JMLR: Workshop and Conference Proceedings 1:1–10, 2020 NeurIPS2019 Competition & Demonstration Track PMLR Post-proceedings

The Causality for Climate Challenge

Jakob Runge JAKOB.RUNGE@DLR.DE
Xavier-Andoni Tibau XAVIER.TIBAU@DLR.DE
Matthias Bruhns MATTHIAS.BRUHNS@DLR.DE
German Aerospace Center, Institute of Data Science, Mälzerstr. 3, 07745 Jena, Germany

Jordi Muñoz-Marí JORDI.MUNOZ@UV.ES
Gustau Camps-Valls GUSTAU.CAMPS@UV.ES
Image Processing Lab (IPL), Universitat de València, C/ Cat. J. Beltrán 2, 46980 València, Spain

Editor: Raia Hadsell and Hugo Jair Escalante

Abstract

Understanding the complex interdependencies of processes in our climate system has become one of the most critical challenges for society with our main current tools being climate modeling and observational data analysis, in particular observational causal discovery. Causal discovery is still in its infancy in Earth sciences and a major issue is that current methods are not well adapted to climate data challenges. We here present an overview of a NeurIPS 2019 competition on causal discovery for climate time series. The *Causality 4 Climate* (C4C) competition was hosted on the benchmark platform `www.causeme.net`. C4C offers an extensive number of climate model-based time series datasets with known causal ground truth that incorporate the main challenges of causal discovery in climate research. We give an overview over the benchmark platform, the challenges modeled, how datasets were generated, and implementation details. The goal of C4C is to spur more focused methodological research on causal discovery for understanding our climate system.
Keywords: Causality, climate, time series, machine learning

1. Introduction

Understanding and predicting our climate system has become one of the most critical challenges for society. Climate change is affecting weather patterns and the frequency and intensity of extreme events, therefore it is more crucial than ever to improve our knowledge of the complex interdependencies of the climate system. To do so, we often rely to modeling and estimations done with climate models and observational data coming from satellite and in situ observational data measurements of essential climate variables such as temperature. These sources of information are complementary and help in the scientific discovery process. Observational causal discovery is a major current topic in machine learning, but still in its infancy in many applied fields, such as Earth sciences (Runge et al., 2019). But perhaps more important is the fact that current causal discovery methods are not adapted to the climate data challenges, and importantly they have not exhaustively evaluated in representative climate data challenges in terms of accuracy and robustness. In this sense, benchmark datasets and competitions have been a major driver of innovation in machine learning, and we believe they should also play a role in causal discovery. Here we present the *Causality 4 Climate* (C4C) challenge as part of the NeurIPS 2019 competitions track.

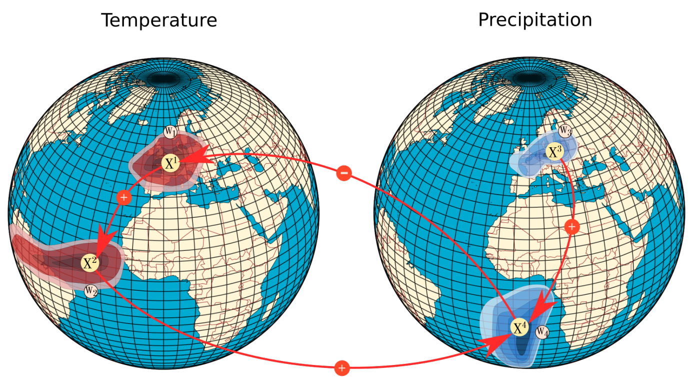

Figure 1: Teleconnections between climate modes of variability (X^i) describe regional subprocess interactions.

C4C offers an extensive number of climate model-based time series datasets with known causal ground truth that incorporate the main challenges of causal discovery in climate research. The focus is on the discovery of global climate teleconnections that causally connect far-away major climate subprocesses (modes of variability) such as El Niño-Southern Oscillation (ENSO) in the tropical Pacific and the North Atlantic Oscillation that strongly drives European and North American climates. This paper also gives an overview of the benchmark platform `www.causeme.net` where the C4C competition run. In particular, we review the modeled challenges, how datasets were generated, the accuracy and robust scores adopted, and some implementation details. Further contributions to these proceedings are on the winning methods. The datasets, platform and results will be curated and freely available, with the aim to spur more focused research and contribute to connecting the machine learning and climate science communities to better understand one of the main challenges of humanity – climate change.

2. Causal discovery and challenges in Earth system science

An overview of causal discovery is presented in a recent *Nature Communications* Perspective paper (Runge et al., 2019). A plethora of methods for causal discovery exist, all based on connecting assumptions about properties of the data with statistical inference techniques. Concepts range from Granger causality time series modeling (Granger, 1969), via nonlinear dynamics inspired methods (Sugihara et al., 2012) to structural causal models (Peters et al., 2017; Pérez-Suay and Camps-Valls, 2019) and the frameworks of conditional independence-based discovery algorithms (Spirtes et al., 2000). Each of these frameworks and their individual methods has its strengths and weaknesses and the goal of C4C is to identify promising candidates for climate data challenges.

In a typical real life climate research scenario (Kretschmer et al., 2016), a climatologist will test a causal hypothesis by investigating dependencies between several index time series describing relevant climatic subprocesses (such as that ENSO's influence on rainfall in California). The index time series are often reconstructed from gridded spatio-temporal satellite data fields of climate variables (temperature, pressure, rainfall, etc.) by either spa-

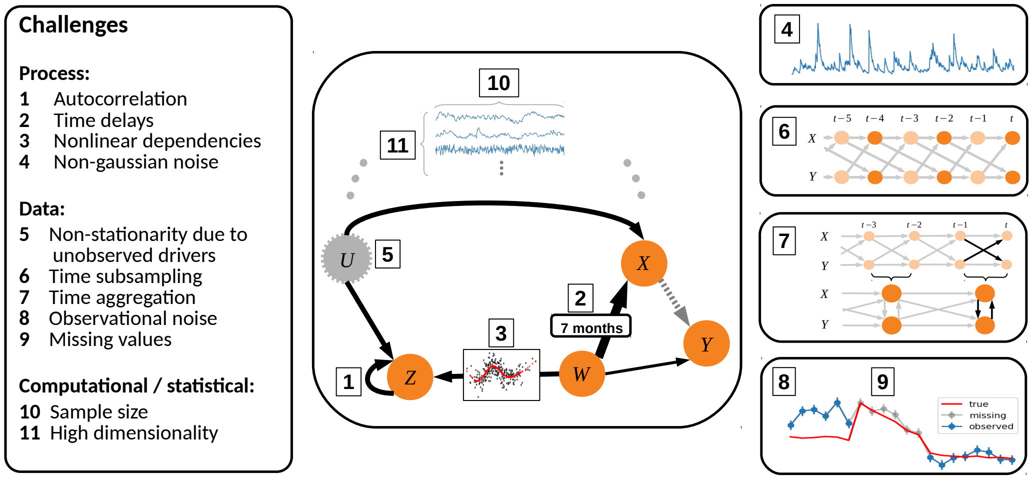

Figure 2: Methodological challenges for causal discovery featured in C4C.

tially averaging the data field over conventionally defined regions, or by using dimensionality reduction methods such as principal component analysis or rotated principal component analysis (e.g., Varimax) (Hannachi et al., 2007). Figure 1 illustrates such scenario.

The challenges of such a causal discovery analysis tackled in C4C (see numbered list in Fig. 2) are based on those presented in (Runge et al., 2019). The time-dependent nature of the physical processes gives rise to strong autocorrelation (1) in the data and time delays (2) by which far-away processes are connected can be very large. Not least since Lorenz famous chaotic weather model we know that nonlinear dynamics (3) are behind weather and climate processes which poses a challenge for statistical modeling techniques. Further, the data distributions are often highly non-Gaussian (4) such as precipitation.

Based on these ubiquitous challenges of the underlying processes themselves, we here also model typical challenges that emerge by the way the data is acquired and processed. These include that important drivers may be (partially) unobserved or undersampled, and here we model the common case of non-stationarity (5) due to such unobserved drivers, for example, slow oceanic processes modulating fast atmospheric dynamics. Further, time-subsampling (6) results from satellites measuring a particular quantity in a region only every few days, while time-aggregation (7) comes from the standard procedure to average climate variability measured at fast time resolution to a monthly time-resolution to reduce the ‘weather noise’. On the data quality side, satellites, as well as station instruments, are plagued by observational measurement noise (8) and also missing values (9) (notably cloud occlusions or sensor malfunctioning). The typical computational and statistical challenges concern sample size (10) due to the limited past availability of satellite records and high-dimensionality (11) emerges since climate researchers face the dilemma that including more variables make a causal discovery analysis more credible (since more potential common drivers are included), but at the same time the increased dimensionality leads to lower detection power and true causal links might be overlooked.

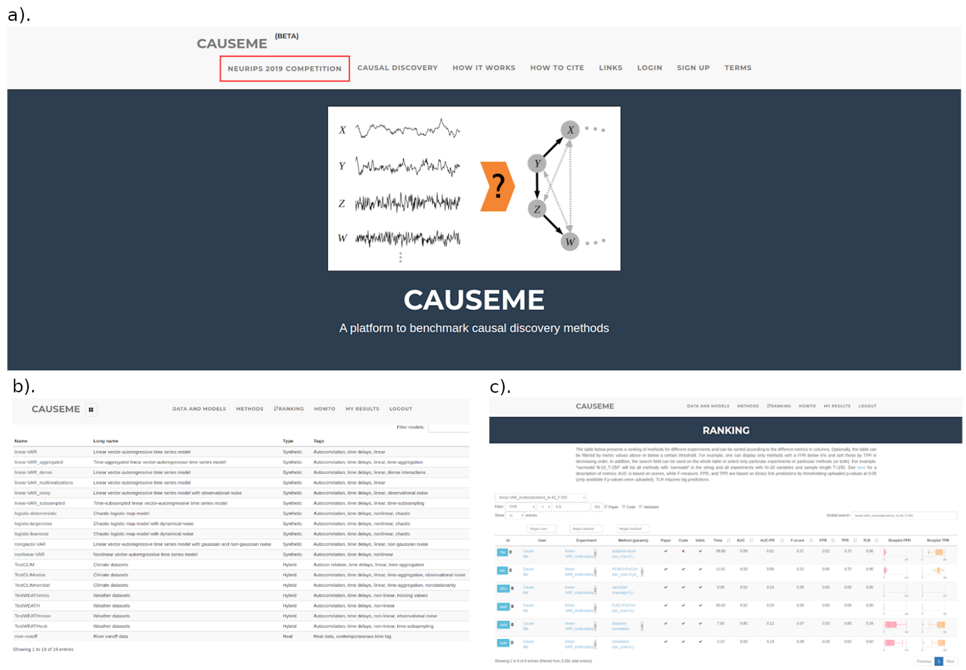

Figure 3: The C4C challenge is available at `causeme.net`. The main front page (a) contains information about the platform and describes how to use it. The data/models page (b) overviews all available datasets tagged by challenge. The ranking page (c) summarizes the results by user, methodology and evaluation metric.

3. The CauseMe platform

`Causeme.net` is a benchmark platform designed to compare the performance of causal discovery methods. It contains a growing number of multivariate time series datasets, some real and some generated from synthetic and hybrid models, but all with ground truth. These datasets model a large number of different real world challenges, those described above, and many more (Runge et al., 2019). In most cases, the datasets for particular challenges are available with different numbers of variables (dimensionality) and time-length to also cover these important challenges for causal discovery. There are two ways to contribute. Either by downloading datasets and uploading predicted causal relations, or by contributing further datasets with known causal ground truth. The workflow to upload estimated causal networks on the CauseMe platform is as follows: (1) a new method is registered and described, (2) datasets are downloaded and the method applied, and (3) results are uploaded. To contribute new datasets the maintainers of `Causeme.net` can be contacted.

`Causeme.net` is structured as follows: The section *HowTo* covers the necessary information to use the platform: explanatory videos, examples of methods implemented in Python, R, Octave and Matlab, as well as a detailed description of performance metrics. The section *My results* allows the users to register their methods and upload their results on the experiments. The section *Data and models* contains a list of the available data sets with a description and tags that list the various challenges they include. In the section *Methods*, one can find all the methods registered by users of the platform. By clicking on one method, it is possible to access the information provided by the user. Providing links to description papers and ideally code allows visitors of the platform to obtain more information and get access to high-ranking methods. The section *Ranking* features an advanced ranking system that allows ordering the method performance for different datasets by different performance metrics, and to use various filters, including hiding those methods that have no paper or code information. Likewise, a field 'validated' indicates if the maintainers of `causeme.net` have been able to reproduce the results. The goal is to encourage open access to information about the causal discovery methods.

4. Causality 4 Climate competition setup

4.1. Climate and weather data

It is difficult to obtain ground truth on causal relationships among modes of climate variability since real world causal experiments are infeasible. To obtain realistic ground truth data in a controlled fashion, we use climate model output from so-called *pre-industrial control runs* (Eyring et al., 2016). Specifically, we construct ground truth data as follows:

1. Extract time series representing relevant subprocess components from climate models.
2. Randomly draw N component time series and fit a linear VAR model with truncated coefficients defining the ground truth model among the N time series.
3. Create datasets by generating realizations with the ground truth models.
4. Process these datasets to add further data challenges.
5. Repeat (2-4) to obtain 400 realizations per experiment (as indicated in Table 1), 200 for the training phase and 200 for the final phase.

Table 1 provides an overview of the final experiments. In the following sections we describe these steps in more detail.

4.1.1. Climate model data

The competition used climate simulation data from the fifth phase of *The Coupled Model Intercomparison Project* fifth phase (CMIP5) for the Canadian CanESM2 and the French IPSL-CM5A-MR models. We used the pre-industrial control simulations (piControl) which are performed under conditions chosen to be representative of the period prior to the onset of large-scale industrialization. The main advantage of piControl runs is that they provide very long time series (200-500 model years) of stationary climate system data. Further, since interdependencies are seasonally varying, we de-seasonalize the data. We used the following climate variables: hfls, hfss, huss, rlds, rlus, rlut, ta, tas, tasmax, tasmin, uas, va, vas, wap, zg (see descriptions in (Eyring et al., 2016)). The models from which the original data came and the climate variables were hidden to the participants since that would allow them to generate similar ground truth data. We extract time series representing relevant subprocesses by applying Varimax-rotated principal component analysis to monthly averages of these spatio-temporal datasets. The obtained weights are then used to generate daily component time series. Finally, these are averaged to a 5-day time resolution.

4.1.2. Statistical models and ground truth

As explained above, the time series represent different climate subprocesses in different variables. We randomly pick N component time series and construct the ground truth by fitting a multivariate linear vector autoregressive (VAR) model (LIN) given by

$$X_t^j = \sum_{\tau=1}^{\tau_{\max}} a_j^\tau X_{t-\tau}^j + \sum_{i=1}^{N} \sum_{\tau=1}^{\tau_{\max}} c_{ji}^\tau X_{t-\tau}^i + \eta_t^j \quad (1)$$

Here a_j^τ are the variable auto-dependency coefficients and c_{ji}^τ are the cross-variable dependency coefficients. To obtain ground truth, coefficients with an absolute value smaller than the threshold $\lambda = .22$ were set to zero. The sparsity of the model was controlled by only keeping those random draws of models with a minimum number of $L = N$ links. A binary ground truth matrix of shape $N \times N$ is then constructed as

$$A_{ij} = \begin{cases} 1 & \text{if } \ |c_{ji}^\tau| > \lambda \ \text{ for any time lag } \tau > 0 \text{, indicating a causality } i \to j \\ 0 & \text{else} \,. \end{cases} \quad (2)$$

For each fitted model, its corresponding residual terms are stored. We create datasets by generating realizations of length T with the ground truth models where we add random and independent draws from the residuals at each time-step. This way, we achieve that the noise of the synthetic data follows the same distribution as the climate model data. Since the minimal time lag is $\tau = 1$ day, there is no ambiguity in the direction of causality and feedbacks between variables, i.e., $A_{ij} = A_{ji} = 1$ are still acyclic in the underlying time-resolved dynamics. That is, our model has no contemporaneous causal dependencies at a 5-daily time scale, but contemporaneous causal relations appear due to time-aggregation, see the data challenges below. These datasets then feature the challenges of autocorrelation, time delays, and non-Gaussian noise for N-dimensional datasets with sample size T. To generate nonlinear data, we also use a nonlinear generalized additive model (NONLIN) with the linear function in model (1) replaced by a nonlinear $f_{ji}^\tau(x) = x + 5x^2 e^{\frac{-x^2}{20}}$. Further, we process these datasets to add further challenges as described below.

4.1.3. Climate and weather modeling settings

Based on these two types of ground truth models (LIN and NONLIN), we construct a number of models featuring data challenges inspired by real world application scenarios (see Tab. 1). We model two types of application scenarios: (1) climate and (2) weather.

Climate variability is typically estimated from monthly data. The challenge for causal discovery comes from the fact that time-aggregation leads to many causal effects being ‘contemporaneous’. This time-aggregation is here modelled by averaging all 5-day measurements of a particular month to obtain $T \approx 100 - 250$ monthly samples. For the climate scenario, we only use the linear models (LIN). The challenge of high dimensionality is modelled by different numbers N of climate components to which the above models are fitted.

We select $N = 5, 10, 40$ to evaluate the method's performance both in a low-dimensional and high-dimensional regime. Climate data can also be non-stationary which we model by adding a term modeled as an Ornstein–Uhlenbeck process to the N components. Last, we model observational noise added to the time series after the data generation. With these challenges, we generate three experiment types as indicated in Table 1. With two different sample sizes T and two different N, we have 12 CLIM experiments, each with 200 dataset realizations.

Weather variability takes place on much shorter time scales. We chose two sample sizes of $T = 1000, 2000$ weekly samples. Also, non-linearity plays a bigger role on these fast time scales (e.g., the chaotic Lorenz system as a simple weather model). Our dataset features both linear and nonlinear dependencies. Data challenges are also slightly different in the weather scenario were we model time sub-sampling at every three weeks. As satellites observations and ground measurement stations also suffer from missing values, here we randomly remove 1% of the values. Last, also observational noise plays a key role in satellite data analysis. Modeling these challenges, we generate four types of experiments as indicated in Table 1. With two different sample sizes T and two different N, we have 16 WEATH experiments, each with 200 dataset realizations.

4.2. Further 'bonus' experiments

We also included 6 further experiments from the main platform that are further described there and listed in Table 1: Linear and nonlinear VAR models with Gaussian noise of different N and coupled chaotic logistic maps for different dynamical noise values to mimic nonlinear chaotic systems.

5. Setup and score metric

There are in total 28 categories of varying numbers of variables and sample sizes in the competition. These correspond to the real world challenges listed in Tab. 1. The task of the competition is to predict the causal connectivity matrices among the N components of each dataset, the time lag doesn't matter for evaluation (some methods may not yield a causal time lag). More precisely, participants will upload matrices C of shape $N \times N$ with non-negative real entries between 0 and 1,

$$C_{ij} = \begin{cases} 1 & \text{indicating a causal link } i \to j \text{ with high confidence} \\ 0 & \text{indicating the absence of a causal link } i \to j \text{ with high confidence} \\ \multicolumn{2}{l}{\text{a number between 0 and 1 to indicate lower confidence for the two cases}\,.} \end{cases} \tag{3}$$

5.1. Metrics

The evaluation of each solution will be based on the standard objective measure of the area-under-the-curve (AUC) of the receiver operating curve (ROC) for each challenge dataset. The AUC is calculated using the trapezoid method. The AUC is well suited for causal discovery since it balances false and true positives. The same metric was also used in the previous Connectomics challenge (Battaglia et al., 2017). As shown in Table 1 there are 28 different types of challenges (process and data challenges, as well as sample size and numbers of variables) with 200 realization datasets for each challenge. We compute one AUC from the 200 realization datasets which provides a robust evaluation of the performance on a particular challenge. This procedure results in an AUC score for each model, sample size, and number of variables, e.g., `CLIMnoise_N-40_T-100`. Participants could win in any of the 28 categories and, in addition, an overall winner was based on the average AUC score across all 28 datasets (counting non-participation in a category as a zero AUC).

6. Competition phases

The competition had two phases: (1) a *feedback/calibration phase* (where we provided a reduced number challenge datasets); and (2) the *submission phase* on the final, complete datasets. The system provides AUC averages as well as a 'Hall of Fame' ranking (leader board) of the best performing teams after submissions:

- The *calibration phase* was aimed to give participants the opportunity to familiarize with the platform, problems/challenges and datasets.
- In the submission phase only the last submission counted towards the final evaluation.

Cheating was prevented by not disclosing which climate models were chosen, by using different models for each realization, by not disclosing the dimensionality reduction method used, and by shuffling the column order of datasets (i.e., the N variables).

7. Conclusions

Establishing causal relations between random variables from observational data is one of the most important challenges in data science. The problem is of paramount relevance, especially in the current scenario of climate change. However, causal discovery methods have not systematically compared in our field. Here we aimed to set up the basis for a consistent evaluation framework, and provided well-curated climate data simulations, including some of the most important data challenges (e.g. non-stationarity, non-linearity, missing data), and a web platform for causal model evaluation and user/methods ranking. More than a hundred participants contributed with many innovative techniques and approaches. In a separate article (Weichwald et al., 2020) some of the winning methods for this competition will be discussed. It goes without saying that the platform is an open-access initiative and a live experiment that will grow in the near future with more problems and methods. Our far-end goal of learning what are the most suitable causal models for each causal data challenge in climate.

Acknowledgments

We are very grateful to Amazon for sponsoring prize money (10,000) in addition to computational resources, and personally thank Cameron Peron and Rebecca Wolff for their support and for publicising C4C. We also thank Neha Goel from Mathworks for providing free MatLab lizences. Veronika Eyring and Peer Nowack guided with their climate knowledge and Andreas Gerhardus and Christoph Kaeding helped in setting up the datasets.

Gustau Camps-Valls was supported by the European Research Council (ERC) through the ERC Consolidator Grant SEDAL (project id 647423).

References

Demian. Battaglia, Isabelle Guyon, Vincent Lemaire, and Javier Orlandi. *Neural Connectomics Challenge*. Springer, 2017.

Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, 2016. ISSN 19919603.

C W J Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

A. Hannachi, I.T. Jolliffe, and D.B. Stephenson. Empirical orthogonal functions and related techniques in atmospheric science: A review. *International Journal of Climatology*, 27: 1119–1152, 2007. doi: 10.1002/joc.1499.

Marlene Kretschmer, Dim Coumou, Jonathan F. Donges, and Jakob Runge. Using causal effect networks to analyze different arctic drivers of midlatitude winter circulation. *Journal of Climate*, 29(11):4069–4081, mar 2016. ISSN 08948755.

A. Pérez-Suay and G. Camps-Valls. Causal inference in geoscience and remote sensing from observational data. *IEEE Transactions on Geoscience and Remote Sensing*, 57(3): 1502–1513, 2019.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference: foundations and learning algorithms*. MIT Press, Cambridge, MA, 2017. ISBN 9780262037310.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, Egbert H van Nes, Jonas Peters, Rick Quax, Markus Reichstein, Marten Scheffer, Bernhard Schölkopf, Peter Spirtes, George Sugihara, Jie Sun, Kun Zhang, and Jakob Zscheischler. Inferring causation from time series in earth system sciences. *Nature Communications*, 10(1):2553, 2019.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Boston, 2000.

George Sugihara, Robert May, Hao Ye, Chih-hao Hsieh, Ethan Deyle, Michael Fogarty, and Stephan Munch. Detecting causality in complex ecosystems. *Science (80-.).*, 338(6106): 496–500, oct 2012. ISSN 1095-9203. doi: 10.1126/science.1227079.

S. Weichwald, M. E. Jakobsen, P. B. Mogensen, L. Petersen, N. Thams, and G. Varando. Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values. In *submitted to the NeurIPS2019 Competition & Demonstration Track Postproceedings*, 2020.

Appendix A. Table of setup of datasets

Model	**Process challenges** (additionally)	**Data challenges**	**Sample size and number of variables**
CLIM		Time aggregation	$T = 100, 250$ $N = 5, 40$
CLIMnoise		Time aggregation, observational noise	$T = 100, 250$ $N = 5, 40$
CLIMnonstat	Nonstationarity	Time aggregation	$T = 100, 250$ $N = 5, 40$
WEATH	Nonlinearity	-	$T = 1000, 2000$ $N = 5, 10$
WEATHsub	Nonlinearity	Time-subsampling	$T = 1000, 2000$ $N = 5, 10$
WEATHnoise	Nonlinearity	Observational noise	$T = 1000, 2000$ $N = 5, 10$
WEATHmiss	Nonlinearity	Missing values (1%)	$T = 1000, 2000$ $N = 5, 10$
Linear-VAR			$T = 150$ $N = 10, 100$
Nonlinear-VAR	Nonlinearity		$T = 600$ $N = 20$
Logistic	Chaos	3 noise levels	$T = 150$ $N = 5$

Table 1: Setup of datasets in seven model categories featuring different process and data challenges (in addition to basis ones: autocorrelation, time delays, non-Gaussian noise) for the climate and weather scenario. Each model is simulated for different sample sizes T and numbers of variables N. For each such setup we simulate 200 ensemble realizations to be able to robustly estimate performance.