. Nowack, P. Braesicke, J. Haigh, N. L. Abraham, J. Pyle, and A. Voulgarakis, "Using machine learning to build temperature-based ozone parameterizations for climate sensitivity simulations," *Environmental Research Letters*, vol. 13, no. 10, p. 104016, 2018.

[2] WMO, "World Meteorological Organization: Scientific assessment of ozone depletion: 2010, Global Ozone Research and Monitoring Project - Report No. 52, 516. pp, Geneva, Switzerland," 2011.

[3] SPARC, "SPARC CCMVal Report on the Evaluation of Chemistry-Climate Models," tech. rep., SPARC, 2010.

[4] P. J. Nowack, N. Luke Abraham, A. C. Maycock, P. Braesicke, J. M. Gregory, M. M. Joshi, A. Osprey, and J. A. Pyle, "A large ozone-circulation feedback and its implications for global warming assessments," *Nature Climate Change*, vol. 5, no. 1, pp. 41–45, 2015.

[5] J. D. Haigh and J. A. Pyle, "Ozone perturbation experiments in a two-dimensional circulation model," *Quart. J. R. Met. Soc.*, vol. 108, pp. 551–574, 1982.

[6] A. I. Jonsson, J. de Grandpré, V. I. Fomichev, J. C. McConnell, and S. R. Beagley, "Doubled CO2-induced cooling in the middle atmosphere: Photochemical analysis of the ozone radiative feedback," *Journal of Geophysical Research*, vol. 109, p. D24103, 2004.

[7] A. Voulgarakis, V. Naik, J. F. Lamarque, D. T. Shindell, P. J. Young, M. J. Prather, O. Wild, R. D. Field, D. Bergmann, P. Cameron-Smith, I. Cionni, W. J. Collins, S. B. Dalsøren, R. M. Doherty, V. Eyring, G. Faluvegi, G. A. Folberth, L. W. Horowitz, B. Josse, I. A. MacKenzie, T. Nagashima, D. A. Plummer, M. Righi, S. T. Rumbold, D. S. Stevenson, S. A. Strode, K. Sudo, S. Szopa, and G. Zeng, "Analysis of present day and future OH and methane lifetime in the ACCMIP simulations," *Atmospheric Chemistry and Physics*, vol. 13, no. 5, pp. 2563–2587, 2013.

[8] S. Solomon, D. J. Ivy, D. Kinnison, M. J. Mills, R. R. Neely, and A. Schmidt, "Emergence of healing in the Antarctic ozone layer," *Science*, vol. 353, no. 6296, pp. 269–274, 2016.

[9] S. Sitch, P. M. Cox, W. J. Collins, and C. Huntingford, "Indirect radiative forcing of climate change through ozone effects on the land-carbon sink.," *Nature*, vol. 448, no. 7155, pp. 791–794, 2007.

[10] C. Williamson and R. Zepp, "Solar ultraviolet radiation in a changing climate," *Nature Climate Change*, vol. 4, pp. 434–441, 2014.

[11] P. J. Nowack, "Cirrus and the Earth system," *Weather*, vol. 70, no. 11, p. 330, 2015.

[12] P. J. Nowack, N. L. Abraham, P. Braesicke, and J. A. Pyle, "Stratospheric ozone changes under solar geoengineering: implications for UV exposure and air quality," *Atmospheric Chemistry and Physics*, vol. 16, pp. 4191–4203, 2016.

[13] P. J. Nowack, P. Braesicke, N. L. Abraham, and J. A. Pyle, "On the role of ozone feedback in the ENSO amplitude response under global warming," *Geophysical Research Letters*, vol. 44, pp. 3858–3866, 2017.

[14] P. J. Nowack, N. L. Abraham, P. Braesicke, and J. A. Pyle, "The impact of stratospheric ozone feedbacks on climate sensitivity estimates," *Journal of Geophysical Research: Atmospheres*, vol. 123, pp. 4630–4641, 2018.

[15] V. Mallet and B. Sportisse, "Ensemble-based air quality forecasts: A multimodel approach applied to ozone," *Journal of Geophysical Research: Atmospheres*, vol. 111, p. D18302, 2006.

[16] C. A. Keller and M. J. Evans, "Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10," *Geoscientific Model Development*, vol. 12, pp. 1209–1225, 2019.

[17] S. J. Silva, C. L. Heald, S. Ravela, I. Mammarella, and J. W. Munger, "A Deep Learning Parameterization for Ozone Dry Deposition Velocities," *Geophysical Research Letters*, vol. 46, no. 2, pp. 983–989, 2019.

[18] T. Sherwen, R. J. Chance, L. Tinel, D. Ellis, M. J. Evans, and L. J. Carpenter, "A machine learning based global sea-surface iodide distribution," *Earth System Science Data*, vol. 11, pp. 1239–1262, 2019.

[19] J. M. Nicely, R. J. Salawitch, T. Canty, D. C. Anderson, S. R. Arnold, M. P. Chipperfield, L. K. Emmons, J. Flemming, V. Huijnen, D. E. Kinnison, J. F. Lamarque, J. Mao, S. A. Monks, S. D. Steenrod, S. Tilmes, and S. Turquety, "Quantifying the causes of differences in tropospheric OH within global models," *Journal of Geophysical Research*, vol. 122, no. 3, pp. 1983–2007, 2017.

[20] B. Kravitz, A. Robock, O. Boucher, H. Schmidt, K. E. Taylor, G. Stenchikov, and M. Schulz, "The Geoengineering Model Intercomparison Project (GeoMIP)," *Atmospheric Science Letters*, vol. 12, no. 2, pp. 162–167, 2011.

[21] V. Eyring, S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor, "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization," *Geoscientific Model Development*, vol. 9, no. 5, pp. 1937–1958, 2016.

[22] E. Esentürk, L. Abraham, S. Archer-Nicholls, C. Mitsakou, P. Griffiths, and J. Pyle, "Quasi-Newton Methods for Atmospheric Chemistry Simulations : Implementation in UKCA UM Vn10 . 8," *Geoscientific Model Development*, vol. 11, no. February, pp. 3089–3108, 2018.

[23] D. Rind, J. Jonas, N. K. Balachandran, G. A. Schmidt, and J. Lean, "The QBO in two GISS global climate models: 1. Generation of the QBO," *Journal of Geophysical Research: Atmospheres*, vol. 119, no. 14, pp. 8798–8824, 2014.

[24] V. Silverman, N. Harnik, K. Matthes, S. W. Lubis, and S. Wahl, "Radiative effects of ozone waves on the Northern Hemisphere polar vortex and its modulation by the QBO," *Atmospheric Chemistry and Physics*, vol. 18, pp. 6637–6659, 2018.

[25] S. Haase and K. Matthes, "The importance of interactive chemistry for stratosphere-troposphere coupling," *Atmospheric Chemistry and Physics*, vol. 19, no. 5, pp. 3417–3432, 2019.

[26] H. T. Hewitt, D. Copsey, I. D. Culverwell, C. M. Harris, R. S. R. Hill, A. B. Keen, A. J. McLaren, and E. C. Hunke, "Design and implementation of the infrastructure of HadGEM3: The next-generation Met Office climate modelling system," *Geoscientific Model Development*, vol. 4, no. 2, pp. 223–253, 2011.

[27] O. Morgenstern, P. Braesicke, F. M. O'Connor, A. C. Bushell, C. E. Johnson, S. M. Osprey, and J. A. Pyle, "Evaluation of the new UKCA climate-composition model Part 1: The stratosphere," *Geoscientific Model Development*, vol. 2, no. 1, pp. 43–57, 2009.

[28] A. A. Sellar et al., "UKESM1: Description and evaluation of the UK Earth System Model," *Submitted to Journal of Advances in Modeling Earth Systems*, 2019.

[29] A. E. Hoerl and R. W. Kennard, "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[30] C. M. Bishop, *Pattern recognition and machine learning*. Singapore: Springer Science+Business Media, 2006.

[31] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[33] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[34] A. Martin et al., "TensorFlow: A System for Large-Scale Machine Learning," 2015.

[35] J. A. Pyle, "A calculation of the possible depletion of ozone by chlorofluorocarbons using a two-dimensional model," *Pure and Applied Geophysics PAGEOPH*, vol. 118, pp. 355–377, 1980.

[36] R. A. Plumb, "Stratospheric Transport," *Journal of the Meteorological Society of Japan*, vol. 80, pp. 793–809, 2002.

[37] G. Flato et al., "Evaluation of Climate Models," in *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*, ch. 9, pp. 741–866, Cambridge, UK and New York, USA: Cambridge University Press, 2013.

[38] S. C. Hardiman et al., "Processes controlling tropical tropopause temperature and stratospheric water vapor in climate models," *Journal of Climate*, vol. 28, pp. 6516–6535, 2015.

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

HETEROSCEDASTIC GAUSSIAN PROCESS REGRESSION ON THE ALKENONE OVER SEA SURFACE TEMPERATURES

Taehee Lee[1], Charles E. Lawrence[1]

***Abstract*—To restore the historical sea surface temperatures (SSTs) better, it is important to construct a good calibration model for the associated proxies. In this paper, we introduce a new model for alkenone ($U_{37}^{K'}$) based on the heteroscedastic Gaussian process (GP) regression method. Our nonparametric approach not only deals with the variable pattern of noises over SSTs but also contains a Bayesian method of classifying potential outliers.**

I. INTRODUCTION

The alkenone is a widely used proxy for inferring sea surface temperatures (SSTs) in paleoceanography. Haplophytes make long-chain ketone lipids (alkenones) with 37 carbons changing in response to water temperature. Let $U_{37}^{K'}$ be the relative unsaturation index of these compounds as [1]:

$$U_{37}^{K'} = \frac{C_{37:2}}{C_{37:2} + C_{37:3}} \tag{1}$$

Because SSTs have other sources of inferences, what we are interested in is the distribution of $U_{37}^{K'}$ given SST, not the reverse: once the prior information (based on the latitude, for example) of SST is organized as a prior distribution and the distribution as a likelihood, it is possible to integrate those information by computing the posterior distribution of SST given the observed $U_{37}^{K'}$ by the Bayes' rule. Also, if SSTs are somehow correlated (with respect to their spatial pattern, for example) one another, the given proxies can be exploited better than a set of individual posteriors with an associated graphical model.

[1] has lead the way in the application of Bayesian statistics in paleoclimatology. The paper approaches to the problem with a Bayeian B-spline regression model (BAYSPLINE) on $U_{37}^{K'}$ data as well as considering their seasonality, and it shows improved performance to extant methods. However, the model lacks of a concrete rule of deciding the number of basis splines or their orders, and does not deal with heteroscedastic noises over SSTs neatly.

For the last decade, models based on the neural networks (NN) have been rapidly emerging and soon overwhelming all research fields requiring machine learning, including climatology, for the outstanding performance. Though some theoretical works such as [2], [3] and [4] guarantee the effectiveness of NNs in some senses, it remains as a potential problem in practice that parameters to be learned are often too many compared to the amount of available data. Bayesian statistics can give a possible solution of this problem: its philosophy that parameters are random variables allows to marginalize all such unknown values out to get the posterior predictive distributions.

In this point of view, among various nonparametric regression models, Gaussian processes (GP) ([5]) have some advantages which make them distinctive from all the others. Besides their explicit predictive posterior distributions, with some specific kernels corresponding to the choice of activation functions, it becomes a marginalized version of associated NN models: details can be found in [6] and [7]. The point is that only a very few hyperparameters (for example, only three hyperparameters are needed in the squared-exponential kernel) are now controlling the GP regression models, and let data explain themselves to make them free from the structure.

Though tuning hyperparameters can be done efficiently in the homoscedastic (i.e., noises are assumed to have a constant variance.) GP models ([5]), heteroscedastic GP models are not yet clear to learn. While [8] tries to model logarithms of variances from empirical variances based on another GP regression, [9] adapts a variational method which cannot avoid breaking up the completeness of the models. In this paper we suggest a heteroscedastic GP model which is not only intuitive but also easy to learn and apply it to

[1]Division of Applied Mathematics, Brown University, Rhode Island, USA

constructing a new calibration for $\mathrm{U}_{37}^{\mathrm{K}'}$.

There is one golden rule: every Gaussian model is sensitive to outliers in learning. Also, we should be aware of the model misspecification for the robustness, unless the model is guaranteed to follow a certain distribution based on the underlying theoretical arguments. GP models cannot be free from these limitations. Student's t-processes ([10]) can be an alternative, but they do not have explicit posterior predictive distributions if noises are also considered. Instead, classifying and excluding outliers in learning seems to be a better idea. We also introduce a Bayesian method of classifying and treating outliers automatically in the learning procedure.

II. METHOD

Let $\mathcal{X} = \{x_n\}_{n=1}^{\mathrm{N}}$ and $\mathcal{Y} = \{y_n\}_{n=1}^{\mathrm{N}}$ be the SSTs and $\mathrm{U}_{37}^{\mathrm{K}'}$ observations, respectively. What to construct is the regression model which returns the distribution $p\left(y|x,\mathcal{X},\mathcal{Y}\right)$ of $\mathrm{U}_{37}^{\mathrm{K}'}$, y, at an arbitrary query SST, x. Let β be the regression function of $\mathcal{Y}$ at $\mathcal{X}$. Then, what we assume are the following:

$$p\left(\mathcal{Y}|\beta,\mathcal{X}\right) = \mathcal{N}\left(\mathcal{Y}|\beta,\Lambda(\mathcal{X})\right) \tag{2}$$

$$p\left(\beta|\mathcal{X}\right) = \mathcal{GP}\left(\beta\middle|\vec{0},\mathbb{K}\left(\mathcal{F}(\mathcal{X}),\mathcal{F}(\mathcal{X})\right)+\xi^2\mathbb{I}\right) \tag{3}$$

, where Λ is a function returning variances of $\mathrm{U}_{37}^{\mathrm{K}'}$ observations in the form of a diagonal matrix at the query SSTs given their regression vector, $\mathcal{F}$ is a feature function, and $\mathbb{K}$ is the GP kernel function defined as follows: for a kernel function k,

$$\mathbb{K}(\mathcal{X}^{(1)},\mathcal{X}^{(2)}) \triangleq \left[k\left(\mathcal{F}(\mathcal{X}_m^{(1)}),\mathcal{F}(\mathcal{X}_n^{(2)})\right)\right]_{m,n} \tag{4}$$

, where $\mathrm{A} \triangleq \mathrm{B}$ means that A is defined by B.

Because feature functions may not be injective, the term $\xi^2\mathbb{I}$ is inevitable to guarantee (3) to be well-defined by making it nondegenerate.

Then, for a scalar regression b of y at x, the regression distribution $p\left(y|x,\mathcal{X},\mathcal{Y}\right)$ is derived as follows:

$$\begin{aligned} &p\left(y|x,\mathcal{X},\mathcal{Y}\right) \\ &= \int p\left(y|b,x\right)\int p\left(b|\beta,x,\mathcal{X}\right)p\left(\beta|\mathcal{X},\mathcal{Y}\right)d\beta db \end{aligned} \tag{5}$$

As consistent with (2), we have the likelihood of y given b and x as a Gaussian $p\left(y|b,x\right) = \mathcal{N}\left(y|b,\Lambda(x)\right)$.

Because the density of regression vector given SSTs is assumed to be a GP, we have the following extension:

$$\begin{aligned} &p\left(\beta,b|\mathcal{X},x\right) \\ &= \mathcal{GP}\left(\begin{pmatrix}\beta\\ b\end{pmatrix}\middle|0,\begin{pmatrix}\mathbb{K}_{11}+\xi^2\mathbb{I} & \mathbb{K}_{12}\\ \mathbb{K}_{21} & \mathbb{K}_{22}+\xi^2\end{pmatrix}\right) \end{aligned} \tag{6}$$

, $\mathbb{K}_{11} \triangleq \mathbb{K}\left(\mathcal{F}\left(\mathcal{X}\right),\mathcal{F}\left(\mathcal{X}\right)\right)$, $\mathbb{K}_{22} \triangleq \mathbb{K}\left(\mathcal{F}\left(x\right),\mathcal{F}\left(x\right)\right)$, $\mathbb{K}_{12} \triangleq \mathbb{K}\left(\mathcal{F}\left(\mathcal{X}\right),\mathcal{F}\left(x\right)\right)$ and $\mathbb{K}_{21} \triangleq \mathbb{K}\left(\mathcal{F}\left(x\right),\mathcal{F}\left(\mathcal{X}\right)\right)$ are the abbreviations. Thus, the conditional distribution of b given β, $\mathcal{X}$ and x can be computed explicitly as a Gaussian.

Finally, the posterior distribution of β given $\mathcal{X}$ and $\mathcal{Y}$ is derived from the likelihood (2) and prior (3), which is also a Gaussian.

I.e., three terms in (5) are all Gaussian, so $p\left(y|x,\mathcal{X},\mathcal{Y}\right)$ can be computed analytically as follows: note that $\left(\mathbb{K}_{11}+\xi^2\mathbb{I}+\Lambda\left(\mathcal{X}\right)\right)^{-1}$ is not a function of x or y, so only one matrix inversion is required to compute the following μ and ν.

$$\begin{aligned} &p\left(y|x,\mathcal{X},\mathcal{Y}\right) \\ &\quad= \mathcal{N}\left(y|\mu\left(x,\mathcal{X},\mathcal{Y}\right),\Lambda\left(x\right)+\nu\left(x,\mathcal{X},\mathcal{Y}\right)\right) \end{aligned} \tag{7}$$

$$\begin{aligned} \mu\left(x,\mathcal{X},\mathcal{Y}\right) &\triangleq \mathbb{K}_{21}\left(\mathbb{K}_{11}+\xi^2\mathbb{I}+\Lambda\left(\mathcal{X}\right)\right)^{-1}\mathcal{Y} \\ \nu\left(x,\mathcal{X},\mathcal{Y}\right) &\triangleq \mathbb{K}_{22}+\xi^2 \\ &\quad-\mathbb{K}_{21}\left(\mathbb{K}_{11}+\xi^2\mathbb{I}+\Lambda\left(\mathcal{X}\right)\right)^{-1}\mathbb{K}_{12} \end{aligned} \tag{8}$$

Heteroscedasticity comes from the choice of Λ: if it is defined to be a constant function, the model becomes homoscedastic. Some papers, such as [8], suggest modelling Λ as another GP regression on the logarithms of the residues. This approach assumes that such logarithms follow Gaussian distributions so symmetric, and always underestimates variances by the Jensens inequality applied to the concave log function: the average of logarithms is always smaller than or equal to the logarithm of average!

Here, we use a more intuitive and direct form of Λ inspired by Nadaraya-Watson kernel regression ([11], [12], [13] and [14]) as follows: let $\mu_n \triangleq \mu\left(x_n,\mathcal{X},\mathcal{Y}\right)$ and $\nu_n \triangleq \nu\left(x_n,\mathcal{X},\mathcal{Y}\right)$ as abbreviations.

$$\Lambda\left(x\right) \triangleq \frac{\sum_{n=1}^{\mathrm{N}}\left(\left(y_n-\mu_n\right)^2+\nu_n\right)\mathcal{K}_h\left(x-x_n\right)}{\sum_{n=1}^{\mathrm{N}}\mathcal{K}_h\left(x-x_n\right)} \tag{9}$$

, where $\mathcal{K}$ is a density kernel and h is a tuning parameter which is called the bandwidth. I.e., $\Lambda\left(x\right)$ is defined as a weighted average of squares of residues, where weights are determined by how much the query SST x is departed from the data. (9) is derived from the following expectation over $\beta|\mathcal{X},\mathcal{Y}$:

$$\mathbb{E}_{\beta|\mathcal{X},\mathcal{Y}}\left[\frac{\sum_{n=1}^{\mathrm{N}}\left(y_n-\beta_n\right)^2\mathcal{K}_h\left(x-x_n\right)}{\sum_{n=1}^{\mathrm{N}}\mathcal{K}_h\left(x-x_n\right)}\right] \tag{10}$$

The choice of $\mathcal{K}$ does not substantially affect to the regression model, but the model does substantially depend on the value of h. One suggestion is to adapt the K-nearest neighbor bandwidth ([15] and [16]).

A GP with arcsine kernel in (11) is interpretable as a one-layer neural network with infinite number of marginalized hidden nodes with a sigmoid function as the activation ([6]). Now we have a scalar η and a square matrix Σ as hyperparameters to be tuned. To avoid overfitting, it is common to assume in addition that Σ is a diagonal matrix. Tuning hyperparameters can be done by maximizing the marginal likelihood (ML) $p(\mathcal{Y}|\mathcal{X})$ from (2) and (3) or by the leave-one-out cross-validation (LOO-CV): more details can be found in [5].

$$k_{\mathrm{NN}}(x,\tilde{x}) \triangleq \eta^2 \sin^{-1}\left(\frac{2f^{\mathbb{T}}\Sigma\tilde{f}}{\sqrt{\left(1+2f^{\mathbb{T}}\Sigma f\right)\left(1+2\tilde{f}^{\mathbb{T}}\Sigma\tilde{f}\right)}}\right) \quad (11)$$

, where $f \triangleq \mathcal{F}(x)$ and $\tilde{f} \triangleq \mathcal{F}(\tilde{x})$.

In our model, outliers are represented as a hidden variable $\mathcal{H} = \{\mathrm{H_n}\}_{n=1}^{\mathrm{N}}$, where $\mathrm{H_n} = 0$ if y_n is an inlier and 1 an outlier. $\mathrm{H_n}$'s are assumed to be independent and each has the following prior distribution:

$$p(\mathrm{H_n}) = \mathrm{Bernoulli}(\mathrm{H_n}|q) \quad (12)$$

, where $q \in (0,1)$ is a hyperparameter not to be learned. A small q implies that most of the observations are believed not to be outliers. This reflects a point of view that outliers are in essence subjective.

Once $\mathcal{H}$ is given, we specified each $\mathrm{U}_{37}^{\mathrm{K}'}$ observation in $\mathcal{Y}$ as follows: let $\Lambda_n \triangleq \Lambda(x_n)$ as abbreviation.

$$p(y_n|x_n,\mathrm{H_n}=0) \triangleq \mathcal{N}(y_n|\mu_n,\nu_n+\Lambda_n) \quad (13)$$

$$p(y_n|x_n,\mathrm{H_n}=1) \triangleq \frac{1}{2}\mathcal{N}\left(y_n\middle|\mu_n+d\sqrt{\nu_n+\Lambda_n},\nu_n+\Lambda_n\right) + \frac{1}{2}\mathcal{N}\left(y_n\middle|\mu_n-d\sqrt{\nu_n+\Lambda_n},\nu_n+\Lambda_n\right) \quad (14)$$

, where $d > 0$ is a hyperparameter for how much outliers are deviated from the inlier distribution. Note that the above outlier classification is working because outputs are defined on the one-dimensional space in this problem.

By considering (12) as the prior and (13) and (14) as the likelihoods given H_n, respectively, we derive the posterior distribution of H_n given x_n and y_n by Bayes' rule:

$$p(\mathrm{H_n}|x_n,y_n) \propto p(\mathrm{H_n})\,p(y_n|x_n,\mathrm{H_n}) \quad (15)$$

Now we are prepared. Algorithm 1 summarizes the learning procedure of our heteroscedastic GP regression (HGPR) on $\mathrm{U}_{37}^{\mathrm{K}'}$ over SSTs.

Algorithm 1: HGPR on $\mathrm{U}_{37}^{\mathrm{K}'}$ over SSTs

```
initialize hyperparameters and H ≡ 0.
while convergence do
    tune kernel hyperparameters in (3) and (11) with X, Y | H = 0.
    compute μ and ν in (8) with X, Y | H = 0.
    choose the bandwidths in (9).
    update Λ in (9) with X, Y | H = 0.
    sample H | X, Y by (15).
end
return μ, ν and Λ.
```

One possible way of utilizing the obtained GP regression model in (7) is plugging it in the following Bayesian inversion:

$$p(\tilde{x}|\tilde{y}) \propto p(\tilde{y}|\tilde{x})\,p(\tilde{x}) \quad (16)$$

, where $\tilde{y}$ is the observed $\mathrm{U}_{37}^{\mathrm{K}'}$ data and $\tilde{x}$ is the associated SSTs to be inferred. The prior $p(\tilde{x})$ can be given as a distribution of SSTs given their geographical information, for instance.

In general, (16) does not have a closed form. Because SSTs are of one dimension, applying the Markov-chain Monte Carlo (MCMC) is enough to sample from the posterior $p(\tilde{x}|\tilde{y})$; if the event to infer is of high dimension, a variational inference to approximate the posterior with a known distribution must be considered.

III. DATA

We used the dataset same with [1], after discarding those in the locations that it excluded from their analysis. The rest of data with 1274 observations were used. We could discard more from them but did not do so for checking whether or not our model could classify apparent outliers desirably. Details about data are in [1].

For the density kernel and bandwidth in (9), we adapted a Gaussian kernel having an unbounded support and the K-nearest neighbor bandwidth where K is selected by the LOO-CV among 10 candidates, from 1% to 10% of the data per 10 iterations. Hyperparameters q in (12) and d in (13) and (14) were set to be 0.065 and 2.48, respectively: these values lead the marginal likelihood of y_n from (12), (13) and (14) to approximating the Student's t-distribution with 6 as the degree of freedom.

We also regularized the raw data $\mathcal{X}$ and $\mathcal{Y}$ by $x \leftarrow x/6.5656 - 3.0205$ and $y \leftarrow y/0.2104 - 3.3642$ so that they fit to the Student's t-distribution with 6 as the degree of freedom, and then took a feature function

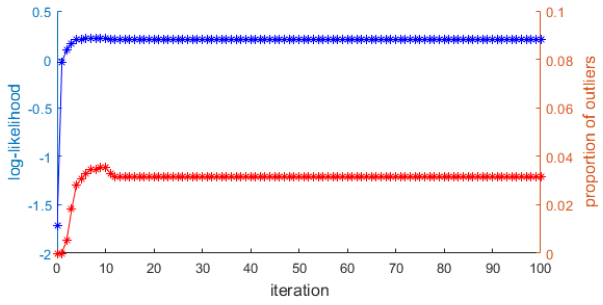


Fig. 1. Average log-likelihoods of inliers and proportions of outliers over iterations.

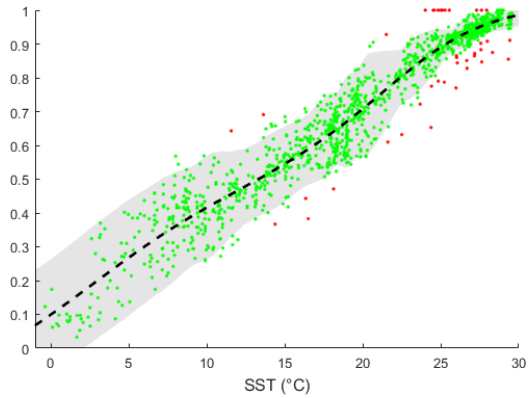


Fig. 2. The learned regression over SSTs. Green and red points are inliers and outliers, respectively. The dashed line shows the inferred means over SSTs and shaded region is the 95% confidence band.

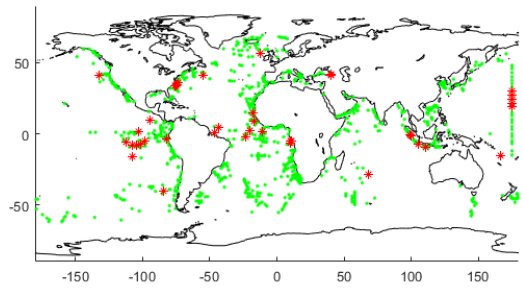


Fig. 3. The world map with data locations. Green and red points are inliers and outliers, respectively.

$\mathcal{F}(x) = (1, x)^{\mathbb{T}}$ on the regularized $\mathcal{X}$. To identify and regularize the model, we defined Σ in (11) to be a diagonal matrix, where the first entry is fixed to be 1.

IV. Results

After 100 iterations in about 7 minutes, we obtained the results shown in the eight figures, from 1 to 8. Finally selected K was 4%. Figure 1 shows the average log-likelihoods of inliers (blue) and proportions of inferred outliers (red). It shows convergence of the learning procedure after around the 15^{th} iteration. Only about 3% of the data were classifed as outliers. Learned kernel hyperparameters are the following:

$$\begin{aligned} \eta &= 2.1962 \\ \Sigma &= \begin{bmatrix} 1 & 0 \\ 0 & 0.4712^2 \end{bmatrix} \\ \xi &= 2.7253 \times 10^{-12} \end{aligned} \tag{17}$$

Figure 2 represents the inferred regression on SSTs with the classification of inliers (green) and outliers (red), and figure 3 those on the world map. It clearly shows the heteroscedasticity of observations and the associated model, as figure 4. Also, apparent outliers at 22-26°C of the upper boundary of the plot are classified as so. Classified outliers above 25°C below the model, however, could be from another cluster. This can be adjusted by tuning hyperparameters in section III. One advantage of GP regression is that the inferred mean at each input is expressed in a closed form and differentiable as much as the adopted kernel. Curvatures of the inferred means are shown in figure 5. [1] suggests that slopes of the means start to change at 23.4°C, which is roughly consistent with our inference, but some more changes are also captured in the inferred model.

To visually check how residuals are treated appropriately by our heteroscedastic model, figures 6 to 8 were also plotted. A normalized residual r_n at x_n is defined as follows:

$$r_n \triangleq \frac{y_n - \mu_n}{\sqrt{\nu_n + \Lambda_n}} \tag{18}$$

I.e., if means and variances of the GP regression model are properly inferred, normalized residuals at inliers must follow the standard normal distribution.

In figure 6, most of the normalized residuals seem to follow the standard normal distribution, and figure 7 supports that assertion as the Q-Q plot of inliers. In addition, figure 8 shows that normalized residuals are barely correlated with SSTs: the correlation between the pairs classified as inliers is 0.0019. These results strongly suggest that our GP regression model appropriately infer the heteroscedasticity of the model.

V. Conclusion

Our GP regression on $\mathrm{U}_{37}^{\mathrm{K}'}$ appropriately explains the heteroscedasticity of data and provides a probabilistic model with explicit distributions. It converges quickly,

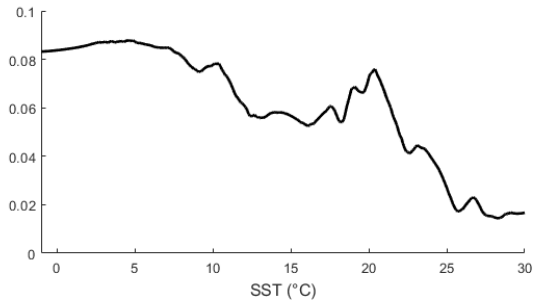


Fig. 4. Inferred standard deviations over SSTs.

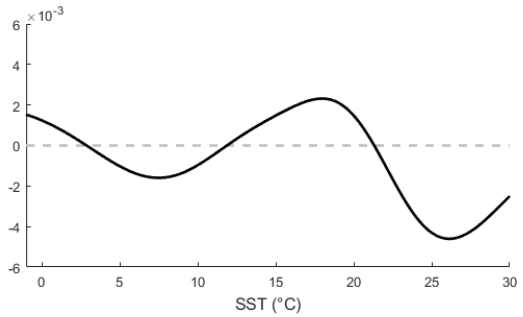


Fig. 5. Inferred curvatures over SSTs.

in about 7 minutes, even if we do not assume any informative prior structure (see $\vec{0}$ in (3)) on the regression. Thus, this work can be considered as a success of the nonparametric approach on the real data.

However, a GP has several disadvantages to overcome. Firstly, it requires at least one matrix inversion in constructing distributions, where its size is the same as the number of observations. There are only 1274 data, which is affordable in the currently available computing power, so it was not problematic in this case. We used two-dimensional features, which do not suffer from the curse of dimensionality.

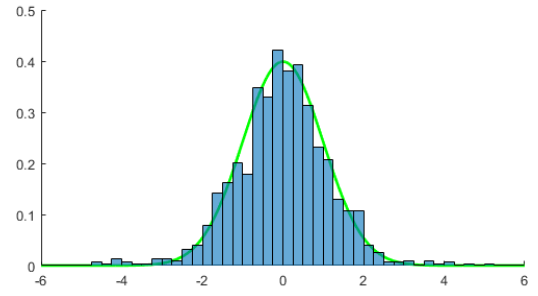


Fig. 6. The histogram of normalized residuals. The green graph is the standard normal distribution.

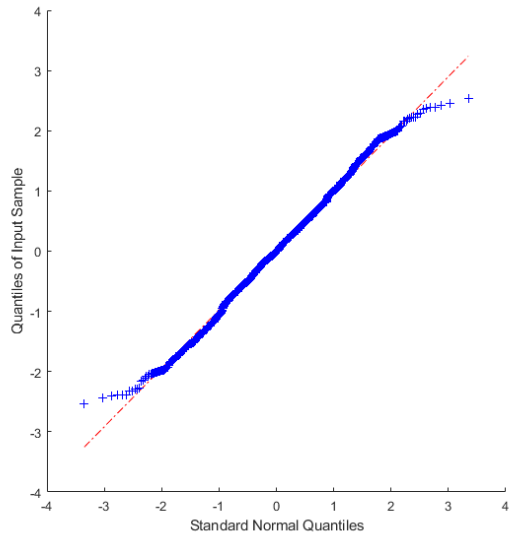


Fig. 7. The Q-Q plot of inliers.

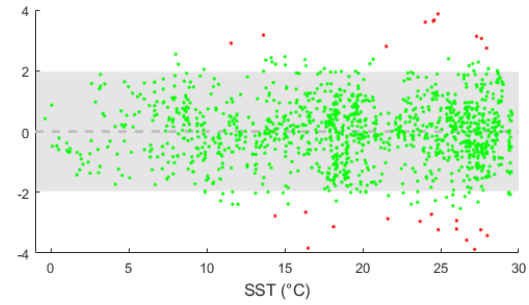


Fig. 8. A plot of normalized residuals over SSTs. Green and red points are inliers and outliers, respectively. The shaded region covers the 95% confidence intervals $[-1.96, 1.96]$.

Nonetheless, in this paper a model based on GPs shows its effectiveness on explaining $U_{37}^{K'}$ observations over SSTs, where relatively moderate amount of data are given and inputs are of low-dimensional. One more advantage of the nonparametric approaches is that it does not depend much on the specific characteristics of data: it is possible to adapt the same approach to any data to do regression.

Codes which run on MATLAB can be found in https://github.com/eilion/HGPR_SST_Proxy_Cal.

ACKNOWLEDGMENTS

This paper is based on the works supported by the Division of Applied Mathematics in Brown University, by the National Science Foundation (NSF) under a grant number OCE-1760838, and by the Kwanjeong Educational Foundation.

REFERENCES

[1] J. E. Tierney and M. P. Tingley, "Bayspline: A new calibration for the alkenone paleothermometer," *Paleoceanography and Paleoclimatology*, vol. 33, no. 3, pp. 281–301, 2018. 1, 3, 4

[2] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp. 303–314, 1989. 1

[3] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Networks*, vol. 4, no. 2, pp. 251–257, 1991. 1

[4] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang, "The expressive power of neural networks: A view from the width," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (USA), pp. 6232–6240, Curran Associates Inc., 2017. 1

[5] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press, 2005. 1, 3

[6] C. K. I. Williams and D. Barber, "Bayesian classification with gaussian processes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998. 1, 3

[7] Y. Cho and L. K. Saul, "Large-margin classification in infinite neural networks," *Neural Computation*, vol. 22, no. 10, pp. 2678–2697, 2010. 1

[8] K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard, "Most likely heteroscedastic gaussian process regression," in *Proceedings of the 24th International Conference on Machine Learning*, pp. 393–400, 2007. 1, 2

[9] M. Lázaro-Gredilla and M. K. Titsias, "Variational heteroscedastic gaussian process regression," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, pp. 841–848, 2011. 1

[10] A. Shah, A. Wilson, and Z. Ghahramani, "Student-t Processes as Alternatives to Gaussian Processes," in *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, vol. 33, pp. 877–885, 2014. 2

[11] E. A. Nadaraya, "On estimating regression," *Theory of Probability and its Applications*, vol. 9, no. 1, pp. 141–142, 1964. 2

[12] G. S. Watson, "Smooth regression analysis," *Sankhya: The Indian Journal of Statistics, Series A*, vol. 26, pp. 359–372, 01 1964. 2

[13] H. J. Bierens, *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge University Press, 1994. 2

[14] N. Langren and X. Warin, "Fast and stable multivariate kernel density estimation by fast sum updating," *Journal of Computational and Graphical Statistics*, vol. 28, no. 3, pp. 596–608, 2019. 2

[15] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, no. 3, pp. 1049–1051, 1965. 2

[16] G. R. Terrell and D. W. Scott, "Variable kernel density estimation," *Ann. Statist.*, vol. 20, no. 3, pp. 1236–1265, 1992. 2

9th International Workshop on Climate Informatics
October 02-04, 2019
Hosted by École Normale Supérieure, Paris, France

DOWNSCALING NUMERICAL WEATHER MODELS WITH GANS

Alok Singh[1], Adrian Albert[1], Brian White[1]

***Abstract*—Numerical simulation of weather is resolution-constrained due to the high computational cost of integrating the coupled PDEs that govern atmospheric motion. For example, the most highly-resolved numerical weather prediction models are limited to approximately 3 km. However many weather and climate impacts occur over much finer scales, especially in urban areas and regions with high topographic complexity like mountains or coastal regions. Thus several statistical methods have been developed in the climate community to downscale numerical model output to finer resolutions. This is conceptually similar to image super-resolution (SR) [1] and in this work we report the results of applying SR methods to the downscaling problem. In particular we test the extent to which a SR method based on a Generative Adversarial Network (GAN) can recover a grid of wind speed from an artificially downsampled version, compared against a standard bicubic upsampling approach and another machine learning based approach, SR-CNN [1]. We use ESRGAN ([2]) to learn to downscale wind speeds by a factor of 4 from a coarse grid. We find that we can recover spatial details with higher fidelity than bicubic upsampling or SR-CNN. The bicubic and SR-CNN methods perform better than ESRGAN on coarse metrics such as MSE. However, the high frequency power spectrum is captured remarkably well by the ESRGAN, virtually identical to the real data, while bicubic and SR-CNN fidelity drops significantly at high frequency. This indicates that SR is considerably better at matching the higher-order statistics of the dataset, consistent with the observation that the generated images are of superior visual quality compared with SR-CNN.**

I. MOTIVATION

Global climate models are limited to ~100 km resolution, while numerical weather prediction models that produce daily forecasts and severe weather warnings are limited to ~3 km. However, accurate assessment of climate and extreme weather impacts near human populations would benefit substantially from finer resolution. While several methods have been developed for downscaling climate models output to finer resolutions, they consist for the most part of complex interpolation methods (see e.g. [3]). In this paper we explore a machine learning method to downscale weather model output using a Generative Adversarial Network (GAN) developed originally for the purpose of image super-resolution (ESRGAN).

Machine learning approaches have only recently started to receive attention in the earth sciences community [4]. Over traditional numeric-based approaches, they could address some key issues in climate modeling:

1) A ML pipeline trained end-to-end that automatically learns optimal filters and transformations between inputs (i.e., remote-sensing, in-situ, and simulation data) and their relationship to the spatiotemporal estimate of parameters of interest (e.g., wind intensity, precipitation), can drastically accelerate the creation of practical ad-hoc relationships between observational datasets and models.
2) An end-to-end differentiable model will allow for the exploration of climate model sensitivities that lead to bias including the influence of the meteorological forcing dataset. Slight perturbations in precipitation phase, intensity, and/or location, shortwave and longwave radiation, wind speed and direction, humidity, and, temperature, could be used to understand the downstream implications on variables that are difficult to measure or model directly.
3) Generative models can be run in parallel, and do not necessarily require iterative schemes to model data, allowing them to run quickly even on low-grade consumer hardware.

II. METHODS AND DATA

A. Dataset

We use 15 years of wind velocity fields from a numerical simulation of the WRF (Weather Research and Forecasting) model over Southern California (see [5] for details on model setup - we use region d04 from

[1]Terrafuse, Inc

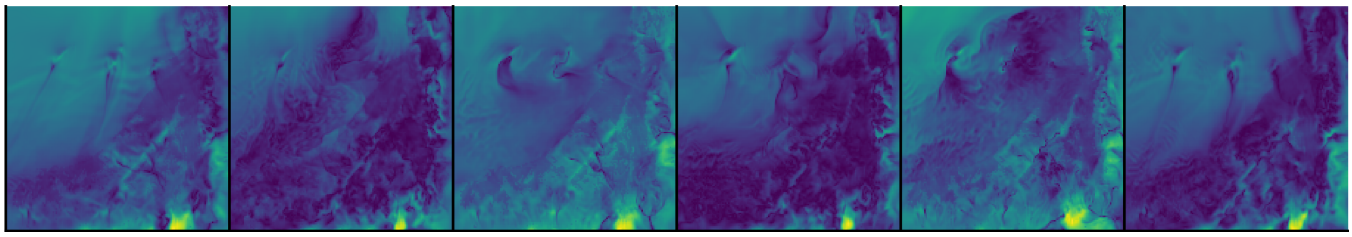
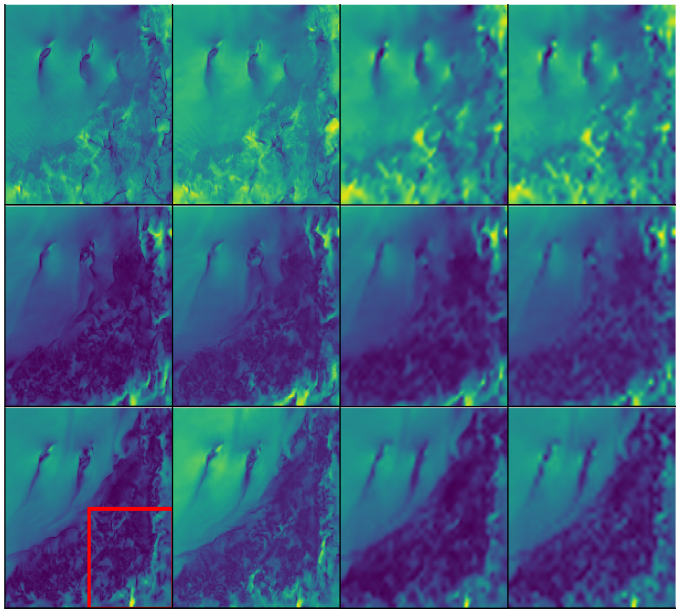
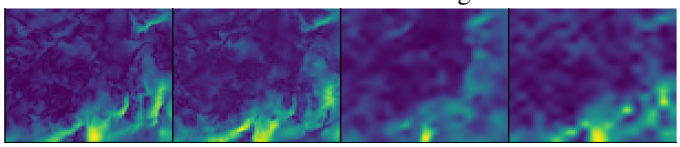


Fig. 1: Example wind speeds from the WRF model output, shown every 4 hours over 1 day.



(a) Real vs. ESRGAN vs. SR-CNN vs. Bicubic Upsampling. Each row is for one image.



(b) Zoomed view of marked region above.

2001-2015), gathered hourly, for a total of about 60,000 data points, represented as grids. In the rest of this work, we shall call these grids "images".

Each image is of size 153×153, where each pixel represents the average wind speed over a 1.5 km $\times$ 1.5 km region. They are stored as a 2-D array of 32-bit floats, linearly scaled to $[0, 1]$ to be compatible with image processing frameworks. For ease of upsampling, we clip a single row and column to resize our images to 152×152. We also combine the south to north and east to west components of the wind vector to model total wind speed rather than modeling each direction's velocity separately.

The dataset is then shuffled and split, with 5% (3,000 images) held out for validation.

B. Method

Note that all experiments were run on a stock HP Z420 with an NVIDIA GeForce RTX 2070 GPU.

1) Bicubic Upsampling: Bicubic interpolation is a common algorithm used for upsampling, useful as a baseline to compare against. There are no trainable parameters, so we simply upsample our validation set.

2) SR-CNN: SR-CNN is a popular deep-learning based approach to image SR. A low-resolution image is first upsampled to the desired size by another method, such as bicubic interpolation. It's then passed through a CNN, which outputs an image that is compared against the ground truth image via mean-squared error (MSE). (See 3 for a high-level pictorial overview.)

We train for 100 epochs with a batch size of 128, the Adam optimizer, and a learning rate of 0.001. For the upsampling step, we use bicubic upsampling. Validation is performed at the end of training.

The model has about 8,000 trainable parameters, a relatively small number by modern standards, so training is quick. It can process about 730 images per second, so training takes about 2 hours with an RTX Titan 2070.

3) ESRGAN: ESRGAN is an optimized version of SRGAN, which we shall describe here. SRGAN is a conditional GAN designed for image SR. Its training procedure involves passing the generator G a batch of low-resolution images, which are upsampled by G and then passed to the discriminator D. D is also given the ground truth images for the batch, and attempts to distinguish between them. An optimization particular to SRGAN is that it also has a "content loss", where in addition to the discriminator, there is another network, typically a CNN pretrained on ImageNet. This "feature network" passes both the generated and ground truth images through it, and then both are compared via MSE. The idea is that a pretrained network will have captured the higher-level dynamics of perceptual similarity. However, since our images do not represent natural images and we do not have an equivalent of ImageNet to perform supervised learning on, it is inappropriate to use content loss, so we remove it. Otherwise, our

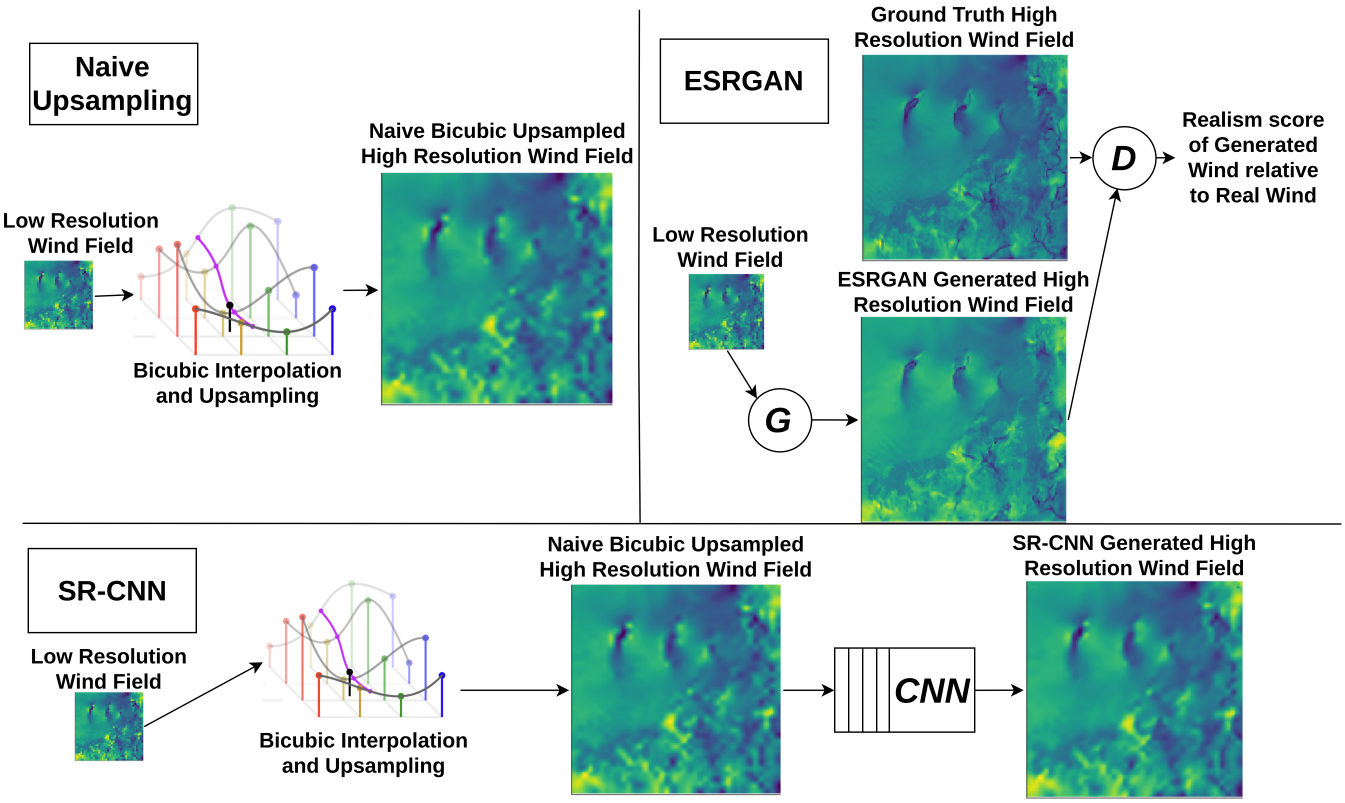


Fig. 3: Overview of models used.

TABLE I: Results

Model	PSNR	MSE	MAE	KL
ESRGAN	32.74	0.00053	0.0148	0.008
SR-CNN	**36.06**	**0.00024**	**0.0091**	0.015
Upsampling	35.52	0.00027	0.0097	**0.006**

training is virtually identical to the methodology outlined in [2].

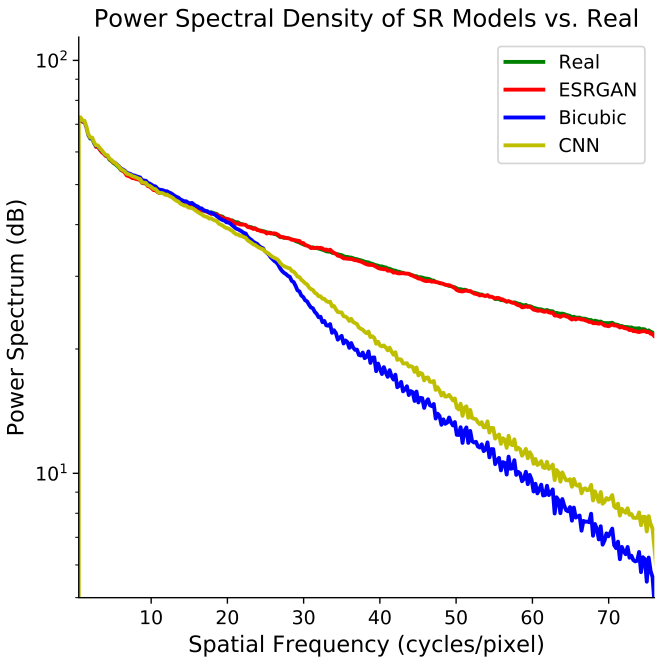
We use a batch size of 12 images and train for 100 epochs. The model has about 25 million trainable parameters, so training is far slower than SR-CNN. Each epoch takes about 35 minutes to complete, and training takes about 2.5 days.

III. EVALUATION

Table I gives an overview of final performance on the validation set. PSNR (peak signal to noise ration), MSE and MAE (mean absolute error) are averaged over all images in the validation set. "KL" represents the KL divergence between the empirical distributions of the generated images and the ground truth.

Notice that SR-CNN performs best on most metrics, and ESRGAN the worst.

Fig. 4: Note that ESRGAN tracks the true data far more closely than the other models.



But the generated images tell a quite different story. As we see in Figure 2a, ESRGAN generates clearer images than the other methods. Zooming in on the highlighted red box (Figure 2b) reveals that the image generated by ESRGAN are sharper and less prone to artifacts. PSNR and MSE have been noted in other works to be poor indicators of image quality, as they fail to capture the underlying dynamics of images well.

A key metric that illustrates the spatial resolution and higher moments of the data distribution is the power spectral density, shown in Figure 4. The power spectrum reveals the power of ESRGAN in capturing the high frequency information present in the wind field. In fact, remarkably ESRGAN's spectrum is so close to that of the true data that they are nearly indistinguishable, whereas SR-CNN and bicubic upsampling fall off significantly at higher frequencies. This is perhaps not surprising as the upsampling and SR-CNN are fundamentally methods of interpolation, whereas ESRGAN is learning the data distribution at all scales. These results suggest that ESRGAN is able to capture far more of the underlying spatial structure, and that while SR-CNN may be doing better than bicubic upsampling, it is not learning the distribution of the true data, but rather that of the upsampled version.

IV. FUTURE WORK

As seen in this work, ESRGAN does a good job reproducing single images. However, we do not currently deal with a sequence of images over time or space, even though both are useful. For example, knowing the wind of surrounding regions helps in prediction, as does knowing what the wind was like an hour ago.

We also plan to incorporate additional variables such as temperature and pressure, and to see if models based on attention mechanisms can improve accuracy..

REFERENCES

[1] T. Vandal, E. Kodra, S. Ganguly, A. R. Michaelis, R. R. Nemani, and A. R. Ganguly, "Deepsd: Generating high resolution climate change projections through single image super-resolution," *CoRR*, vol. abs/1703.03126, 2017.

[2] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," September 2018.

[3] G. Bürger, S. Sobie, A. Cannon, A. Werner, and T. Murdock, "Downscaling extremes - an intercomparison of multiple methods for future climate," *Journal of Climate*, vol. 26, 2013.

[4] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

[5] P. Vahmani and A. D. Jones, "Water conservation benefits of urban heat mitigation," *Nature Communications*, vol. 8, no. 1, p. 1072, 2017.