Universität Bremen

Machine-Learning Based Observational Cloud Products for Process-Oriented Climate Model Evaluation

Doctoral Dissertation of

Arndt Kaps

December 2023

University of Bremen

Institute of Environmental Physics (IUP)

Machine-Learning Based Observational Cloud Products for Process-Oriented Climate Model Evaluation

Doctoral Dissertation of

Arndt Kaps

A thesis submitted in fulfillment of the requirements for the degree

Doktor der Naturwissenschaften (Dr. rer. nat.)

Primary Examiner: Prof. Dr. Veronika Eyring
Secondary Examiner: Prof. Dr. Hartmut Bösch

Submission: 11 December 2023

Abstract

The importance of clouds in regulating the Earth's energy balance as well as moisture and heat distributions cannot be overstated. Consequently, clouds have a considerable influence on the trajectory of anthropogenic climate change, of which possible scenarios are being studied with global climate models (GCMs). Uncertainties from the representation of clouds in GCMs have been identified as a leading cause of inter-model spread in climate projections. Our current understanding of clouds and the processes relevant to their formation and effect on climate is informed partly by observations from remote sensing instruments aboard orbital satellites. This thesis introduces new methods of characterizing clouds from space with the help of machine learning and neural networks. The purpose of these methods is to improve the understanding of and reduce the uncertainties in climate projections by providing satellite products that are objectively interpretable and consistently comparable to GCM output.

In a first study, the lack of interpretability in existing products is addressed with a newly developed framework to assign cloud classes to satellite data and GCM output. A neural network and a Random Forest are combined and trained on observations from both active and passive satellite sensors to infer cloud class distributions from low-resolution cloud property data. During training, the models use cloud properties from the Moderate Resolution Imaging Spectroradiometer (MODIS) as inputs. The ground truth classes - eight cloud types defined to be similar to those established by the World Meteorological Organization (WMO) - are obtained from CloudSat radar and Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) lidar measurements. The generalization performance of the framework is assessed using the European Space Agency (ESA) Climate Change Initiative cloud dataset (Cloud_cci). Throughout all stages of machine learning, the predicted cloud-type distributions are physically consistent with the WMO definitions and comparable to those of the ground truth dataset. This allows cloud-related data to be presented in the phase space of cloud classes, which makes the data more easily interpretable and usable for GCM evaluation.

Based on this, the trained framework was used to create a new Cloud Class Climatology (CCClim) from the complete ESA Cloud_cci AVHRR-PMv3 (ESA-CCI) dataset. CCClim contains daily mean values of the cloud properties from ESA-CCI and the predicted cloud-type distributions globally at 1° resolution over 35 years. Compared to existing cloud-type datasets, CCClim provides comparable or better resolution and coverage and as it is based on active sensor data, allows for more objective downstream studies. Applying the machine-learning framework to the output of GCM and comparing the simulated to observed cloud-type

distributions is demonstrated as one of the use cases of CCClim using output from a simulation of the Icosahedral Nonhydrostatic Atmosphere model (ICON-A) climate model. CCClim acts as a new basis for process-based analysis of clouds and can be valuable for evaluating similar cloud class distributions in GCMs.
The limited comparability between GCMs and observations is addressed in a third study by employing neural-network-based generative domain adaptation, tailored specifically to satellite observations of clouds. This process aims to produce synthetic observations - similar to those of instrument simulators - from existing GCM output. For this purpose, a cycle-consistent generative adversarial network (CycleGAN) is trained to convert ICON-A scenes to ESA-CCI observations and vice-versa and compared to output from an established satellite simulator package.
The methods explored in this thesis highlight that machine learning and especially neural networks have the potential to improve multiple aspects of climate science. The presented results show that cloud classes can be reliably obtained from low-resolution data to improve their interpretability. They further show that comparison between climate models and observations can potentially be simplified with machine learning.

Integrated Author's References

The following publications and their supplements, both submitted to peer-reviewed journals, constitute parts of this thesis. The chapters where this applies are given in Section **1.3** with further details stated in the chapters themselves.

Kaps, **A.**, A. Lauer, G. Camps-Valls, P. Gentine, L. Gomez-Chova, and V. Eyring (2023a). "Machine-Learned Cloud Classes From Satellite Data for Process-Oriented Climate Model Evaluation". In: *IEEE Transactions on Geoscience and Remote Sensing* 61, pp. 1–15. DOI: 10.1109/TGRS.2023.3237008.

Kaps, **A.**, A. Lauer, R. Kazeroni, M. Stengel, and V. Eyring (2023c). "Characterizing clouds with the CCClim dataset, a machine learning cloud class climatology". In: *Earth System Science Data Discussions*. in review. DOI: 10.5194/essd-2023-424.

Contents

Notation

x	scalar
$\mathbf{x}$	vector
x_i	scalar element of vector $\mathbf{x}$ at index i
$\mathbf{x}^{(i)}$	vector $\mathbf{x}$ at index i in a set of vectors
$\mathbf{X}$	2D matrix or higher-dimensional tensor
X_{ij}	scalar element of matrix $\mathbf{X}$ at index (i, j)
$f(x)/\mathbf{f}(x)$	scalar- / vector-valued function of x
$\mathbf{F}(x)$	Matrix or tensor-valued function of x
$\mathbb{X}$	set of any-dimensional objects
$\mathbb{E}_{x\sim P}[x]$	expected value of random variable x sampled from distribution P
$*$	element-wise multiplication operator
$\circ$	convolution operator
$\bar{\mathbf{x}}$	arithmetic mean of $\mathbf{x}$: $\frac{1}{N}\sum_i^N x_i$
$\lvert\mathbf{x}\rvert_p$	L_p-norm of $\mathbf{x}$: $\sqrt[p]{\sum_i^N \lvert x_i\rvert^p}$
$\log(\mathbf{x})/\exp(\mathbf{x})$	element-wise natural logarithm/ exponential function

1. Introduction

1.1. Motivation

Observations from various sources have shown evidence of climate change for decades now, making the connection between human activities and climate change unequivocal (IPCC, 2021). Using observations to link extreme weather events, which have become increasingly frequent and devastating in recent years, to climate change has further moved the climate crisis into the public eye (Otto, 2023). Due to the unprecedented impact climate change is expected to have on humankind, its causal mechanisms as well as its consequences will remain important subjects of further studies. Informed decisions are crucial for developing effective mitigation and adaptation strategies, and require projections of possible climate scenarios. The future climate scenarios are mainly projected by global climate models (GCMs), which are numerical representations of the climate system that simulate the extension of past and current climate into the future using equations approximating known physical relationships (Gettelman and Rood, 2016, ch. 4). Many different GCMs are developed with varying strategies and goals and are used to produce an ensemble of possible future scenarios (Eyring et al., 2016).
The development of GCMs relies on information from theory, small-scale experiments, high-resolution simulations and observations of the Earth system (Plant and Yano, 2014, ch. 16). Where approximations of physical processes are required to limit computational costs of GCMs, specifically when the numerical grid of the simulation is not fine enough to resolve a process, *parametrizations* approximates the effects of subgrid-scale processes on the explicitly resolved variables. Parametrizations are often adjusted to match expected behavior in a *tuning* process which is constrained by observations (Hourdin et al., 2017). Also, observations are used to assess the performance of GCMs to simulate current and historical climate, to *evaluate* if a GCM is suited for projecting future climate (e.g. Eyring et al., 2019; Lauer et al., 2023). Evaluation and subsequent development of GCMs have enabled progress in the representation of multiple processes in current GCM, e.g. for supercooled liquid cloud droplets or the response of Arctic sea ice to carbon dioxide emissions (Arias et al., 2021). Nonetheless, the uncertainty associated with the projections of future climate has not decreased (Meehl et al., 2020; Schlund et al., 2020b). While incorporating other sources of information, e.g. from paleoclimate, has reduced the cloud-related uncertainty, clouds remain major contributors to the overall uncertainty in climate projections (Arias et al., 2021). Therefore, evaluating how processes

governing cloud formation and evolution are represented in current GCMs is essential to reduce the uncertainty in future projections of climate change. Consequently, it is important to ensure that the observations used for tuning and evaluation are of sufficient quality and are fit for purpose. For this, observational datasets aim to meet the following basic requirements:

- Accuracy: the measured quantities provide an accurate picture of the real state
- Comparability: the measured quantities are comparable to those simulated by GCMs
- Coverage: the observations sample a large area in space, time and phase-space
- Interpretability: the measured quantities can objectively be attributed to physical processes

Accuracy and comparability are subject to the methods (*retrieval algorithms*) used to compute the physical quantities from raw measurements, which are often fundamentally different from how they are computed in GCMs. Sufficient coverage is important because global model performance can not be fully evaluated from a small number of local samples. Interpretable observations allow for straightforward inference of related physical processes with minimal subjective bias. Naturally, observations can not flawlessly fulfill all of these requirements simultaneously but are typically designed specifically to perform well on a subset. This thesis focuses on the improvement of interpretability and comparability as these can be optimized using existing products and are less dependent on the characteristics of the instrument than coverage and accuracy.

Observations of clouds are obtained from in situ measurements from balloons or airplanes, from remote-sensing instruments on the surface or similar instruments aboard orbital satellites. While in situ measurements can be helpful for process analysis (Stevens et al., 2021) they provide little information about clouds on a global scale. Surface-based measurements are very localized and while this makes them very useful to obtain temporally continuous measurements, they are spatially sparse, especially over oceans, and thus cannot fully capture large-scale climate-relevant cloud properties. Measurements from remote-sensing instruments aboard spaceborne satellites can measure cloud properties almost globally. However, satellite observations and GCM-simulated quantities are produced fundamentally differently and their comparability is therefore limited. A strategy to mitigate this problem is to use software that simulates the observed view and characteristics of a satellite instrument for a GCM scene, so-called *satellite simulators*. Additional limitations of satellite observations remain and will be a subject of this thesis, but a variety of observational cloud products from satellite instruments have helped advance the understanding of clouds. The International Satellite Cloud Climatology Project (ISCCP) (Schiffer and Rossow, 1983) is a long-standing provider of cloud observations, including a categorization into *morphological* cloud types (Rossow and Schiffer,

1999), i.e. types defined by their visual appearance (WMO, 2023). Addressing cloud-related processes through the lens of cloud types can be useful because different types vary in their effects, for example on the Earth's radiation budget. Low-level clouds tend to have a cooling effect because they reflect solar radiation while having a small impact on the net thermal emission to space, but high clouds act as warming agents because they trap more thermal radiation in the atmosphere. It has been proposed that an observation-based decomposition of clouds by type or regime can enhance cloud analysis and evaluation of the cloud representation in GCMs (Stephens, 2005). Improving the computational methods by which cloud observations are categorized and analyzed could therefore increase the efficiency and effectiveness of GCM evaluation and ultimately lead to improved GCMs providing enhanced reliability of future climate projections.

In climate modeling, processing of observations and GCM evaluation, applying machine learning (ML) has become part of the state-of-the-art (Rasp et al. (2018), Groenke et al. (2020), Schlund et al. (2020a), respectively). This development is driven by the recent increase in available computing power and the performance of new algorithms. This is especially the case for new graphics processing unit (GPU) generations designed specifically for deep learning (DL) applications. While ML has a long history in climate science (Malone, 1955), the new developments have enabled much more expansive applications (Camps-Valls et al., 2021). ML is ideally suited to process large amounts of data from GCMs and observations, distilling them down to their relevant content and making them easier to work with.

To address uncertainties in the representation of clouds in GCMs, this thesis introduces new methods that aim to make the evaluation of cloud processes in GCMs more effective through the use of ML and satellite observations. By training the ML methods on a combination of multiple satellite products, improvements of both *coverage* and *accuracy* over the individual products are achieved. The resulting ML framework is designed to be applicable to GCM output, producing a more *interpretable* picture of clouds and their related processes via the assignment of cloud classes. To achieve the same for observations, this method is applied to satellite data creating an extended observational dataset. This Cloud Class Climatology (CCClim) dataset provides long-term information on cloud classes with global coverage. Leveraging ML and active sensor data results in a cloud classification method that is arguably more objective and consistent than currently existing frameworks. In addition to addressing the interpretability of cloud-related data, improving the *comparability* between existing GCM output and observations is attempted with a DL domain adaptation method as an "offline" alternative to satellite simulators.
Together, the classification framework, the CCClim dataset and the domain adaptation method represent new ML-based and data-driven options for the analysis of cloud-related processes in observational data and GCMs. Expanding on the typically retrieved quantities and leveraging the known relationships between cloud types and the processes involved in their formation

and evolution, the options can help to identify and understand error sources in GCMs and eventually reduce the uncertainty of climate projections.

1.2. Central Scientific Questions

This thesis aims to answer three overarching scientific questions:

- **Question 1** : "Can physically robust and self-consistent cloud-type distributions be obtained from data at resolutions typical for global climate models?"
- **Question 2** : "Does the explicit addition of cloud-type labels benefit the analysis and understanding of cloud-related processes to improve climate model evaluation?"
- **Question 3** : "Can the systematic bias (domain shift) between satellite data and climate model output be quantified or possibly reduced using generative domain adaptation?"

1.3. Content and Structure of this Thesis

Two of the chapters in this thesis are in large parts already published in or under review at peer-reviewed journals in the form of two first-author papers. To account for the involvement of all co-authors of these publications, the pronoun "we" is used in sections from these papers instead of the passive voice. The scientific content as well as all text, figures and tables shown are the results of the work performed by the author of this thesis, unless explicitly stated otherwise, with specific contributions by others to the content of the included publications being clearly stated as well. The scientific background of the methods and related research is provided in Chapter **2** and the observational and GCM data used here are described in Chapter **3**, which contains tables from Kaps et al. (2023a). In Chapter **4** the ML framework for cloud classification is presented, based on a paper published in *Transaction in Geoscience and Remote Sensing* (Kaps et al., 2023a). The method is used to create CCClim, which has been made available to the public to complement existing cloud-class products (Kaps et al., 2023b). CCClim and examples of potential applications are presented in Chapter **5**, based on a manuscript that at the time of submission of this dissertation is under review at *Earth System Science Data* (Kaps et al., 2023c). Chapter **6** introduces a method for generative domain adaptation (DA), with the goal of making observations and GCM data more comparable. The results are discussed in the context of the scientific questions in Chapter **7**, which also provides an outlook.

2. Scientific Background

This chapter provides the physical basis of the data, methods and terms employed in this thesis. First, Section **2.1** provides an overview of the physics involved in the formation of specific types of clouds which are at the center of this thesis. The interactions of clouds with radiation introduced in Section **2.1.1** highlight aspects of the relevance of cloudiness for the climate system while also providing context for Section **2.2**, which explains the methods involved in obtaining clouds properties from space. The context of these satellite retrieval methods and the data they provide is important to correctly interpret the results obtained with the ML methods applied downstream. These methods were the primary application by which data was produced, processed and analyzed for this thesis and Section **2.3** contains the related theoretical background. Lastly Section **2.4** provides a synopsis of the state-of-the-art in research on cloud classification for GCM evaluation, focusing on observations from space and ML methods.

2.1. Physics of Clouds

Clouds are subject to complex interconnected processes governing their formation and dynamical behavior. Different atmospheric conditions therefore lead to both qualitatively and quantitatively different cloudiness, which in turn can have varying climate feedbacks (Sherwood et al., 2020; Zelinka et al., 2020). The analysis of morphological cloud types is therefore inextricably linked to that of atmospheric processes. Section **2.1.1** will introduce important processes and Section **2.1.2** will relate them to specific cloud types.

2.1.1. Cloud-related Processes

Cloud Formation

While the definition of a cloud is not trivial (Spänkuch et al., 2022), it is safe to say that a cloud forms when atmospheric water vapor condenses, with liquid or ice particles staying suspended

in the air. While infinitesimal amounts of atmospheric water always condense and evaporate, persistent condensation occurs when the air is supersaturated with water, i.e. the water vapor pressure ν_v becomes larger than the saturation vapor pressure $\nu_s(T)$. Conceptually, cloud formation is therefore most easily thought of as cooling of subsaturated air, which reduces ν_s. Specifically, the dependence of ν_s on the temperature T is given by the Clausius-Clapeyron equation (Eq. 2.1), via the specific latent heat l_v and gas constant R_v of water vapor.

$$\frac{d\nu_s}{dT} = \frac{l_v \nu_s(T)}{R_v T^2}. \tag{2.1}$$

Essentially, Eq. 2.1 states that ν_s always increases with T and thus warmer air can contain more water vapor before becoming saturated. This relationship is highly relevant in global warming conditions as air can transport more moisture and with it latent heat, potentially increasing the severity of extreme weather events (Coumou and Rahmstorf, 2012).
A cloud can appear once the temperature cools to the dew point $T = T_d$, which most commonly happens through expansive cooling of ascending air, or to a lesser extent via radiative cooling or advection. At this point, cloud formation is governed by an interplay of the work W_A required to create a droplet and the latent heat L that is available from condensation. The change in energy induced by creation of a droplet is thus $\Delta E = W_A - L$. While $W_A = 4\pi r^2 \sigma$ is proportional to the surface area of the droplet, L increases with its volume. In supersaturated conditions, there is therefore a critical value r_K above which the energy released from condensation becomes larger than the energy needed to increase the surface of the droplet. The radius r_K, at which *homogeneous nucleation* of cloud particles is possible, is given by Kelvin's equation Eq. 2.2.

$$r_K = \frac{2\sigma}{n k_B T \log \frac{\nu_v}{\nu_s}}. \tag{2.2}$$

Homogeneous nucleation therefore relies on chance collision of subcritical droplets to reach r_K before they evaporate. As a consequence, clouds are rarely formed from homogeneous nucleation, but more often aerosols act as condensation nuclei, essentially eliminating the need to create subcritical droplets and thus reducing the amount of "activation energy" (*heterogeneous nucleation*).
After formation, the trajectory of a cloud's evolution can vary strongly with coinciding atmospheric conditions. The following Sections **2.1.1** and **2.1.2** will explain the relevant physics and the effects concerning different cloud types.

Clouds and Radiation

Clouds absorb, scatter and emit electromagnetic radiation in a broad spectrum. The strength of each of these processes strongly depends on the light's wavelength λ, as well as the micro-

physical composition and temperature of the cloud. These dependencies can be exploited to design specialized sensors that detect all cloud types. Moreover, measurements of the impact clouds have on the radiative energy distribution at the surface and in the atmosphere inform GCMs development (e.g. Loeb et al., 2009; Smith et al., 2015).
Throughout this section radiation will be assumed to be isotropic and emitted and received at normal angles. This eliminates any angular dependencies in the equations unless otherwise stated.

Two different spectral ranges are prevalent in the atmosphere: the reflection of shortwave (SW) radiation emitted by the sun, and longwave (LW) radiation from thermal emission inside the atmosphere. While the spectra from both sources are wide, most of the energy of solar radiation is confined to the range $\lambda_{SW} \in [0.4, 0.8]$ μm (visible), while for the terrestrial thermal radiation this range is $\lambda_{LW} \in [8, 12]$ μm (infrared) (Siebesma et al., 2020, ch. 4).
While the sun's emission is near-constant, the Earth's rotation causes a strong diurnal cycle in the incident SW radiation and the LW emissions from especially land surfaces. This affects cloud formation as discussed in Section **2.1.2**. The thermal emission of the surface and of clouds can be approximated using the Planck law for the spectral radiance of a black body at temperature T:

$$I_B(\lambda, T) = \frac{2hc^2}{\lambda^5 \exp(hc/\lambda k_B T) - 1}, \tag{2.3}$$

with Boltzmann constant k_B, Planck constant h and speed of light c. Equation 2.3 states that, for a given wavelength λ, the power of the radiation emitted from a unit area per unit solid angle per unit wavelength increases with T. Integrating Eq. 2.3 over the angular component and all wavelengths results in the Stefan-Boltzmann law of power P emitted per area A:

$$\frac{P}{A} = \frac{2\pi^5 k_B^4}{15c^2h^3} T^4 = \sigma T^4. \tag{2.4}$$

It follows therefore that due to their lower temperature, cloud tops emit less energy to space than the Earth's surface would in cloud-free conditions. However, Planck's law only holds for an idealized black body, while true materials are characterized as gray bodies, where Eq. 2.3 and Eq. 2.4 are only approximations. For gray bodies, the spectral emissivity ε_λ is defined as the ratio of the emitted radiance to the radiance of a black body at the same temperature (Eq. 2.5).

$$\varepsilon_\lambda = \frac{I(\lambda, T)}{I_B(\lambda, T)}. \tag{2.5}$$

The actual emission of a cloud top therefore amounts to $\varepsilon\sigma T^4$ with $\varepsilon \in (0, 1)$. From conservation of energy, it is evident that in thermal equilibrium a substance must emit as much radiative energy as it absorbs such that

$$\varepsilon_\lambda = \alpha_\lambda. \tag{2.6}$$

Equation 2.6 is known as Kirchhoff's law and holds for any wavelength λ with ε and the absorptivity α being a fundamental property of each material. Clouds act as strong greenhouse

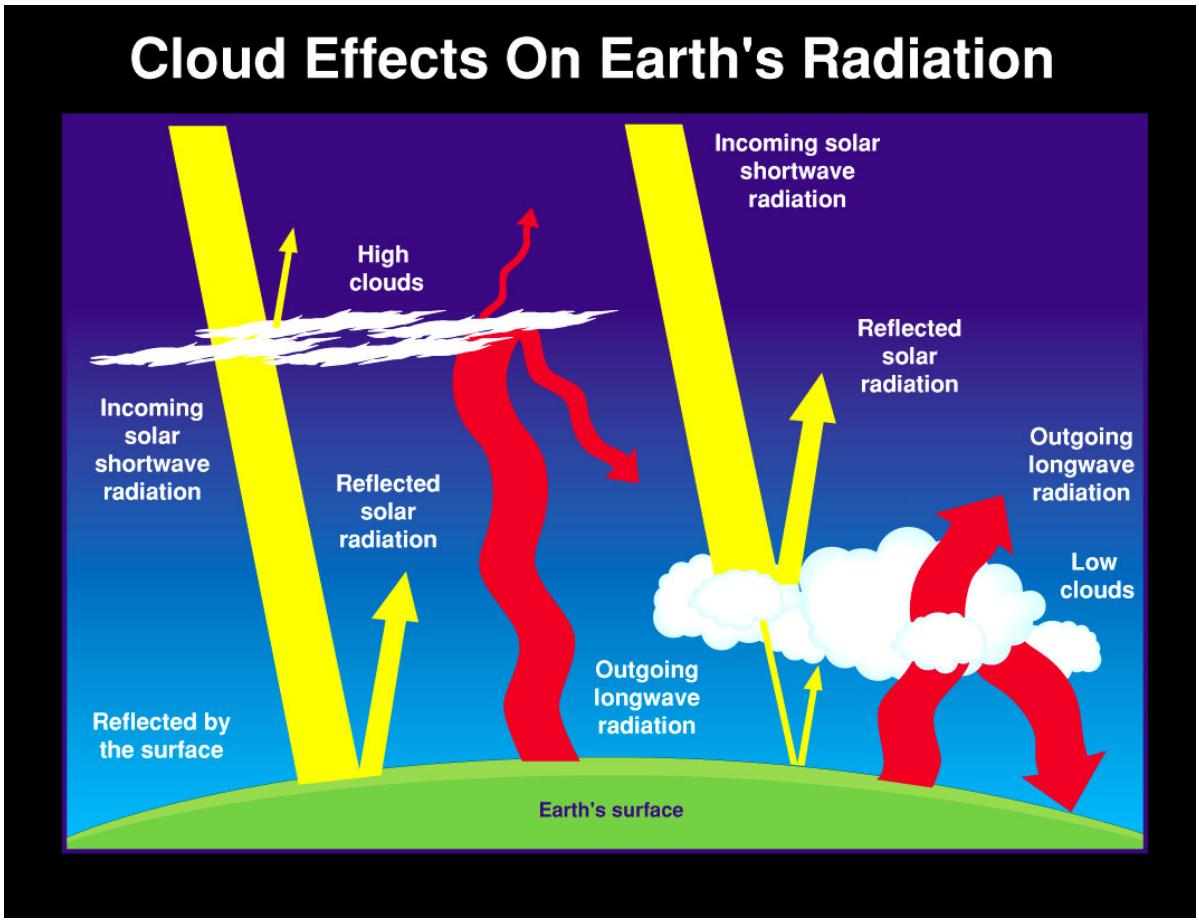


Figure 2.1.: Illustration of how reflection and absorption of incoming solar radiation and subsequent reemission of thermal radiation form the Earth's radiative budget. the height of a cloud affects its impact on the Earth's energy budget, with high clouds emitting less thermal radiation to space than low clouds due to their lower cloud-top temperature. Additionally, high clouds like cirrus are often optically thin and allow for a higher amount of solar radiation to be transmitted to the surface. Obtained from National Aeronautics and Space Administration (NASA)'s Visible Earth webpage[1].

agents and contribute to a warming of the atmosphere because cloud bases absorb and re-emit LW radiation coming from below. Also, clouds efficiently reflect incoming SW radiation directly

[1] `https://visibleearth.nasa.gov/images/54219/cloud-effects-on-earths-radiation`, last accessed on 7th of December 2023.

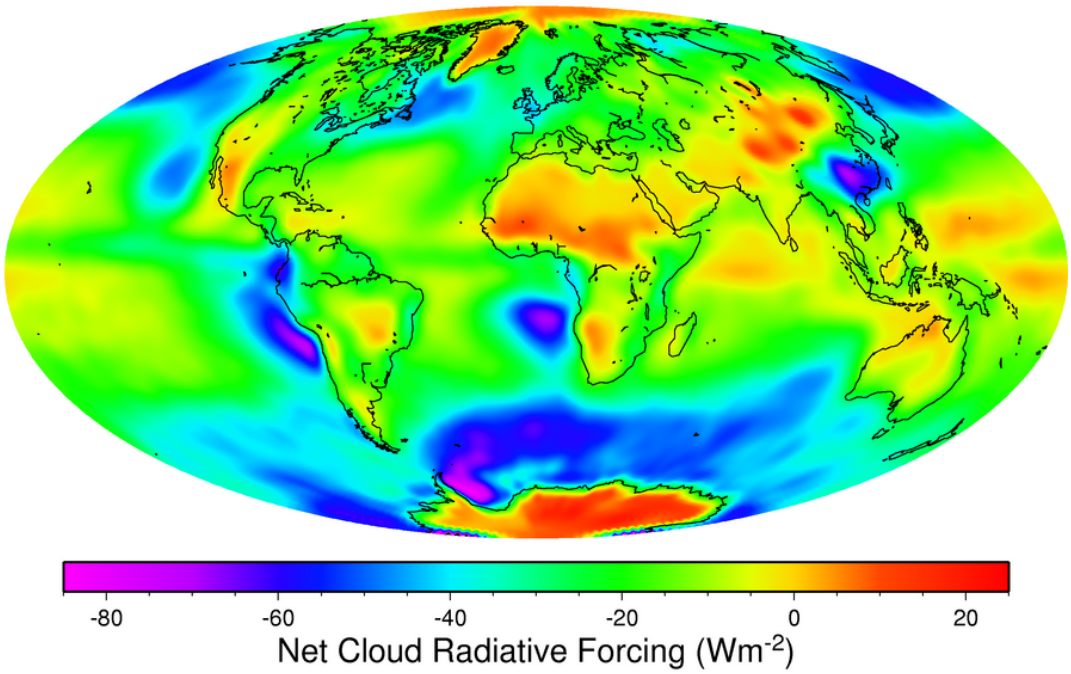


Figure 2.2.: Sum of LW and SW (net) CRE as measured by Clouds and the Earth's Radiant Energy System (CERES) averaged over observations between March 2000 and February 2001. The largest net CRE is found for clouds over the subtropical oceans, where persistent decks of low stratocumuli are known to occur. Image credit: Image courtesy of the CERES Science Team at NASA Langley Research Center in Hampton, Virginia, USA.

back to space which has a cooling effect, which is illustrated in Fig. **2.1**. The cloud radiative effect (CRE) clouds have on the radiative energy budget in the atmosphere is given by:

$$\mathrm{CRE} = F_{clear} - F_{all}. \tag{2.7}$$

Here, F_{all} and F_{clear} are the net top of the atmosphere (TOA) irradiance under average observed (*all-sky*) conditions and under the assumption of a cloud-free sky, respectively. Taking into account SW and LW contributions, the observed net CRE < 0, i.e. clouds act to cool the Earth. The global distribution of average observed CRE is shown in Fig. **2.2**, showing that especially over the oceans, clouds have a cooling effect. Projecting the future development of this value is an important goal of climate science (Forster et al., 2021).

Since clouds are comprised of liquid and/or frozen water particles, the particle properties, such as their number, size distribution, habit and temperature determine the CRE. For spherical particles, the relationship of the size of a particle to the incident wavelength provides a good indication of the scattering behavior via the *size parameter* $x = \frac{2\pi r}{\lambda}$. The distributions of particle size and shape are hard to determine, such that in practice the radius r is often replaced by the effective radius of cloud particles (*cer*), which reproduces the observed volume

to surface area fraction. Assuming spherical water droplets, *cer* is equal to (Hansen and Travis, 1974):

$$cerl = \frac{\int_0^\infty r^3 N(r) dr}{\int_0^\infty r^2 N(r) dr}. \tag{2.8}$$

While the assumption of spherical shape is not too strong for liquid particles, ice particles are distinctly non-spherical. In this case, the general form of *cer* applies:

$$ceri = \frac{\int_0^\infty V N(l) dl}{\int_0^\infty A N(l) dl}, \tag{2.9}$$

for the volume V and surface area A, where l is the maximum size of the particle in any dimension (Liou, 2002).
Depending on the size parameter x, three scattering regimes are distinguished:

- Rayleigh scattering for $x << 1$
- Mie physics for $x \approx 1$
- Geometric optics for $x >> 1$

Water droplets that remain suspended in the air typically have a size in the single- to double-digit micrometer range, while the effective radii of liquid particles (*cerl*) are of $\mathcal{O}(100\ \mu\text{m})$. Clouds therefore interact with solar radiation chiefly in the Mie regime, in which most of the light is scattered in its incident direction such that clouds are rarely fully opaque. The strength of scattering and absorption are determined respectively by the real and imaginary part of the complex index of refraction $b(\lambda) = b_r(\lambda) + ib_c(\lambda)$ of, in this case, water. In the SW range, $b_c \approx 0$, such that scattering dominates, while in the LW spectrum $b_c \approx 10^7 b_r$ is observed (Siebesma et al., 2020, ch. 4). Clouds therefore absorb almost all of the incoming infrared radiation emitted by the surface, and re-emit it according to Eq. 2.6, which causes a positive LW CRE especially for high-top clouds.

The physics by which clouds interact with radiation through absorption or scattering are not only important for understanding the CRE, but also to design instruments and algorithms that can use measured radiation to detect clouds and their composition. The amount by which a substance of density ρ absorbs or scatters radiation of a given wavelength λ is given by the *mass attenuation coefficient* $k_\lambda = a\,(b_c(\lambda)) + s\,(b_r(\lambda))$, decomposed into *mass absorption/scattering coefficient* a_λ / s_λ. The radiation transferred through the atmosphere with spectral radiance I_λ is attenuated by an atmospheric constituent with density ρ and mass attenuation coefficient k_λ

according to Eq. 2.10 (Beer-Lambert law). Equation 2.10 denotes the incremental attenuation dI_λ along a path of length dz.

$$dI_\lambda = -I_\lambda k_\lambda \rho dz, \tag{2.10}$$

$$\ln I_{\lambda,z_0} - \ln I_{\lambda,z_\infty} = \int_{z_0}^{z_\infty} k_\lambda \rho dz, \tag{2.11}$$

$$I_{\lambda,z_0} = I_{\lambda,z_\infty} \exp\left(\int_{z_0}^{z_\infty} k_\lambda \rho dz\right). \tag{2.12}$$

The exponent in Eq. 2.12 is called the optical thickness, or often cloud optical depth (*cod*) when dealing with clouds, defined as

$$\tau_\lambda \equiv cod = \int_{z_0}^{z_\infty} k_\lambda \rho dz. \tag{2.13}$$

Using Eq. 2.13 the transmissivity is defined as

$$\gamma_\lambda = e^{-\tau_\lambda}. \tag{2.14}$$

By definition, all of the incident radiation is either transmitted(γ), absorbed(α), or scattered (ϕ).

$$\gamma_\lambda + \alpha_\lambda + \phi_\lambda = 1. \tag{2.15}$$

Equation 2.15 means that, for LW radiation, which is almost completely either transmitted or absorbed by clouds, the optical thickness and the absorptivity α_λ are directly related. Since α is in turn almost always equal to ε, measurements of the emissivity in the LW spectrum inform on the optical thickness of a cloud, which is a useful relationship for remote sensing of clouds.

Cloud Dynamics and Thermodynamics

Through condensation and consecutive evaporation of water, clouds transport large amounts of heat and moisture. The mechanisms involved include movements of air masses induced by large-scale circulations, mesoscale pressure differences, local thermals (convection) over warm surfaces and orographic lifting caused by advection over mountains. Due to the range of relevant scales - horizontal, vertical and temporal - convective dynamics are among the most difficult to simulate with GCMs, thus having a large influence on the uncertainty of projected cloudiness and cloud feedback. The dynamics and thermodynamics of atmospheric convection are often described in terms of an *air parcel* of infinitesimal volume, being lifted initially by external forces such as large-scale flows. During ascent, the parcel cools dry-adiabatically as

its pressure decreases with the surrounding air (Eq. 2.16) until the temperature reaches T_d and condensation sets in at the so-called *lifting condensation level* (LCL).

$$\Gamma_d \equiv -\frac{\mathrm{d}T}{\mathrm{d}z} = \frac{g}{c_{p_d}}. \tag{2.16}$$

Γ_d is called the dry-adiabatic lapse rate and describes the adiabatic decrease in temperature T of the parcel with height z, depending on the gravitational acceleration g and the isobaric specific heat of dry air c_{p_d}. Usually, the atmospheric lapse rate is smaller than Γ_d, meaning that it is stable to infinitesimal perturbations and external forces are required to lift a parcel to the LCL. Generally, the height of the LCL decreases with the moisture available at the surface but can increase over cold water surfaces due to the smaller sensible heat flux (Haiden, 1997).

Above the LCL, since condensation of water releases latent heat, the rising parcel now cools more slowly, the temperature now following a moist-adiabatic lapse rate. At this smaller cooling rate, the parcel will eventually become warmer than the surrounding atmosphere and thus independently buoyant at the level of free convection (LFC). The buoyancy force B is the difference between the upward displacement force and the downward gravitational force, both depending on $g = 9.81\,\mathrm{m\,s^{-2}}$. For an air parcel of unit volume and density ρ

$$B = g(\bar{\rho} - \rho), \tag{2.17}$$

such that the buoyancy is positive (upward) if the density $\bar{\rho}$ of the surrounding air is larger than that of the parcel. Using the gas constant R_d for dry, the gas equation can be written as

$$p = R_d \rho T_v, \tag{2.18}$$

with the virtual temperature T_v, which eliminates a dependence on the moisture content in Eq. 2.18. With Eq. 2.18 the densities in Eq. 2.17 now only differ in their respective T_v. From the LFC a parcel keeps rising until it loses all its buoyancy. The work E_{CIN} required to lift the parcel to the LFC is given by the convective inhibition (CIN) (Eq. 2.19) which determines if free convection can start in the first place. E_{CIN} is found by integrating over the buoyancy acceleration from the level z_0 at which lifting starts (e.g. the surface) to the LFC. Using Eq. 2.17 and Eq. 2.18, E_{CIN} can computed via Eq. 2.20.

$$E_{CIN} \equiv \int_{z_0}^{z_{LFC}} \frac{B}{\rho}\, dz, \tag{2.19}$$

$$\Leftrightarrow E_{CIN} = \int_{z_0}^{z_{LFC}} g\frac{T_v - \bar{T}_v}{\bar{T}_v}\, dz. \tag{2.20}$$

CIN can be interpreted as a measure of stability with respect to vertical lifting. A large CIN results in a low probability of an air parcel reaching the LFC, often resulting in vertically small, non-precipitating clouds often called fair weather cumulus which are a result of this *shallow* convection. If convection reaches higher - roughly above the 500 hPa level - this is typically

termed *deep* convection (Dc). Once the convection reaches the LFC, the height the cloud can reach depends on the convective available potential energy (CAPE), given by E_{CAPE} analog to E_{CIN}.

$$E_{CAPE} = \int_{z_{LFC}}^{z(B=0)} \frac{B}{\rho}\, dz. \tag{2.21}$$

The larger the CAPE, the higher the cloud top will be and remaining kinetic energy may even cause the parcel to overshoot the level at which $B = 0$. Furthermore, in a stable boundary layer, e.g. when large-scale lifting cannot overcome CIN, CAPE may build up for a long time. Especially stable layers below a temperature inversion where the sign of the temperature gradient turns positive at the inversion level, prevent further convection. If the CAPE is large above a stable layer, dissolution of the stable layer may result in strong convection and thunderstorms. The CAPE is an important quantity that can be used in the closure of the equations describing subgrid mass fluxes (*parametrization*) in GCMs.
Beyond larger-scale convective flows, clouds are also strongly affected by turbulent processes on smaller scales. The mixing of the moist cloud air with surrounding dry air is called *detrainment* if moist air leaves the cloud and *entrainment* if dry air enters the cloud. Convective clouds are strongly affected by this mixing, which causes evaporative cooling. This reduces the cloud's buoyancy and condensed water content and moistens the surrounding air. Entrainment at the lower cloud levels reduces buoyancy, slowing down the updraft and thus enabling stronger detrainment at higher levels. In GCMs, de/entrainment are part of parametrizations typically as empirically adjusted rates of which the values are *tuned* so that the convection behavior matches observations (Crueger et al., 2018; Hourdin et al., 2021; Mauritsen et al., 2012; Schmidt et al., 2017). Updrafts always cause a corresponding downdraft, which can be located next to the cloud, where the air cooled from detraining cloud elements sinks. For strongly precipitating clouds the downdraft might also form from the air cooled by evaporating rain. If the downdraft is strong enough it can essentially lift surface air, potentially triggering new convective events. This results in a mixing effect on the troposphere that efficiently redistributes heat and moisture.

2.1.2. Cloud Types

Since many cloud properties are directly and indirectly related to atmospheric height the focus will be separately on low- (max. 2 km), middle- (2 km to 8 km) and high-level (3 km to 18 km) cloud types (WMO, 2023). To be consistent with the observation of clouds from space, the levels will be distinguished by typical cloud top height. The focus of this thesis will be on eight cloud types(see Table **3.2**), each of which will be described in this section in terms of appearance, dynamics and regions of prevalent occurrence. Six of the eight types are defined equivalently to the World Meteorological Organization (WMO) cloud genus of the same name.

However, the three cirriform cloud genera from the WMO are combined here to form cirrus (Ci), and the cumulonimbus genus is roughly equivalent the deep convective (Dc) type. It should be noted that even though this section focuses on individual types, these rarely appear in isolation. Rather, many clouds are part of larger organized systems and/or are products of the evolution of other types.

Low Clouds

Most of the Earth's cloud cover consists of low clouds, especially over the oceans (Hahn and Warren, 2007). Stratocumulus (Sc) is the most common type globally and can occur very persistently over the subtropical oceans (Wood, 2012). Here, Sc occur mostly in the marine boundary layer below inversions, prevalent in regions where large-scale dynamics induce subsidence, like the descending branch of the Hadley cell. The clouds are capped by the inversion and radiative cooling at the cloud top is the main driver of (downward) convection. The appearance of Sc is usually a shallow cloud field with a large horizontal extent of several hundreds of kilometers, with potentially many clear spots in between open cells. Most Sc are relatively optically thin, manifesting as light gray or white, non-precipitating clouds. One way of Sc formation is by radiative cooling of clear sky, such that water vapor condenses in the cooler air, the suspended liquid forming a cloud. If a cloud forms this way in a stable layer, it often first becomes the Stratus (St) cloud type. St is defined as a flat, grayish cloud sheet in a stably stratified environment. However, this St layer can quickly become unstable, e.g. through (further) radiative cooling at the cloud top, causing convection and thus breaking up into Sc. St is rarely encountered as a persistent sheet, as radiative cooling at the top and heating at the bottom become stronger the ticker the St layer becomes. For these reasons, the transition between fields of St and Sc is usually smooth. Dissipation of an Sc layer can happen through loss of moisture to the surrounding air through detrainment or entrainment, where dry air gets mixed into the cloud causing evaporation. If the convection at the cloud top stalls, e.g. because solar heating of the cloud top compensates for the radiative cooling, the moisture flux from the surface is interrupted, causing the cloud to start to dissipate. This can lead to the breakup of the cloud layer and inversion, resulting in cells of cumulus (Cu) convection. In turn, existing Cu can spread horizontally where surface updrafts and a strong inversion above the boundary layer coincide, forming an Sc layer.
Cu are prototypical convective clouds formed from transport of moisture lifted up (from the surface) with air masses driven thermally or by large-scale flows. In the context of this thesis, deep convection is treated as a separate cloud type, while Cu only refers to shallow cumuli. These are typically $\mathcal{O}(1\text{ km})$ in size in any direction. Over the oceans, Cu are prevalent in the trade wind regions, where they tend to form out of the subtropical Sc decks. Small (fair weather) Cu caused by thermal updrafts from solar heating of the surface are often found over land. If the updrafts cease, Cu will eventually dissipate through turbulent mixing with

the ambient air, since due to their convective nature they are susceptible to en/detrainment. All three of these low types (St, Sc, Cu) are primarily composed of liquid particles as the lower atmosphere usually remains warmer than the temperatures required for ice formation, ice particles can, however, form in very cold conditions.

Midlevel Clouds

Three cloud types with mid-level cloud tops are distinguished here: Altostratus (As), Altocumulus (Ac), and Nimbostratus (Ns). Taking the surface-observer definitions of the original cloud atlas, the "alto-" prefix denotes clouds with mid-level base and, as their names suggest, As and Ac are higher-altitude versions of St and Cu/Sc, respectively (Howard, 1803). The As type displays similarities to both Ns and Ac. Ns and As both manifest as contiguous cloud sheets or layers and often occur in mid or high latitudes. While Ns almost always precipitates, this is rarely the case for the thinner and less opaque As type. This lack of precipitation also applies to most Ac, but they are mostly comprised of liquid water particles, while As contains mostly ice. Therefore, As are often found to be formed by freezing of the supercooled droplets in Ac. This is also the reason why Ac is often more shallow than As (Sassen and Wang, 2011). Ns rarely occur alone but are usually the product of convective cloud processes, i.e. originating from Cu, byproducts of deep convection (Dc) or parts of mesoscale convective system (MCS) and midlatitude storms. In Ns, cloud particles of both phases can occur, as ice crystals often fall from the associated Dc system into the stratiform region to then melt and precipitate out. In fact, Ns and As, in combination with Dc contribute the majority of the total atmospheric ice (Austin et al., 2009). While Ns can be $\mathcal{O}(10$ km) deep, they no longer contain the strong updrafts found in Dc and can cover an extensive horizontal area. This makes Ns clouds the primary source of stratiform precipitation in which the vertical velocity of air is much smaller than the typical fall speed of ice particles (Houze, 2014, Ch. 6). Ac can resemble both the more sheet-like structure of Sc as well as the interspersed cells of Cu. Over the continents, orography is a strong driver of Ac formation, while over tropical oceans it is often a product of detrainment from deep convection (Sassen and Wang, 2011).
Because of their frequently mixed phase and potentially extensive thickness, mid-level clouds are difficult to measure remotely, requiring combination of different sensors to accurately measure their properties (Sassen and Wang, 2011).

High Clouds

As mentioned in Section **2.1.1**, when the convection forming Cu clouds reaches heights above 500 hPa, these clouds are typically termed deep convective (Dc). Dc clouds are frequently associated with heavy rain and thunderstorms and redistribute large amounts of heat and

moisture in the atmosphere. Dc is one of the major sources of (convective) precipitation, as the ice particles in the strong Dc updrafts increase in size until they become too heavy to be suspended in the air. Convection above the LFC is therefore self-enhancing, as the latent heating through condensation and the effective decrease in density through precipitation increase buoyancy of the air. As shown by Eq. 2.21, this buoyancy determines the height of the Dc cloud's top, where wind shear, displacement from the continuing updraft and detrainment can cause the development of an *anvil*. The anvil is a thin layer of cloud atop the convective tower with increased horizontal extent. Thicker and lower anvils can form Ns sheets behind the Dc column. Older anvils can develop into the cirrus (Ci) cloud type, which denotes very high ice clouds appearing as white clouds appearing as anything from smaller patches to larger layers. In situ measurements have shown that freezing of haze particles through homogeneous nucleation is the dominant ice crystal formation process (e.g. Cziczo et al., 2013). Ci are usually very thin ($\sim$1.5 km) and fairly transparent, with even the most opaque Ci having a $cod \approx 3$. Ci can occur as high as the tropopause transition layer (TTL) where they are often so thin that they are invisible to the naked eye. These *subvisible* Ci and also slightly more opaque Ci are only detectable with special sensors such as lidar, making it difficult to comprehensively study these clouds with remote sensing techniques.
Due to higher surface temperatures and large-scale flows enabling deeper convection, Dc and, as a consequence, anvil Ci are frequently found in the deep tropics. In the tropics and subtropics several Dc cores often form deep, organized cloud systems with both convective and stratiform components (MCS) .
High-top clouds like Ci and Dc are important for the Earth's energy budget because they contribute significantly to the greenhouse effect and at the same time emit only a small amount of thermal radiation to space (see Eq. 2.4 and Eq. 2.7). Detecting and distinguishing various cloud types is therefore important for a complete understanding of the climate system.

2.2. Satellite Instruments and Retrieval Methods

This section provides an overview of how some of the cloud radiative properties introduced in Section **2.1.1** can be exploited to determine the composition of clouds with remote sensors. There are three main ways of observing clouds and the atmosphere from space: (1) measuring reflected sunlight, (2) measuring emitted infrared radiation and (3) measuring the reflection of radiation emitted by the instrument. Especially for the first two (passive) methods the retrieval of physical properties usually requires the solution of an inverse optimization problem. This type of retrieval involves finding physical variables that allow a numerical model to closely approximate the measured radiation. The satellites carrying the instruments usually orbit the Earth in a stable low earth orbit (LEO) or a geostationary orbit (GEO). In these stable orbits the centrifugal force matches the gravitational force of the Earth acting on the satellite. In a GEO the satellite is fixed at one longitude at the equator at all times, enabling consistent

measurements of roughly a third of the Earth's surface as a disk projection. The majority of Earth observation satellites are in LEO, specifically a sunsynchronous orbit, where the satellite always crosses the equator at the same local time. When the equator is crossed from south to north, this is called ascending node, while crossing it from north to south is called descending node.

2.2.1. Retrievals from Passive Sensors

Passive sensors are detectors of electromagnetic radiation of various wavelengths. For most physical properties of interest, retrievals for passive sensors rely on inverting a forward model, given by a solution of the radiative transfer equation:

$$\frac{dI(\tau)}{d\tau} = I(\tau) - J(\tau), \tag{2.22}$$

which is - as stated above - formulated without radial coordinates, such that the source term J can be expressed only in terms of the scattered sunlight as well as the solar irradiance. For reflection-based methods solutions to Eq. 2.22 are usually found by assuming single scattering and a plane-parallel atmosphere such that the reflectance (measured radiance relative to solar irradiance) can be expressed in terms of viewing geometry as well as *cod* and *cer* (Stephens and Kummerow, 2007). Some of the variables that determine the reflectance are known from instrument geometry or can be parametrized to constrain the solutions, but *cod* and *cer* need to be retrieved. The reflectance is therefore computed for various combinations of *cod* and *cer*, and the solutions to this forward model are stored in lookup tables (LUTs), which are later used for the inversion.

As the optical properties of atmospheric particles and gases depend highly on the wavelength, detectors for different wavebands are combined to be able to accurately observe a wide range of atmospheric phenomena. Using multiple wavelengths is helpful because scattering and absorption properties depend on wavelength (Eq. 2.15). In the bispectral reflectance method (Nakajima and King, 1990) the correct reflectance is found by exploiting that it varies most strongly with *cod* at smaller wavelengths and with *cer* at larger wavelengths. Furthermore, it uses the relationship between *cod* and *cer* to compute the cloud water path, as liquid water content (*lwc*) and liquid water path (*lwp*) depend on the liquid water density ρ_{lq} and particle size distribution $N(r)$ (adapted from Liou (2002, p. 373)):

$$lwp = \Delta z \cdot lwc = \Delta z \frac{4\pi\rho_{lq}}{3} \int_0^\infty r^3 N(r) dr, \tag{2.23}$$

while the *cod* (Eq. 2.13), depends on the attenuation efficiency Q_λ through k_λ, which is assumed to be homogeneous. Using the thickness of the atmospheric column Δz and the geometric cross section $A = \pi r^2$ of the attenuating particles gives:

$$k_\lambda \rho_{lq} = \int_0^\infty Q_\lambda \pi r^2 N(r) dr, \tag{2.24}$$

$$\Rightarrow \tau_\lambda = \int_{z_0}^{z_\infty} \rho_{lq} k_\lambda dz = \Delta z \int_0^\infty Q_\lambda \pi r^2 N(r) dr. \tag{2.25}$$

With $Q_\lambda \approx 2$ outside of the Rayleigh regime and using the definition of *cer* (Eq. 2.8):

$$\frac{lwp}{cod} = \frac{4\pi\rho_{lq} \int r^3 N(r) dr}{3 \cdot 2\pi \int r^2 N(r) dr}, \tag{2.26}$$

$$lwp = \frac{2\rho_{lq}}{3} cer \cdot cod. \tag{2.27}$$

For ice (ice water path (*iwp*)), an analog to Eq. 2.27 can be obtained to compute the total amount of condensed water in the columns (cloud water path (*cwp*)) as the sum of *lwp* and *iwp*. The problem posed by Eq. 2.27 has no unique solution for a given *lwp*, but retrieving *cer* and *cod* simultaneously using the bispectral method eliminates this issue. Many retrieval algorithms for passive sensors use variations of this technique, with varying ways of finding the best approximation (Stephens and Kummerow, 2007). The difference between radiances computed using the forward model is minimized with methods ranging from least squares (Nakajima and King, 1990) to brute-force iteration (Nakajima and Nakajma, 1995) to Bayesian inversion techniques (McGarragh et al., 2018; Sus et al., 2018).

A widely used technique to retrieve cloud top height using passive sensors is so-called carbon dioxide (CO_2) slicing (Chahine, 1974; Smith and Platt, 2023). With this method, the dependence on cloud fraction and emissivity is eliminated by using two infrared channels in wavebands with high absorption from CO_2 and subtracting the clear sky radiance from the measurements in the respective channels. For large wavelengths, emissivities depend more strongly on temperature than on wavelength, such that identical emissivities can be assumed for both channels. Using CO_2 absorption bands eliminates the dependence on gas mixing ratios in the atmosphere, as the distribution of CO_2 is a known quantity. Notably this method also works at night as it does not depend on reflected solar radiation. Among the sensors using this technique is Moderate Resolution Imaging Spectroradiometer (MODIS), the data of which are the basis of the studies in this thesis (see Section **3.1.1**).

2.2.2. Retrievals from Active Sensors

Active sensors are instruments that measure the return signal of radiation emitted by the instrument itself. Two types of active sensors are distinguished: radio detection and ranging

(radar) and light detection and ranging (lidar), where radar uses radio- or microwaves, while lidar uses laser beams at shorter wavelengths in the visible or infrared. However, the basic principle of both types is the same. The instrument aboard the satellite emits radiation, which then gets absorbed or scattered by the atmosphere after which a detector aboard the satellite measures the backscatter signal. The intensity, time delay and shift in frequency and polarization of the backscatter are then measured to derive information about the atmospheric composition. The type and construction of the sensor therefore determines which phenomena can be detected. Unless the instrument is designed to detect specific trace gases, the operating wavelength is limited to bands in which atmospheric gases are not strongly absorbing. For spaceborne radars, the millimeter range is used for cloud observation, as the antennas are not required to be too large. This means cloud radars operate in the Rayleigh regime, where the scattering cross section for a single particle is

$$\sigma_{s,\lambda} = \frac{\pi^5 |K|^2 D^6}{\lambda^4}, \tag{2.28}$$

with K depending on the refraction index and D being the particle diameter (Liou, 2002, ch. 7.6). This yields the backscattering coefficient

$$\beta = \frac{\pi^5 |K|^2}{\lambda^4} \int D^6 n(D) dD \equiv \frac{\pi^5 |K|^2}{\lambda^4} Z, \tag{2.29}$$

with the reflectivity factor Z. The power of the measured backscatter from an object at distance r can then be calculated as (Liou, 2002, ch. 7.6)

$$P_s(r) = \frac{C |K|^2}{r^2} Z, \tag{2.30}$$

with C an instrument constant depending on transmitter power, wavelength and geometry. The exact value of C is obtained by calibration. Since the details of the composition of the observed atmospheric column are unknown, K needs to be estimated. Then the measured radiation in the form of P_s can be used to infer Z, informing about the observed slice of the atmosphere, e.g. a cloud. Since Z depends only on the distribution of the size of the scattering particles, it can be used to compute cloud- and precipitation-related properties, for example in retrievals for Cloud Profiling Radar (CPR) onboard CloudSat (CS). Also, the CPR reflectivity factor and the Cloud-Aerosol Lidar with Orthogonal Polarization (CALIOP) backscatter coefficient are important inputs to the DARDAR products (Delanoë and Hogan, 2010), which provide cloud properties from both of these active sensors. Note that Eq. 2.30 assumes Rayleigh-scattering as well as a homogeneous composition of the atmosphere at each distance r, where the former is not valid for lidar wavelengths and the latter is rarely the case in the real atmosphere. Equation 2.30 is therefore only an approximation of operational retrieval algorithms. However, Eq. 2.29 shows that the backscatter decreases with the fourth power of the wavelength, which is why radar is more suitable than lidar for detecting large droplets and precipitation-sized particles, while lidar can detect very thin and subvisible clouds.

For small wavelengths in or near the visible spectrum and large droplets or ice crystals with $cer \gtrsim 10\,\mu m$ the size parameter is $x >> 1$, such that geometric optics apply for lidar. In this regime, change of polarization through in-particle refraction plays a significant role. For spherical particles light refracted back in the incident direction will not change its polarization, but for non-spherical ice particles, depending on the exact shape and the incident light's direction, a change in polarization is likely. This fact is exploited in depolarization lidars, where the laser is vertically polarized. Measuring the power of backscatter signal in this polarization plane and the orthogonal (horizontal) plane is used to infer the amount of non-spherical particles in the atmosphere. Similarly, radars can use depolarization effects to distinguish larger droplets from slightly non-spherical raindrops or snowflakes to detect the presence of precipitation. The presented theory holds only for the single-scattering approximation, where it is assumed that each beam only undergoes a single scattering event before reaching the detector. Measured reflectances deviate from this theory even for passive sensors, but active depolarization measurements are especially affected. Properties of mixed-phase clouds are therefore hard to determine even with lidar measurements.

Many of the methods presented in this section, while they are being tuned and adapted to take advantage of ever-improving instruments, are decades old at their core. Some are therefore already being replaced by new methods from the field of ML (see also Chapter **3**).

2.3. Machine Learning

As a subcategory of artificial intelligence (AI), ML encompasses algorithms that can adapt to data when implemented on a machine without hard-coded knowledge (Goodfellow et al., 2016a, p. 2). The process of adaptation in which the learning algorithm - hereafter *model* - learns to optimize for defined targets, is called *training*. An important distinction is made for training structure: in *supervised* training the model $\mathbf{f}$ learns to approximate an unknown mapping $\mathbf{f}_{data}$ from **available** inputs $\mathbf{x}$ and outputs $\mathbf{y} \equiv \mathbf{f}_{data}(\mathbf{x})$, such that $\mathbf{f}(\mathbf{x}) \approx \mathbf{f}_{data}(\mathbf{x})$. If samples for $\mathbf{y}$ are **unavailable**, the problem is *unsupervised*, which requires the definition of alternative targets for the model to learn. Note that due to sampling of observations, measurement uncertainties and similar issues $\mathbf{f}_{data}$ is not necessarily equal to the true physical relationship $\mathbf{f}_{true}$ between $\mathbf{x}$ and $\mathbf{y}$. ML problems usually fall into one of two further categories: *classification*, for which $\mathbf{y}$ is categorical/discrete and *regression*, for which $\mathbf{y}$ is continuous. For each problem or small set of problems, specific algorithms are designed and optimized. Solutions to ML problems are typically not available in closed form, with few exceptions like linear regression. Instead, some criterion is optimized iteratively. While convergence of the iterative process to a global optimum is not always guaranteed, finding a local optimum is usually sufficient to solve a given problem (Choromanska et al., 2015; LeCun et al., 2015).

There exist a vast amount of ML models, which are further categorized into so-called classical ML and DL. Classical ML contains methods that can learn without the need for neural networks and backpropagation (see Section **2.3.2**). The only relevant classical method for this thesis is the Random Forest regression model (RFRM) model, explained in Section **2.3.1** (Breiman, 2001). Since DL models are a much larger part of this work, a number of DL models will be presented with the necessary background in Section **2.3.2**.

2.3.1. Random Forests

Simply put, a Random Forest is a large number of slightly different decision trees, each applied to the same problem. This design is based on the assumption that an ensemble of weak learners can act as a strong learner, which is called *boosting* (Schapire et al., 1998). A decision tree is a fairly simple method, in which a dataset is split iteratively in a way that optimizes a specified splitting criterion (Kotsiantis, 2011). The solution that optimizes the criterion is typically found via a full search of all possibilities of splitting the data. During inference, a new sample is passed through the tree following the learned decisions until it ends up in a *leaf node*, in which the splitting process terminates for a given sample. While decision trees are easy to build and easily explainable, they lack accuracy and can easily overfit to the training data. For Random Forests, the addition of stochastic modifiers considerably increases the generalization capability by achieving low-bias predictions with minimal variance (Breiman, 2001). One of these modifiers is *bootstrapping*, where the training data are randomly resampled (with replacement) individually for each tree, such that the total amount of samples seen by each tree is equal to the number of available samples but with a modified distribution. If a smaller subset is sampled without replacement, the technique is called bagging (**b**ootstrap **agg**regat**ing**) (Breiman, 1996). For each tree, a random subset of the elements (*features*) of $\mathbf{x}$ is selected as input for that tree. The number of selected features is often chosen to be close to $\sqrt{N}$ for classification and $N/3$ for regression, where N is the maximum available number of features (Breiman, 2001). Finally, the output of the Random Forest model is the aggregate of the output of all individual trees. For Random Forest classifiers, this aggregate is determined by majority voting while for Random Forest regression models (RFRMs) it is simply the arithmetic mean. From here on, only the RFRM variant will be discussed.
During training the splitting is performed such that each split optimizes a score $\mathcal{S}$ given by the error function $\mathcal{C}$. The splitting process in a branch terminates when each partition contains a previously defined minimum number of samples, thus becoming a leaf. The maximum number of splits per branch (*depth*) can be specified in advance to terminate splitting even earlier. Simple functions like mean squared error (MSE) or mean absolute error (MAE) are usually

sufficient as error function $\mathcal{C}$. The score for a split separating the training dataset with inputs $\mathbf{x} \in \mathbb{X}$ and outputs $\mathbf{y} \in \mathbb{Y}$ into subsets i and j is then given by:

$$\mathcal{S}\left[\mathbb{X}, \mathbb{Y}\right] = \frac{1}{2K} \sum_{y_k \in \mathbb{Y}_i} \mathcal{C}(\mathbf{y}_k, \bar{\mathbf{y}}_i) + \frac{1}{2M} \sum_{y_m \in \mathbb{Y}_j} \mathcal{C}(\mathbf{y}_m, \bar{\mathbf{y}}_j), \tag{2.31}$$

where the overbar denotes the mean and K and M are the number of elements in the respective split. In each branch, the split s divides the previous partition $(s-1)$ such that it minimizes $\mathcal{S}_s = \mathcal{S}\left[\mathbb{X}_{s-1}, \mathbb{Y}_{s-1}\right]$, or equivalently maximizes the decrease in score $\Delta\mathcal{S}_s$ induced by the split s with respect to the mean error in the data at $s-1$ (Eq. 2.32).

$$\Delta\mathcal{S}_s = \frac{1}{2K} \sum_{y_k \in \mathbb{Y}_{s-1}} \mathcal{C}(\mathbf{y}_k, \bar{\mathbf{y}}_i) - \mathcal{S}_s. \tag{2.32}$$

Each individual split is performed on a single feature n, such that for a threshold value $z_s \in \mathbb{R}$:

$$x_n < z_s, \ \forall \ \mathbf{x} \in \mathbb{X}_i,$$
$$x_n > z_s, \ \forall \ \mathbf{x} \in \mathbb{X}_j.$$

As the score S is only a measure of the impurity of the leaf nodes, it is not useful for evaluating the RFRM performance. A more appropriate way to evaluate an RFRM is by using the coefficient of determination, also known as R^2 or R2-score:

$$R^2 = 1 - \frac{\sum \left(|\mathbf{f}(\mathbf{x}_i) - \mathbf{y}_i|_2\right)^2}{\sum \left(|\mathbf{y}_i - \bar{\mathbf{y}}|_2\right)^2}, \ (\mathbf{x}_i, \mathbf{y}_i) \ \in \ \{\mathbb{X}, \mathbb{Y}\}. \tag{2.33}$$

R^2 quantifies to what extent the variance of the data is accounted for by the predictions. Note that for vector-valued targets $\mathbf{y}$ as is the case here, R^2 can be defined in two ways. The definition in Eq. 2.33 computes the score by feature, i.e. element n of $\mathbf{y}$, while switching the order of sum and L2-norm would result in the average the score per sample i. Applying both definitions can offer additional insight into the performance of $\mathbf{f}$. Using Eq. 2.33 or a similar score, the generalization capabilities of an RFRM can be estimated without the need of a holdout set by applying it to the out-of-bag samples. That means each tree in the RFRM is applied to the respective samples that were excluded during the bagging procedure, which are different for each tree. These results can be used to compute performance metrics R_{oob} on the out-of-bag samples, i.e. the coefficient of determination.
In contrast to a single decision tree, the predictions coming from an RFRM are hard to track and explain, as there are typically $\mathcal{O}(100)$ trees with varying structures. A straightforward way to explain an RFRM's predictions is to assess the *feature importance* by computing the sum of all $\Delta\mathcal{S}$ induced by splitting on each feature. While this measure is easy to compute, it is also subject to several problems and therefore rarely used to explain an RFRM predictions. Instead, Breiman (2001) suggests computing the *permutation importance* on the out-of-bag

samples to obtain an estimate of which features have the highest influence on the predictions produced by the RFRM. To compute the permutation importance of each feature n, the values of only this feature for the different samples are randomly permuted between samples, after which R_{oob} is computed again. The change in R_{oob} for each feature indicates its importance for the RFRM to make its predictions.
The RFRM therefore combines several qualities that make it an attractive model for many applications: the trees can easily be trained in parallel, bagging ensures small susceptibility to noise, good generalization and little bias, while the permutation importance offers some measure of explainability without requirements for external algorithms. Furthermore, the predictions of the RFRM are normalized if the training samples are normalized because outputs are just averages over all samples in a leaf node. Downsides include the high number of hyperparameters and that the trained RFRM can easily require several gigabytes on disk if the number of trees is large or the depth is high.

2.3.2. Deep Learning

Similarly to RFRMs, neural networks are designed by combining many instances of a simple model to achieve vastly increased performance. While for an RFRM the individual objects are trained separately and act as an ensemble, the individual *neurons* in a neural network (NN) can influence each other. A neuron is defined as a function f (Eq. 2.34) that determines the output y from the input x by multiplying the weight w and adding bias b, and finally applying a nonlinear activation function z. Equation 2.34 is easily vectorized for multidimensional inputs $\mathbf{x} \in \mathbb{R}^n$ and outputs $\mathbf{y} \in \mathbb{R}^m$ to obtain a neural layer (Eq. 2.35), with weights $\mathbf{W} \in \mathbb{R}^{m \times n}$ and biases $\mathbf{b} \in \mathbb{R}^m$.

$$y = z(w \cdot x + b), \tag{2.34}$$

$$\mathbf{y} \equiv \mathbf{f}(\mathbf{x}) = \mathbf{z}(\mathbf{W} \cdot \mathbf{x} + \mathbf{b}). \tag{2.35}$$

Thus, a neural layer is a basic linear model, deliberately made nonlinear through the application of $\mathbf{z}$. The requirement for $\mathbf{z}$ becomes apparent when several neural layers are combined as

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}_3\left(\mathbf{f}_2\left(\mathbf{f}_1(x)\right)\right), \tag{2.36}$$

to obtain a multi-layer perceptron (MLP) (Rosenblatt, 1961), in which any layer that is not directly connected to input or output, $\mathbf{f}_2$ in this case, is called a *hidden layer*. With $\mathbf{z} \equiv \mathbf{z}_{id}(\mathbf{x}) = \mathbf{x}$, any output of the MLP would be a linear combination of the weights, biases and inputs. Then, if the data $\mathbf{x}$ are not linearly separable in terms of $\mathbf{y}$, neither will $\mathbf{f}(\mathbf{x})$, thus limiting its complexity. The possibility for $\mathbf{f}$ to closely approximate a nonlinear function is therefore only given if both multiple layers and a nonlinearity $\mathbf{z}$ are used (LeCun et al., 2015). Under this condition, an MLP can in theory serve as a universal function approximator, provided that

the number of hidden neurons is large enough (Hornik et al., 1989). The choice of function for $\mathbf{z}$ influences the performance of the NN, and specific choices are sometimes mandated by the application. A simple yet popular choice for $\mathbf{z}$ is the rectified linear unit (RELU)

$$z_{RELU}(x) = \begin{cases} x, & \text{if } x > 0 \\ 0, & \text{if } x \leq 0 \end{cases}. \tag{2.37}$$

An MLP is synonymous with a *dense* or *fully connected* NN because all neurons (or *nodes*) are densely connected to all nodes in the adjacent layers via weight multiplication. The more layers the NN has, the *deeper* it is, the more neurons each layer has, the *wider* it is.

The weights and biases are the free parameters $\boldsymbol{\theta}$ of the NN $\mathbf{f}_{\boldsymbol{\theta}}$. In the following, "weights" refers to all elements of $\boldsymbol{\theta}$. For $\mathbf{f}_{\boldsymbol{\theta}}$ to achieve a desired mapping, the weights need to be trained. This is achieved through *gradient descent* (GD) and *backpropagation*. GD specifies the process of updating the network's weights in the direction of the gradient of some loss function $\mathcal{L}$ with respect to the weights. $\mathcal{L}$ provides a measure of how well $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x})$ approximates the ground truth $\mathbf{y}$. Backpropagation refers to the way the gradients are computed in a deep neural network, by applying the chain rule to iteratively compute each layer's derivative. The trainable weights $\boldsymbol{\theta}$ are randomly initialized and then updated from step i to the next step $i+1$ as:

$$\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \lambda \nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})), \tag{2.38}$$

where the *learning rate* $\lambda \in (0, 1)$, typically, helps to smooth the convergence towards a minimum of $\mathcal{L}$. To compute the gradient in Eq. 2.38, the chain rule is applied, iteratively obtaining the derivatives layer by layer. For the MLP in Eq. 2.36 the gradient of the loss becomes:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\mathbf{f}_{\boldsymbol{\theta}}, \mathbf{y}_t)) = \frac{\partial \mathbf{y}_1}{\partial \boldsymbol{\theta}} \frac{\partial \mathbf{y}_2}{\partial \mathbf{y}_1} \frac{\partial \mathbf{y}_3}{\partial \mathbf{y}_2} \frac{\partial \mathcal{L}}{\partial \mathbf{y}_3}. \tag{2.39}$$

In Eq. 2.39, the derivatives with respect to previous layers $\frac{\partial \mathbf{y}_i}{\partial \mathbf{y}_{i-1}}$ can be further expanded in terms of weights and activation function:

$$\frac{\partial \mathbf{y}_{i+1}}{\partial \mathbf{y}_i} = \sum_j \theta_{ij} \frac{\partial \mathbf{z}}{\partial \mathbf{y_i}} \tag{2.40}$$

To minimize $\mathcal{L}$, backpropagation steps are repeated until $\mathcal{L}$ converges, where each repetition is called an *epoch*. Epochs are usually further split into minibatches (or simply *batches*), as in most applications, the training data do not fit into memory. As the gradients for the batches differ, the computed gradients will be slightly noisy, which is why this procedure is called

stochastic gradient descent (SGD)[2]. The forward and backward passes, i.e. the computation of $\mathcal{L}(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}))$ and $\nabla_{\boldsymbol{\theta}}\mathcal{L}$ are then consecutively applied to each batch of the data. Under the assumption that the training data are independent and identically distributed, the expected value of the SGD gradient is equal to the deterministic gradient.

GD methods are iterative algorithms to find a solution for $\nabla_{\boldsymbol{\theta}}\mathcal{L}(\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{y})) = 0$ in terms of $\boldsymbol{\theta}$, i.e. to find an extremum of the loss, which is not available in closed form for most NNs. However, Eq. 2.38 does not come with a guarantee that the iteration converges or that convergence would find the global minimum. However, it has been shown for several sets of assumptions that both algorithms converge to useful solutions: GD converges to a global minimum for overparametrized networks, i.e. networks with redundant parameters (Du et al., 2019), and cases in which SGD converges include convexity of $\mathcal{L}$ in terms of $\boldsymbol{\theta}$, or if the learning rate decreases sufficiently quickly (Bottou, 1998). Furthermore, the optimization getting stuck in a local minimum or plateau is unlikely due to the slightly noisy nature of SGD. This means that in theory, SGD will almost surely find a near-perfect approximation of the mapping $\mathbf{f}_{data} : \mathbb{X} \to \mathbb{Y}$ represented by the training data. It is therefore commonly assumed that SGD and related algorithms are suitable to optimize the parameters of most NN architectures. Since the gradients differ between the batches, SGD often cannot find the global minimum of $\mathcal{L}$ (e.g. overfit), but that is actually a desirable feature that usually increases generalization performance.

Optimizing Neural Networks

Training ML models is subject to two major challenges, which are especially relevant for NNs: (1) lack of convergence, when the NN struggles to improve the results beyond a certain point, and (2) overfitting, where the model learns to reproduce the training data instead of the underlying function $\mathbf{f}_{true}$. Improving the overall results through faster and more stable convergence can usually be achieved through modifications of SGD and associated hyperparameters. The Adam algorithm is a popular optimizer that improves SGD by introducing two moment-decay parameters (Kingma and Ba, 2015). These control the effect of previous gradients on the magnitude of the gradient update, and thus smooth out and stabilize the stochastic optimization process. This way, Adam achieves better convergence properties than comparable algorithms for convex functions (Kingma and Ba, 2015). Hyperparameters that can be tuned to achieve more effective training include decreasing learning rate λ, cyclic learning rate (Smith, 2017), the choice of nonlinearity function $\mathbf{z}$ and hyperparameters of the optimizer, such as the decay parameters of the Adam optimizer.

[2]Some literature calls GD over the complete data set "batch GD", in contrast to "SGD on **mini**batches". This batch vs minibatch nomenclature seems confusing. Therefore (deterministic) GD and (batch) SGD are distinguished here and the word minibatch is not used.

Techniques that counteract overfitting are called *regularizers* and ensure that the trained model is able to generalize to unseen data. A popular method is L2-regularization, where the L2-norm of the weights θ is added as a penalty to the loss $\mathcal{L}$. This ensures weights of small magnitude and therefore a smoother response to a wide range of inputs. Similar results can be achieved with *dropout*, in which individual weights θ_i are set to 0 with a certain probability during each forward pass (Srivastava et al., 2014). This prevents the NN from memorizing individual input/output pairs. Another way to achieve this is by interrupting the training procedure once the loss on an independent validation set no longer improves (*early stopping*). Additionally, some problems can have problematic regions in the loss hyperplane, e.g. plateaus with vanishing gradients. In this situation, an appropriate initialization of the weights θ is crucial to find a good optimum with any form of GD. Therefore it might be required to run the same model with various random initializations to achieve good results.

Convolutional Neural Networks

While dense NNs can in principle learn arbitrary mappings between data, they cannot exploit their fundamental structures, such as correlations of nearby values in time series or images (LeCun and Bengio, 1995). An example of this is image classification, where objects should be recognized independently of the background they appear on, where in the image they appear, or at which angle. This issue is addressed in convolutional neural networks (CNNs) by using shared weights that are applied equally at every location in the sample. Due to their ability to leverage the structure in image-like data, CNNs are part of multiple architectures used throughout this work. The core of a CNN is a usually square *convolutional kernel* $\mathbf{K} \in \mathbb{R}^{k \times k}$ that is applied element-wise to each $k \times k$ sized patch of the input. In the context of CNNs, the 2D discrete convolution at the pixel (i, j) is defined as:

$$(\mathbf{X} \circ \mathbf{K})_{i,j} = \sum_{h,l}^{k,k} X_{i+h,j+l}\ K_{h,l}. \tag{2.41}$$

A basic CNN layer $\mathbf{f}_{CNN}$ usually includes a nonlinearity $\mathbf{Z}$ and a bias $\mathbf{B}$:

$$F_{CNN}(\mathbf{X})_{i,j} \equiv \mathbf{Z}\left((\mathbf{X} \circ \mathbf{K})_{i,j} + B_{i,j}\right). \tag{2.42}$$

The shape of the result of Eq. 2.42, often called a *feature map* is defined by the valid indices of Eq. 2.41, i.e. the indices $i + h$ and $j + l$ that are not out of bounds. The fundamental operations of a CNN are illustrated in Fig. **2.3**, which shows schematically how the kernels move over all possible patches of the input image. The elements of the kernel $\mathbf{K}$ and bias $\mathbf{B}$ are the trainable parameters of the CNN. Since $\mathbf{K}$ is applied equally everywhere on $\mathbf{X}$, it is learned such that a sharp signal is obtained for the same structures, no matter where they appear. This *translation equivariance* is one of the primary advantages of a CNN. To detect

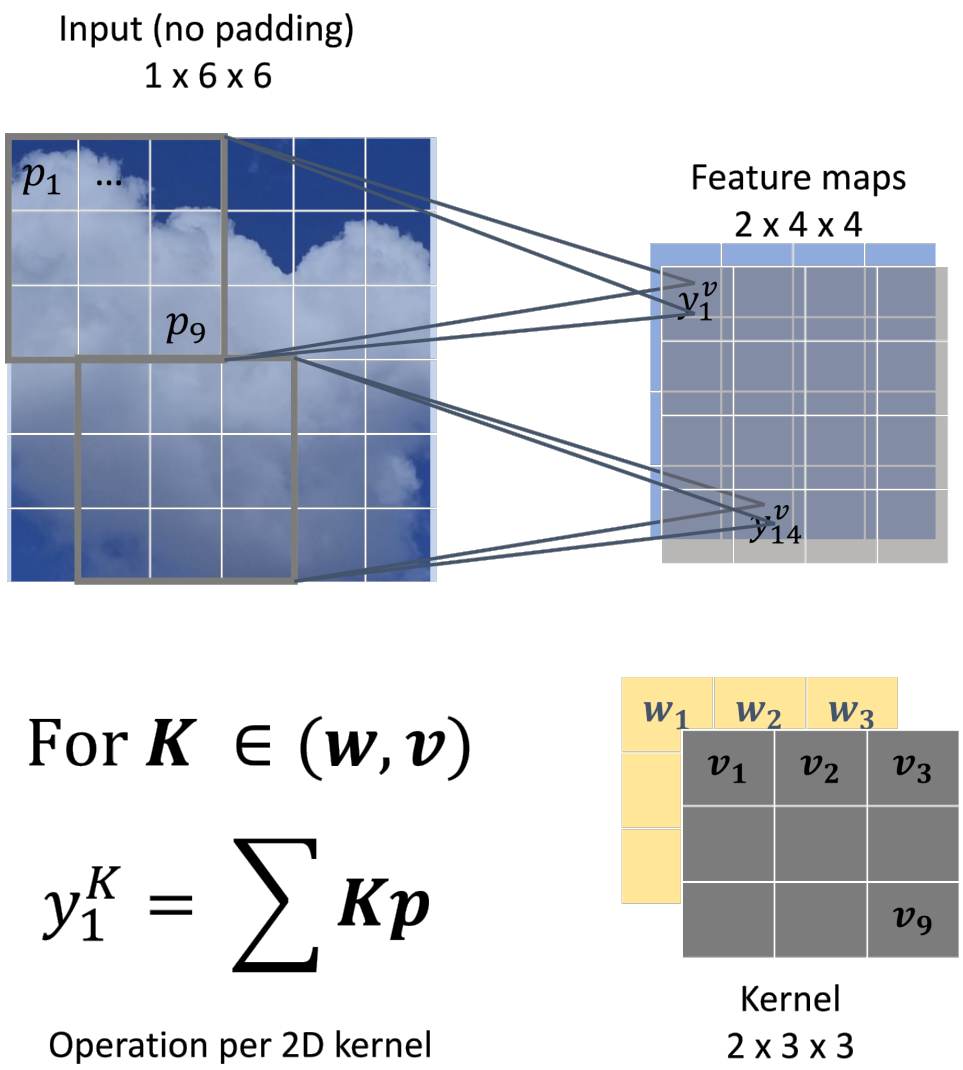


Figure 2.3.: Operation of a single CNN layer with two kernels on a two-dimensional image, without padding. Because the kernels are applied to exactly each valid patch of the input image, the feature maps are reduced by two pixels in each dimension.

multiple kinds of features, Eq. 2.42 is extended to produce several feature maps with different kernels. Other adaptations to Eq. 2.42 include *strides*, where the kernel is applied with less overlap, and *dilation*, where the kernel is not applied contiguously, i.e. kernel element $\mathbf{K}_{i,j}$ is multiplied with $\mathbf{X}_{i,j}$ and $\mathbf{K}_{i+1,j+1}$ with $\mathbf{X}_{i+d,j+d}$, with dilation width $d > 1$. Both of these strategies help detect features at different scales while reducing the number of operations and trainable parameters.

The feature maps are smaller than the input along the dimensions of the kernel, with the exact shape depending on kernel size, stride and dilation. To avoid this reduction in size, the input may be padded with zeros at the edges, but this can induce substantial edge artifacts. Another way to recover the input shape is to follow up one or more convolution layers with *transpose convolutions*, in which individual pixels in the input correspond to multiple entries in the output matrix. Transpose convolutions are an important part of the popular Residual

Network (ResNet) architecture, where the layers are trained to approximate $\mathbf{g}_{res}(\mathbf{x}) = \mathbf{g}(\mathbf{x}) - \mathbf{x}$ instead of the true underlying mapping $\mathbf{g}(\mathbf{x})$ (He et al., 2016). Therefore, $\mathbf{g}$ has to preserve the shape of $\mathbf{x}$, which is achieved by constructing *bottleneck blocks* that follow up each convolution with their transpose. To minimize only the residual, the input of each block is added to the output via *shortcut connections*. Using the residual formulation allows for the use of deeper networks without introducing training instabilities. The increased depth in turn increases the representation power of the network, which can then achieve better accuracy.

Generative Adversarial Networks

The NNs discussed above are trained to solve problems by optimizing a single cost function, but their fundamental architectures can be used to produce any kind of output. A prime example of this are generative adversarial networks (GANs) (Goodfellow et al., 2014), a framework in which multiple networks are trained in an adversarial way, i.e. with opposing objectives, to generate data similar to the training samples. Usually, these data are images and the GAN contains two CNNs, one called the generator $\mathbf{G}$, one the discriminator $\mathbf{D}$. The adversarial training is essentially a minimax game in which $\mathbf{G}$ is trained to generate convincing "fake" samples, while $\mathbf{D}$ is trained to discriminate between real and fake samples (Goodfellow et al., 2014). For generator and discriminator, training maximizes the respective log-likelihood $\mathcal{C}_{\mathbf{G}}$ and $\mathcal{C}_{\mathbf{D}}$, given by:

$$\mathcal{C}_{\mathbf{G}} = \mathbb{E}_{\mathbf{q}\sim P_{rand}} \left[\log\left(\mathbf{D}\left(\mathbf{G}\left(\mathbf{q}\right)\right)\right)\right], \tag{2.43}$$

$$\mathcal{C}_{\mathbf{D}} = \mathbb{E}_{\mathbf{x}\sim P_{data},\mathbf{q}\sim P_{rand}} \left[\log\left(\mathbf{D}\left(\mathbf{x}\right)\right) + \log\left(1 - \mathbf{D}\left(\mathbf{G}\left(\mathbf{q}\right)\right)\right)\right]. \tag{2.44}$$

In this basic GAN, the fake samples are generated from random noise $\mathbf{q}$ sampled from P_{rand} and compared to real samples from P_{data}. Instead of random noise, $\mathbf{q}$ can also be meaningful data related to $\mathbf{x}$ or class labels $\mathbf{y}$. This is then called a conditional GAN as the generated data distribution is conditioned on the distribution of $\mathbf{q}$ (Mirza and Osindero, 2014).

The optimal solution for simultaneously maximizing $\mathcal{C}_{\mathbf{G}}$ and $\mathcal{C}_{\mathbf{D}}$ exists and is exactly equal to the generated data distribution being identical to the training data distribution (Goodfellow et al., 2014). Convergence to this optimal solution is not guaranteed and many optimization techniques and architecture adaptations exist to help convergence. For more stable training, several gradient ascent updates can be performed for $\mathbf{D}$ for each update of $\mathbf{G}$. If the generated data are to be used for a downstream DL task, additional *task loss* functions may be included. These evaluate the performance of a pretrained NN on the generated data. Another application of GANs is generative domain adaptation, in which the generator is trained to perform a specific mapping between two distributions. A famous example of this is turning images of horses into images of zebras with the same background (Isola et al., 2017). This can be

achieved with a cycle-consistent GAN (CycleGAN), which uses two GANs, which are trained simultaneously to each perform one direction $\mathbf{G}_1 : \mathbb{P}_\mathcal{S} \rightarrow \mathbb{P}_\mathcal{T}$ or $\mathbf{G}_2 : P_\mathcal{T} \rightarrow P_\mathcal{S}$ of the transfer between the source and target domains with data distributions $P_\mathcal{S}$ and $P_\mathcal{T}$, respectively (Zhu et al., 2017). The method is termed "cycle-consistent" because both generators are trained to transfer output from the other generator back to the original domain. The framework can also accommodate additional losses to ensure faster convergence and more consistent results. Another way to improve training is to replace Eq. 2.44 and Eq. 2.43 with loss functions that provide a more informative gradient. A popular choice here is the Wasserstein- or earth-movers-distance $\mathcal{W}(P_i, P_j)$ (Eq. 2.45). The Wasserstein distance quantifies the cost of the optimal transport plan between two distributions. Equation 2.45 denotes the computationally tractable form of computing the Wasserstein distance $\mathcal{W}$.

$$\mathcal{W}(P_i, P_j) \propto \max_{\mathbf{w}} \mathbb{E}_{\mathbf{x}\sim P_i}[\mathbf{F}_\mathbf{w}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim P_j}[\mathbf{F}_\mathbf{w}(\mathbf{y})]\,. \tag{2.45}$$

$\{\mathbf{F}_\mathbf{w}\}$ is a family of Lipschitz-continuous functions parametrized by $\mathbf{w}$. The distributions P_i and P_j denote a fixed and parametrized distribution, respectively. For a Wasserstein-GAN (WGAN) using Eq. 2.45, trained to adapt $P_\mathcal{S} \rightarrow P_\mathcal{T}$ Eq. 2.45 becomes

$$\mathcal{W}(P_\mathcal{S}, P_\theta) \propto \max_{\mathbf{w}} \mathbb{E}_{\mathbf{x}\sim P_\mathcal{S}}[\mathbf{F}_\mathbf{w}(\mathbf{x})] - \mathbb{E}_{\mathbf{y}\sim P_\mathcal{T}}[\mathbf{F}_\mathbf{w}(\mathbf{G}_\theta(\mathbf{y}))]\,, \tag{2.46}$$

$$\Leftrightarrow \mathcal{W}(P_\mathcal{S}, P_\theta) \propto \max_{\mathbf{w}} \mathbb{E}_{\mathbf{x}\sim P_\mathcal{S}}[\mathbf{F}_\mathbf{w}(\mathbf{x})] - \mathbb{E}_{\mathbf{q}\sim P_\theta}[\mathbf{F}_\mathbf{w}(\mathbf{q})]\,, \tag{2.47}$$

such that eventually $P_\theta \approx P_\mathcal{S}$.

2.4. Cloud Classification

2.4.1. Observational Products

The importance of surface-based observations is evident in the nomenclature of clouds, which inherently emphasizes the cloud base. Even though modern surface-based observation techniques, e.g. those using active sensors (Stevens et al., 2016) help understand local cloud dynamics, they are not best suited for providing global climatologies of clouds as they cannot provide uniform global coverage (e.g UniData2003).

The requirement and possibility to analyze satellite observations of clouds in terms of classes were already noted by Schiffer and Rossow (1983) in the outline of the ISCCP. At the end of the 20^{th} century, most of the satellite-based cloud products only provided cloud cover, which alone is of limited value (Norris, 1998; Rossow et al., 1993). The need for a more detailed dataset resulted in the creation of the D-Series dataset of ISCCP, which included a classification

of nine morphological cloud types (Rossow and Schiffer, 1999). In ISCCP-D the clouds are classified via thresholds in *cod*/cloud top pressure (*ptop*)-space for daytime measurements at 3 h temporal and 280 km/2.5° horizontal resolutions. Furthermore, low and middle clouds are further distinguished by thermodynamic phase such that per-sample averages of *cod*, *ptop* and *cwp* are attributable to 15 classes. This approach is extended in the ISCCP H-Series with increased horizontal resolution (1° equal area), and all 18 ice/liquid classes (Young et al., 2018). The ISCCP data can serve as input to ML methods to produce a classification of large-scale cloud structures, termed *regimes* (Gordon et al., 2005; Jakob and Tselioudis, 2003; Williams and Webb, 2008). These studies use k-means clustering (Lloyd, 1982) of cloud properties from satellite products and GCM satellite simulator output for unsupervised classification of clouds. The clusters are then labeled according to how their mean physical properties resemble morphological cloud types. Such a classification offers insights into how clouds are represented in individual GCMs, different from what is possible with climatologies of physical variables, such that processes related to specific cloud types can be analyzed more directly (see Section **2.4.2**). However, only passive sensors are used and sensor limitations can affect comparisons to GCMs, even if instrument simulators are used (Pincus et al., 2012; Swales et al., 2018). More accurate results are expected for application of these clustering methods to the ISCCP-H series and GCMs as performed by Tselioudis et al. (2021). Figure **2.4** shows the *cod* and *ptop* distributions of the clusters that are found by Tselioudis et al. (2021), which are each assigned labels corresponding to the cloud types that are likely to be prevalent in each of these clusters. Nevertheless, the classes obtained through clustering methods are not necessarily connected to established WMO classes, even if more advanced clustering methods are employed (e.g. Denby, 2020). Also, the aforementioned methods are based on passive sensor data and come with the limitations intrinsic to these sensors. Using radar and lidar data has helped provide more objective and accurate classifications (Huang et al., 2015), but these sensors lack the spatial coverage required for comprehensive studies. Improving comparability to established classes, accuracy and coverage might be achieved with (supervised) NN approaches. Moving from general image classification to NN-based cloud classification is a fairly natural idea and has already been attempted decades ago (e.g. Lee et al., 1990). Large-scale application of this idea is however only slowly making its way into the state of the art as computing power as well as more efficient and powerful architectures are becoming available. These methods for cloud classification are not necessarily supervised, as plenty of architectures and training schemes not requiring labeled data are available (e.g. Grill et al., 2020; Kingma and Welling, 2013; Kramer, 1991). To emphasize the distinction between class-average properties, unsupervised methods based on NNs have been developed, whereas in the clustering methods, the resulting classes are compared to established cloud types (Denby, 2020; Kurihana et al., 2021). Since the WMO cloud classes are defined with surface observations in mind, they could be considered unsuitable in the context of satellite observations, to the extent that new classes have been manually extracted and used to train NNs (Marais et al., 2020; Rasp et al., 2020; Stevens et al., 2019). In contrast, supervised classification assumes that the assigned classes are fit for purpose. Also, a set of labeled data is required for supervision, which is difficult

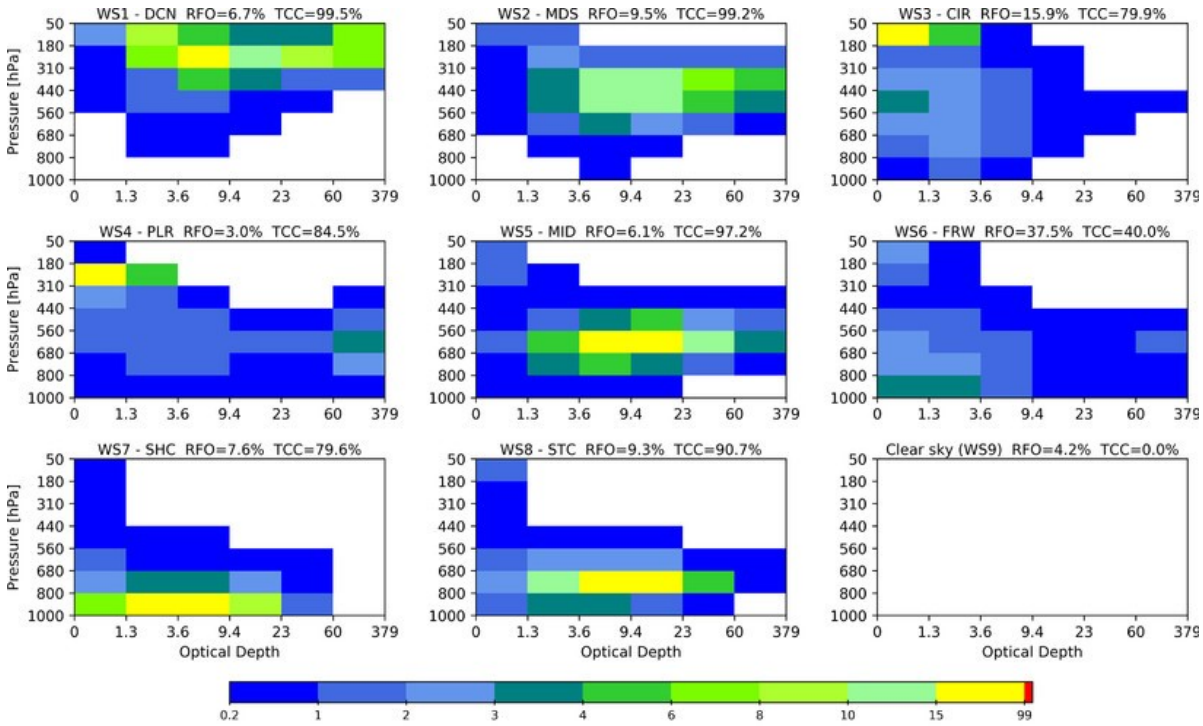


Figure 2.4.: Distribution of *cod* and *ptop* in the nine clusters found from ISCCP-H data by Tselioudis et al. (2021). The clusters are each assigned names that correspond with the prevailing weather states, as diagnosed from their property distributions and areas of occurrence: DCN - deep convective and anvil, MDS - midlatitude storms, CIR - thin high cirrus, PLR - polar clouds, MID - middle-top clouds, FRW - fair weather, SHC - shallow cumulus, STC - stratocumulus, CLR -clear sky. Adapted from Tselioudis et al. (2021),©American Meteorological Society. Used with Permission.

and expensive to create for clouds, and therefore not readily available on a global scale. With WMO-like cloud class labels from active spaceborne sensors (see Chapter **3**), supervised NN approaches can be developed (Gorooh et al., 2020; Zantedeschi et al., 2019). Such methods are however still affected by the sparsity of the active sensor data and have to rely on passive sensors to achieve global coverage at sufficient temporal resolution. This lack of available labeled satellite data appears to be the main reason why DL classification using observations labeled by surface observers has received more attention as of now (Liu et al., 2020; Sedlar et al., 2021; Zhang et al., 2018).

2.4.2. Application to Climate Model Evaluation

As mentioned in the previous section, some cloud classification methods have already been applied to evaluate GCMs. Many of these are K-means clustering (Lloyd, 1982) schemes using *cod* and *ptop* as inputs, sometimes additionally combined with total cloud fraction (*clt*). For GCM evaluation, typically the clusters are either predefined using observational data (Tse-

lioudis et al., 2021; Tsushima et al., 2012; Williams and Webb, 2008) or GCM clustering is performed separately (Chen and Genio, 2008; Williams et al., 2005). When clustering is performed separately, there is no guarantee that the sets of clusters found in each dataset are equivalent. Also, many ISCCP clustering methods use the full range of 42 parameters suggested by ISCCP (Rossow and Schiffer, 1999), which leads to noticeable dependence of the final clusters on the initial configuration of the cluster (Gordon et al., 2005), further increasing the dissimilarity between observational and GCM clusters. However, assigning GCM output to existing observational clusters requires comparable data, which is usually produced using satellite simulators. Recently, this was done for clusters obtained from ISCCP or MODIS with the corresponding instrument simulators by Tselioudis et al. (2021) and Cho et al. (2021), respectively. Comparing the ISCCP and MODIS clusters, Oreopoulos et al. (2014) find some common regimes but also many differences, highlighting both the deviations that can be introduced by different retrieval methods as well as the lack of objectivity inherent to unsupervised clustering.
Cloud regimes are assigned to observational and GCM data in an effort to inform future field campaigns and modeling efforts (Jakob and Tselioudis, 2003) and to attribute deviations between models and observations to either an incorrect representation of the frequency of a regime or of the regime's physical impact. This idea is implemented by Williams and Tselioudis (2007) and Williams et al. (2005) to determine the CRE response of the cloud regimes to global warming and partitioning this response into the change of the distribution of the cluster properties, the change in the relative frequency of occurrence (RFO) of each cluster, and the co-variation of both. This way of partitioning the results had previously been applied to regime analysis by Bony et al. (2004), who used sea surface temperature (SST) and 500 hPa vertical velocity as proxies for the thermodynamic and dynamical regime. Of the six GCMs analyzed in Williams and Tselioudis (2007), four show a positive net cloud feedback of the stratocumulus regime to warming, one displayed a negative response and one did not produce such a regime consistent with observations. Williams and Webb (2008) extend the approach of Williams and Tselioudis (2007) to additional GCMs, and find varying RFOs and mean cluster properties between the GCMs, leading to large differences in the average geographical distributions of some regimes. Such differences can be a good starting point for process-based GCM evaluation. Chen and Genio (2008) use ISCCP cluster analysis to pinpoint improvement potentials in a GCM, such as requirements for higher entrainment rates to more accurately simulate mid-level clouds. They also note that in general, a potential source of errors in this approach is the uncertainty of the passive sensors used for the ISCCP dataset in determining the cloud top pressure.

The ISCCP regimes can also be used to analyze the seasonal variation of the CRE associated with each of the regimes. Schuddeboom et al. (2018) use the self-organizing maps clustering technique (Kohonen, 1998, 2013) to identify cloud representation errors in a single atmosphere GCM. They find that the CRE deviation relative to observations is largely based on the difference in RFO of the clusters, not in the simulated CRE of a given cluster. They also conclude that the composition of the clusters can significantly vary over the globe, for example,

a single cluster associated with Sc clouds can contain a high content of Ns at higher latitudes. In an attempt to determine which GCMs provide the most accurate estimate of equilibrium climate sensitivity (ECS), Jin et al. (2016) use several metrics and ISCCP cloud regimes (so-called *weather states* (Tselioudis et al., 2013)) to rank Climate Model Intercomparison Project (CMIP) Phase 5 GCMs in terms of cloud-representation performance. However, no relationship between a GCM's ECS and its rank is apparent, and the two highest ranking GCMs differ strongly in simulated ECS. The CRE response of cloud regimes to warming is further investigated by Zelinka et al. (2022b), who use a detailed decomposition of the net cloud feedback to gain further insight into processes responsible for cloud feedbacks in ten selected GCMs. Similar to earlier studies, they find that most of the global SW cloud feedback is attributable to a change in regime RFO under warming conditions.

The clusters that are found with the above methods are typically assigned labels that correspond to the dominant cloud types present in that cluster in an effort to make the results more interpretable. However, the same studies also recognize that these labels do not apply to every cloud in a regime, as these are usually mixtures of several types. Cloud regimes produced by unsupervised clustering are therefore difficult to interpret objectively. Furthermore, several of the studies mentioned above state that the exact definition of the clusters has little effect on the interpretation of the results (Gibson et al., 2017; Schuddeboom et al., 2018; Zelinka et al., 2022b), which is good in the sense that it allows for some error tolerance but also raises the question of how meaningful the clusters truly are. As mentioned in Section **2.4.1**, supervised approaches offer higher interpretability but supervision is difficult in the context of cloud classes and extending a supervised approach to GCMs comes with additional problems like low resolution and limited comparability of simulated cloud properties to observations. Kuma et al. (2023) used surface-based cloud type observations and satellite-observed SW and LW TOA radiation fields to predict probabilities of occurrence of four predefined cloud types (Middle, High, Cumuliform, Stratiform) from $2.5° \times 2.5°$ cells using NNs. They find that CMIP6 GCMs with higher ECS have a smaller bias overall but tend to overestimate low marine cumuliform clouds and underestimate stratiform clouds in the same regions. They also show that the accuracy with which cloud type distributions are represented in the GCMs could be used to infer the validity of simulated feedback parameters like ECS and cloud feedback for a given GCM, similar to the study mentioned above (Jin et al., 2016). There is an indication that GCMs which represent the cloud types more accurately with respect to observations have a stronger warming response, but due to correlations between related GCMs, which share code to various degrees, the confidence in this is small.
Overall, it is apparent that *cod-ptop* histograms like those provided by ISCCP have found a lot of use and some success in pinpointing potential GCM deficiencies, whereas more clearly defined classes like those also included in the ISCCP data have rarely been used for GCM evaluation. This is most likely due to the lower resolution of GCMs, which makes it very difficult to identify cloud types solely based on grid-box average *cod*, *ptop* and thermodynamic phase.

Furthermore, since the regimes are defined on the grid scale, other properties like CRE can be easily attributed to a regime, while this is more difficult for subgrid-scale class distributions.

3. Instruments and Data

This chapter contains strongly modified text and slightly adapted tables from the integrated publication Kaps et al. (2023a).

3.1. The A-Train and Cumulo

Since 2004, the Afternoon Constellation (A-Train) made it possible to obtain near-simultaneous measurements of the same location from a variety of spaceborne instruments. The A-Train contained a maximum six satellites (today three), of which three feed into the CUMULO dataset, which form the basis for large parts of the work presented in this thesis (Zantedeschi et al., 2019). CUMULO combines data from instruments aboard the Aqua, CloudSat (CS) and CloudAerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) satellites which flew in the A-Train with an interval of less than 3 min from Aqua to CloudSat.

There are three years of CUMULO available of which the year 2008 was downloaded from the GitHub page [1] of CUMULO. In the CUMULO dataset, the MODIS and 2B-CLDCLASS-LIDAR (CC-L) (Section **3.1.2**) data are aligned, such that for near-simultaneous observations one of the eight cloud type labels from CC-L are available in addition to the MODIS retrievals, resulting in a narrow track of labels across the MODIS images. The columns where labels are available are each assigned only a single cloud type, defined as the label that occurs most often in the column, such that the complete CUMULO dataset is comprised of column-integrated values. Even though CS and CALIPSO provide nighttime measurements, CUMULO only contains daytime data as MODIS can not reliably retrieve all physical quantities at night. The collocation provided in CUMULO of the passive and active sensors aboard the three satellites enables the training of ML models to infer active sensor information from the passive sensor data.

[1] `https://github.com/FrontierDevelopmentLab/CUMULO`, last accessed 27th Nov. 2023

3.1.1. Moderate Resolution Imaging Spectroradiometer

MODIS is a passive instrument operating aboard NASA's Terra and Aqua satellites, which are both in sunsynchronous orbit with an orbital period of about 99 minutes. MODIS measures reflected sunlight as well as planetary thermal radiation in 36 channels from the SW, covering 0.4 µm to 3.0 µm, to the infrared, covering 3 µm to 14.5 µm. Some of these channels are so-called heritage channels for consistency with older sensors like Advanced Very High Resolution Radiometer (AVHRR)(Section **3.2.1**). Depending on the channel, the resolution is 250 m, 500 m and 1000 m. This work made use mainly of the Cloud Product MYD06 of the Aqua satellite, with some ancillary information from the cloud mask and geolocation products. MYD06 is provided at 1 km nadir resolution with granules of 1354 across-track and 2030 along-track pixels. It contains retrievals of nine cloud-related quantities (Table **3.3**). The following provides an overview of the relevant retrieval methods employed in collection 6 of MYD06. Cloud properties are only retrieved if the MODIS cloud mask (MYD35) returns that the sample is (probably) cloudy.

MODIS retrievals of the cloud top thermodynamic phase (*cph*) uses two different approaches (Marchant et al., 2016). In the first, three channel pairs in the infrared are each compared to infer the phase through the relative strength of the absorption. The absorption for ice scales more strongly with wavelength than for water, as discussed in Section **2.1.1** (Baum et al., 2012). The results of this first algorithm are part of a weighted voting mechanism, which combines cloud top height and temperature checks with retrievals of *cer* in the shortwave infrared spectrum (SWIR). This last part accounts for the ambiguity of *cod* and *cer* (comp. Eq. 2.27). The result of this voting mechanism can be ice, liquid or undetermined. The *cph* retrieval informs the retrieval of all other cloud top properties. To determine *cod* and *cer*, the MYD06 algorithm compares the radiance of reflected solar radiation to values stored in LUTs, where pre-computed reflectance values for the waveband are stored. The LUTs only store results of radiate transfer calculations with assumed multiple scattering, as the single scattering component can be computed on the fly for each pixel (Platnick et al., 2017). To simultaneously retrieve *cod* and *cer*, the reflectance calculations are performed for smaller wavelength channels (0.66 µm, 0.86 µm and 1.12 µm) that are more sensitive to *cod* and channels at 1.6 µm, 2.1 µm and 3.7 µm, where the reflectance is more sensitive to *cer* (Nakajima and King, 1990; Platnick et al., 2017). As indicated by Eq. 2.27, *cwp* is obtained via the product of *cer* and *cwp*. The retrievals for *cph*, *cwp*, *cod* and *cer* depend in one way or more on reflected solar radiation and are thus only available for daytime measurements. In contrast, *ptop* and cloud effective emissivity (*cee*) are retrieved using the CO_2-slicing technique (see Section **2.2**), which relies and thermal emission and therefore works without solar reflection.

Table 3.1.: The three satellite datasets used at different stages throughout this work. Adapted from Kaps et al. (2023a).

Name	Product/ Version	Purpose	Resolution	Reference
CC-L	2B-CLDCLASS-LIDAR / P1-R05	cloud type	1.4 km × 1.8 km	Wang (2019a), Sassen et al. (2008)
MODIS	MYD06 / 6.1	physical variables	1 km	Platnick et al. (2003)
ESA-CCI	ESA Cloud_cci L3U / AVHRR-PMv3	validation	0.05°	Stengel et al. (2020)

Table 3.2.: WMO-like cloud types from the CUMULO dataset. From Kaps et al. (2023a).

Abbv.	Cloud type
Ci	Cirrus/Cirrostratus
As	Altostratus
Ac	Altocumulus
St	Stratus
Sc	Stratocumulus
Cu	Cumulus
Ns	Nimbostratus
Dc	Deep Convection

3.1.2. CPR and CALIOP

In addition to the passive sensor data from MODIS, CUMULO contains measurements from CALIOP aboard CALIPSO and CPR aboard CS, which were used together to create the CC-L dataset, which expands on the 2B-CLDCLASS dataset (Sassen et al., 2008; Wang, 2019b) by incorporating the lidar measurements. The Cloud Profiling Radar (CPR) is a 94.05 GHz (3.2 mm) radar aboard CloudSat (CS) with a range resolution of 500 m and a horizontal resolution of 1.4 km × 1.8 km (Sassen et al., 2008). CPR is very high-powered and able to detect most clouds with the exception of thin cirrus or clouds with small droplets. CALIOP, on the other hand is precisely designed to detect small particles with a two-wavelength laser in the visible and infrared (532 nm and 1064 nm). Additionally, CALIOP can distinguish between ice and water particles via detecting linear depolarization in the 532 nm channel. The resolution achieved by CALIOP depends on the altitude of the measurement. Below 8.2 km the vertical resolution is 30 m, while it is 60 m in the upper troposphere and TTL. The horizontal footprint is 333 m × 333 m. As discussed in Section **2.2.1**, cloud properties can be inferred from the CPR reflectivity factor Z (see Eq. 2.30) and the attenuated backscatter coefficient from CALIOP. CC-L also uses ancillary temperature profiles from European Center for Medium

Table 3.3.: All physical variables obtained from the MODIS Cloud Product and contained in CUMULO. All can be used for the ML methods, but some are effectively redundant (cloud top height (*htop*),*ptop*, cloud top temperature (*ttop*)). Adapted from Kaps et al. (2023a).

Abbv.	Physical variable
cwp	Cloud water path
cod	Cloud optical thickness
ptop	Cloud top pressure
htop	Cloud top height
ttop	Cloud top temperature
tsurf	Surface temperature
cer	Effective cloud particle radius
ceff	Effective Emissivity
phase	Cloud thermodynamical phase

Range Weather Forecasts (ECMWF) data and cloud mask information from MODIS. These measurements provide the properties

- cloud height and phase
- temperature profile
- radar reflectivity factor Z
- cloud thickness and horizontal extent
- cloud cover
- precipitation

The combination of radar and lidar in CC-L enables retrievals that would be impossible or inaccurate with a single sensor. The different response of ice and water particles to the respective radar and lidar wavelengths owing to their size makes phase determination throughout multiple vertical layers much more effective. Each column sampled in CC-L contains up to ten vertical layers of clouds, with layers separated by at least 500 m. The layers do not have predefined heights and are populated from top to bottom. The derived cloud properties are used to assign a single cloud-type label to the column element (Table **3.2**). The classifier is a combination of rules logic (decision tree) and fuzzy logic, with rules manually set from expert knowledge. Because CPR can penetrate most clouds and CALIOP is sensitive to even the thinnest cloud layers, measurements of both the height of top and base of a cloud are possible, making the resulting classification more refined than classical methods using passive sensors (e.g. in Young et al., 2018).

Variable name	Abbreviation	Unit	Comment
Cloud mask and cloud fraction	CMA and CFC	1%	Binary cloud occurrence classification (ANNmask) and fraction of cloudy pixels
Cloud phase and liquid cloud fraction	CPH and LCF	1%	Binary cloud phase classification (ANNphase) and fraction of liquid clouds
Cloud-top pressure	*ptop*	hPa	OE retrieval result of cloud-top pressure
Cloud-top height	*htop*	km	Derived from CTP and atmospheric profile
Cloud-top temperature	*ttop*	K	Derived from CTP and atmospheric profile
Cloud effective radius	*cer*	µm	OE retrieval result of cloud effective radius
Cloud optical thickness	*cod*	1	OE retrieval result of cloud optical thickness
Surface temperature	surface temperature (*tsurf*)	K	OE retrieval result of surface temperature
Cloud water path	*cwp*	$g\,m^{-2}$	Derived from CER and COT

Table 3.4.: Cloud properties in ESA-CCI. Additional variables are available but were not used here. OE refers to the optimal estimation inversion algorithm used in Community Cloud retrieval for Climate (CC4CL). Adapted from (Stengel et al., 2020) according to creative commons license CC BY 4.0[2].

The small footprint of CPR and CALIOP makes it difficult to distinguish if a cloud structure is consistent on a scale of $\mathcal{O}(10-100)$ km (Wang, 2019a). This affects the classification of the St and Sc cloud types, which have similar properties locally but vary in their horizontal homogeneity (comp. Section **2.1.2**), which is however hard to detect with this small footprint (Wang, 2019a).

3.2. ESA Cloud_cci

The Climate Change Initiative of the European Space Agency (ESA) has published six datasets of cloud properties retrieved from passive spaceborne sensors (Stengel et al., 2017). The dataset with the longest temporal coverage (1982-2016) is based on the AVHRR sensor (Stengel et al., 2019). It is called the AVHRR-PM dataset since it is compiled from measurements made by different AVHRR instruments (versions 2 and 3) on consecutively launched satellites that all orbited in sunsynchronous orbit with an afternoon equatorial crossing time. Since this is the only ESA Cloud_cci dataset used for this thesis, the AVHRR-PMv3 dataset will be referred to as ESA-CCI. Cloud-related variables (Table **3.4**), similar to those MODIS provides in the Cloud Product (Section **3.1.1**) are extracted from the dataset. ESA-CCI is provided at two temporal resolutions, daily in the L3U version of the dataset and as monthly means in L3C. The daily data contains the ascending and descending node of the orbit separately at a horizontal resolution of 0.05°, the monthly data is averaged over the nodes and saved at 0.5°.

[2] `https://creativecommons.org/licenses/by/4.0/`

3.2.1. Instrument and Retrievals

The AVHRR instruments used for ESA-CCI detects reflected solar and thermal infrared radiation in five channels between 0.6 µm to 12 µm. The native resolution of AVHRR is 1.1 km × 1.1 km, but ESA-CCI uses and averaged product with a 1.1 km × 4.4 km footprint size. The AVHRR-carrying satellites are all subject to orbital drift, with the earlier satellites changing equatorial crossing time by more than 2 h by the end of their lifetime. Since change in observed local time induces a shift in the measured radiances and the switching of the satellite usually causes significant jumps in the data. The retrievals do not account for these effects, making the data unsuitable for trend analysis without correction.

All but one of the ESA Cloud_cci datasets, including this AVHRR dataset retrievals are produced using the CC4CL algorithm, which is meant to provide consistency between the data from different sensors (Stengel et al., 2017; Sus et al., 2018) The retrievals are performed at resolution of the sensor and then averaged to the output grid of the L3U/L3C data. In CC4CL, the cloud mask and cloud top phase are provided by a small NNs trained on the respective *cod* and *cph* values from CALIOP. Further cloud top properties *cod*, *cer* and *ttop* are retrieved using the Optimal Retrieval of Aerosol and Cloud (ORAC) algorithm, which uses inversion of the measured reflectance values against those calculated in a forward model. The forward model takes into account only multiple scattering as single scattering computations are simpler and can be performed during the retrieval (McGarragh et al., 2018). The retrieval of *cod* and *cer* even during the night was enabled by using the relative reflectances in the 3.7 µm and the 10.8 µm/12 µm channels (Stengel et al., 2020). Using these retrievals, *lwp* and *iwp* are computed using Stephens (1978). However, this method for nighttime retrievals is considered experimental and might be less accurate. Because AVHRR is missing the LW channels of the MODIS sensor, overlapping clouds are difficult to measure, resulting in inconsistent measurements in comparison to active sensors (Sus et al., 2018).

3.3. ICON Data

In Chapters **5** and **6**, GCM output from an ICON-A (Giorgetta et al., 2018) simulation with specifically expanded output is used. The general information on the physical processes simulated in ICON-A is given Fig. **3.1**, as given in Giorgetta et al. (2018). Specifically, an Atmospheric Model Intercomparison Project (AMIP) (Gates, 1992) simulation with prescribed sea-surface temperatures was run with a 29-year transient and instantaneous 3-hourly output saved for 2008 and 2009. The output is regridded from its native R2B5 icosahedral grid (∼80 km) to a 1° regular latitude-longitude grid using the Climate Data Operators (`cdo`,

[3] `https://creativecommons.org/licenses/by-nd/4.0/deed.en`

Table 2
Key Physical Processes in ICON-A

Process	Base reference	Characteristics	Atm. tendencies
Cloud cover	Sundqvist et al. (1989)	Cloud cover diagnosed from relative humidity	–
RadiationSW and LW	"PSrad,"Pincus and Stevens (2013)	Correlated-k scheme, two streams, originating from RRTMG, scattering in SW, McICA for cloud effects	$\partial T/\partial t$
Vertical diffusion	Mauritsen et al. (2007)	Total turbulent energy scheme, implicitly coupled to the land surface model	$\partial T/\partial t$, $\partial q_{v,c,i}/\partial t$, $\partial u/\partial t$, $\partial v/\partial t$
Land surface	"JSBACH4-lite," Raddatz et al. (2007)	Five layers for water and heat storage, one surface type per cell	-
Cumulus convection	Nordeng (1994)	Mass flux scheme with shallow, deep or midlevel convection. Moisture convergence closure for shallow and mid level convection, CAPE closure for deep convection	$\partial T/\partial t$, $\partial q_{v,c,i}/\partial t$, $\partial u/\partial t$, $\partial v/\partial t$
Stratiform clouds	Lohmann and Roeckner (1996),	Prognostic cloud water and ice; diagnosed liquid or snow precipitation	$\partial T/\partial t$, $\partial q_{v,c,i}/\partial t$
Orographic drag	Lott (1999)	Orographic blocking and orographic gravity wave drag, subgrid-scale orographic parameters based on GLOBE elevation data	$\partial T/\partial t$, $\partial u/\partial t$, $\partial v/\partial t$
Nonorographic gravity wave drag	Hines (1997)	Spectra in eight azimuths, constant, globally uniform and isotropic wave sources at ~680 hPa	$\partial u/\partial t$, $\partial v/\partial t$

Note. The last column shows the tendencies in atmospheric variables resulting from each parameterization. The $\partial q_{v,c,i}/\partial t$ stands for tendencies of water vapor, cloud liquid, and cloud ice mass mixing ratios, respectively.

Figure 3.1.: Physical processes and their parametrizations in Icosahedral Nonhydrostatic Atmosphere model (ICON-A). Reproduced without changes from Giorgetta et al. (2018) in line with creative commons license CC BY-NC-ND 4.0[3].

Schulzweida (2023)) package's nearest neighbor averaging. The variables effective radius of cloud water droplets (*cerl*), *ceri* and *cod* as well as *ptop* for general clouds are not contained in the ICON-A standard output and are specifically added. The *cerl*, *ceri* and *cod* are extracted from the radiation parametrization "PSrad" of Icosahedral Nonhydrostatic model (ICON) (Pincus and Stevens, 2013)[4]. These cloud optical properties in PSrad are computed for several wavebands and saved for the bands centered on 2.05 µm and 2.32 µm. These wavebands are chosen as they are used in the *cod* and *cer* retrievals of ESA-CCI. Since the *cod* values for both bands show no considerable difference, the 2.32 µm band is used for *cod* in the following. The other three missing variables require the definition of a cloud top, which in the standard ICON-A output is only provided for convective clouds. The cloud top is determined here by finding the layer at which the total *cod* viewed from TOA exceeds $\tau = 0.2$, which has the benefit of being comparable with satellite observations. The pressure at this level then serves as *ptop* and the particle radii *cerl* and *ceri* are also extracted from this level.

[4] Adding the variables and running the simulation was performed by Rémi Kazeroni, according to jointly agreed specifications.

4. Machine-Learned Cloud Classes From Satellite Data for Process-Oriented Climate Model Evaluation

4.1. Overview

global climate models (GCMs) are important tools not only to improve our understanding of present-day climate but also to project climate change under different plausible future scenarios. The simulation of clouds and their interactions with the climate system, however, remains a major challenge for GCMs (Vignesh et al., 2020). The representation of clouds in these models has been identified as one of the primary sources of inter-model spread (Dufresne and Bony, 2008; Zelinka et al., 2020). Cloud processes in GCMs are therefore evaluated against observations of clouds to improve their simulations (Bony, 2015; Schneider et al., 2017; Williams and Webb, 2008). Frequently, observations used to assess model performance are obtained from long-term satellite products providing near-global coverage, which have proven to be well suited for the evaluation of GCMs (e.g. Kawai et al., 2019; Lauer et al., 2017). This conventional approach is, however, constrained in part due to limitations and uncertainties of observational products themselves (Jakob and Tselioudis, 2003), such as biases or varying spatial and temporal coverage.

In this chapter, a new approach to presenting cloud-related data is presented, designed to facilitate process-oriented evaluation of clouds in GCMs and to address some of the apparent limitations of using conventional observational data. We use a priori knowledge about the characteristics of different cloud classes based on the cloud type classification of the WMO. By exploiting this a priori knowledge, cloud processes can be highlighted in further evaluation. Our approach extends the recent development of ML based cloud classification methods for satellite data (Denby, 2020; Kurihana et al., 2021; Marais et al., 2020; Rasp et al., 2020; Zantedeschi et al., 2019; Zhang et al., 2019) to GCM-comparable resolutions.

To our knowledge, no high-resolution ($\mathcal{O}$(1 km)) cloud-class-labeled satellite data have been used for analysis and evaluation of GCMs, so far. We argue that labeled datasets allow for a

more detailed and more direct interpretation of cloud classes in the respective satellite data in contrast to comparatively coarse classifications as used for example by the ISCCP (Rossow and Schiffer, 1999; Young et al., 2018).
Our method aims at establishing this connection between observations and models without the requirement to assign cloud classes a posteriori. We instead compute the relative amount of WMO-like cloud classes in coarse grid cells. Statistical analysis can be conducted on these distributions in the same manner as for the traditionally used physical variables but in the phase space of cloud classes.

The contents of this chapter are largely already published in a peer-reviewed first-author paper and its supplements (Kaps et al., 2023a). All scientific work contained in this chapter was performed by the author of this thesis, including tuning of the ML models, analysis of the results and generation of figures. The co-authors of Kaps et al. (2023a) contributed to outlining the concept and assisted in writing the published text. Some parts of the code are modified versions of code previously published by others and this is clearly stated where applicable. All novel code can be found on `DOI:10.5281/zenodo.7248773`.

The chapter is outlined as follows: In Section **4.2**, we describe the satellite products used and introduce the two ML methods applied. In Section **4.3**, we use our results to establish (1) that an NN can be used to accurately assign *physically robust* cloud class labels to satellite data, (2) that this labeled satellite data provides a sufficient basis to train a regression model relating physical variable retrievals from satellites to cloud class distributions in coarse grid cells and (3) that the application to a coarse-grained version of the alternative ESA Cloud_cci satellite product Stengel et al. (2017) is possible, showing the potential of the framework. Our findings are summarized in Section **4.4** and discussed in the context of related work and the scientific questions of Chapter **1** in Section **4.5**.

4.2. Methods

4.2.1. Overview

Our goal is to evaluate clouds in GCMs using observational data labeled with cloud types. For this, we need to

1. obtain or create a cloud-labeled dataset

2. enable a comparison to GCM output

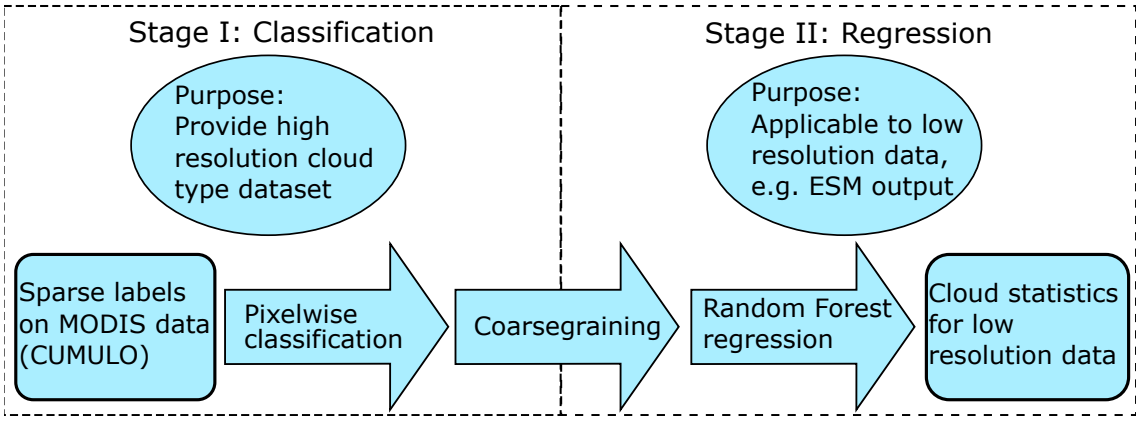


Figure 4.1.: Two stages of ML - a classifier and a regression model - are required to obtain cloud-type predictions on datasets with low horizontal resolution. From Kaps et al. (2023a).

As a starting point, we use the partially labeled CUMULO dataset Zantedeschi et al. (2019) (hereafter Z19), which is then fully labeled by using an NN classifier. This procedure proposed by Z19 could result in a long-term high-resolution and full coverage cloud-type dataset. The concept outlined here takes a further step by adapting the resolution to be comparable to current GCMs. We train a regression model to predict the relative cloud-type amounts of larger areas (grid cells). This concept is outlined in Fig. **4.1**.
In Section **4.2.2** (Stage I in Fig. **4.1**), we outline how the CUMULO dataset was created using MODIS data and cloud type labels from CS. The training of the regression model on a coarse-grained version of CUMULO is outlined in Section **4.2.3** and Stage II in Fig. **4.1**. Section **4.2.4** explains the steps applied to validate the regression model's performance (see Fig. **4.5**) on coarse-grained data from ESA-CCI, which are independent of the training data.
The workflow of this framework is shown in Fig. **4.2**, which illustrates how the different datasets and the two ML models contribute to providing cloud-type distributions for low-resolution data.

4.2.2. Pixel-wise Classification

We create a fully labeled cloud-type dataset by applying a pixel-wise classifier network to the sparsely labeled CUMULO dataset Z19. The result is a high-resolution, high-coverage, cloud-labeled dataset (see Fig. **4.3**). The classification scheme applied here is largely based on the code published for the classification algorithm used in Z19, available from GitHub[1]. Since the CS and CALIPSO swaths are quite narrow, most of the pixels in the resulting CUMULO dataset are not assigned a label, which is why an NN is trained using the labeled part of the dataset to predict cloud-type labels for the unlabeled pixels.
The NN used in Z19 and here to classify clouds in the CUMULO dataset is a semisupervised

[1] `https://github.com/FrontierDevelopmentLab/CUMULO`, last accessed 27th Nov. 2023

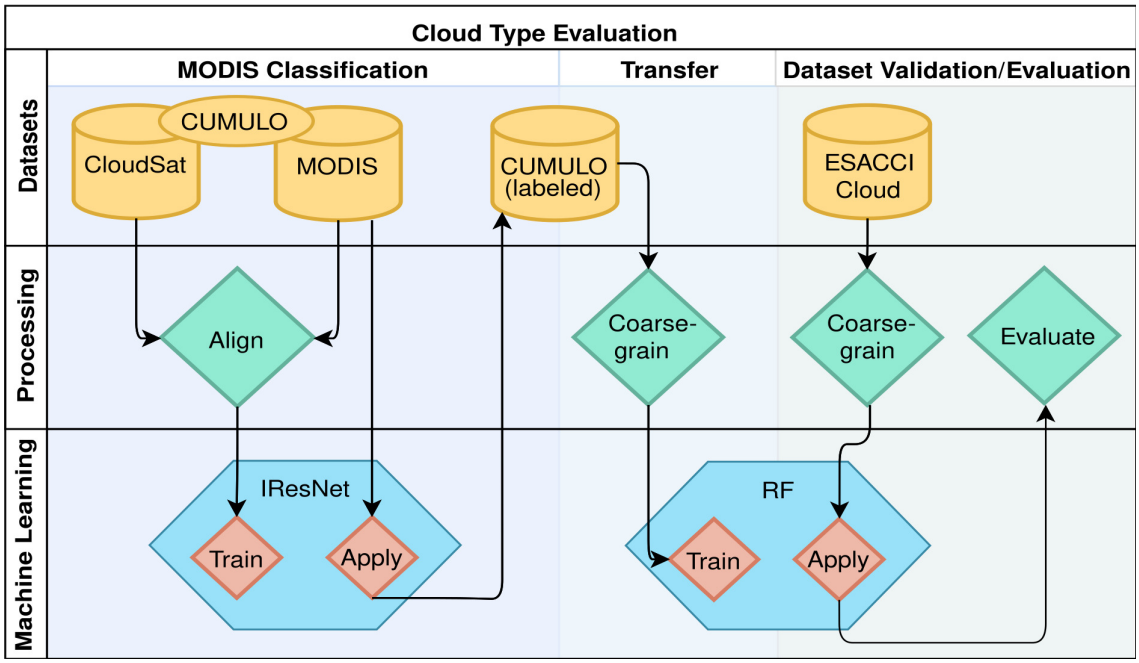


Figure 4.2.: Workflow schematic: (1) The Invertible Residual Network (IResNet) is trained on the CUMULO dataset and then applied on the unlabeled full-swath MODIS yielding the fully labeled CUMULO dataset. (2) An RFRM regression model is trained on a coarse-grained version of this data to provide cloud class distributions. (3) The RFRM is applied to unseen data, allowing validation of the method's performance or evaluation of the target dataset. From Kaps et al. (2023a).

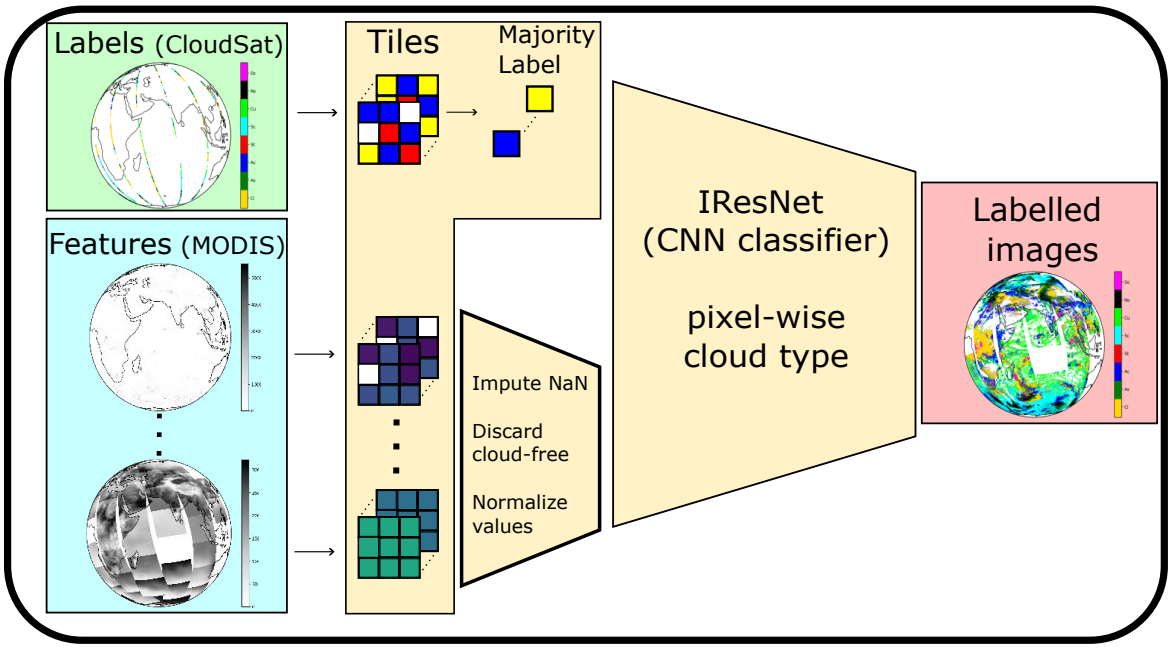


Figure 4.3.: Schematic of the pixel-wise classifier, which is a CNN trained on features from MODIS and one of eight cloud type labels from CC-L per pixel. Adapted from Kaps et al. (2023a).

CNN based on the IResNet (Behrmann et al., 2019). Residual networks (He et al., 2016) have become the baseline for many image-related tasks and the IResNet additionally allows for semisupervised training. The training is termed semisupervised as both labeled and unlabeled samples are fed to the network. The model learns to minimize the cross-entropy for the labeled parts (Eq. 4.2) as well as the negative log-likelihood (Eq. 4.1) of the latent representation $\mathbf{z}$ of all (labeled and unlabeled) samples.

$$\mathcal{L}_u = -\sum_{\substack{\mathbf{z}_k=\mathbf{F}(\mathbf{x}_k),\\ \mathbf{x}_k\in\mathbb{X}}} \log(p(\mathbf{z}_k)) + \mathrm{Tr}(\mathbf{J}_\mathbf{F}), \tag{4.1}$$

$$\mathcal{L}_l = \sum_{\mathbf{x}_k,\mathbf{y}_k\in\mathbb{X}_l} \log(\mathbf{x}_k)\mathbf{y}_k. \tag{4.2}$$

Here, $\mathbf{z}_k = \mathbf{F}(\mathbf{x}_k)$ is the latent output of the IResNet $\mathbf{F}$ without its classifier head, and $\mathbf{J}_\mathbf{F}$ is its Jacobian, with Tr denoting the trace operation. $\mathbb{X}$ contains all samples $\mathbf{x}$, $\mathbb{X}_l$ contains only the samples with labels $\mathbf{y}$. The IResNet is applied to tiles consisting of 3×3 MODIS pixels to determine the cloud type of the central pixel. The training target is the cloud class that occurs most often in the tile, with ties being resolved by random draw. Note that this way, the class label is predicted such that it is representative of the whole tile, even though the label is only assigned to the central pixel. This is a design choice that possibly introduces a bias towards more frequent cloud classes but increases the number of usable tiles both for training and prediction by allowing for overlapping tiles. This is also one of the major modifications to the code used in Z19, where tiles were completely labeled and thus not allowed to overlap. Tiles that contain less than six cloudy pixels according to the MODIS cloud mask are discarded. Therefore, the NN is agnostic to such cases including clear sky situations. By applying the trained model, pixels in the CUMULO data that are yet unlabeled are assigned class labels, resulting in a set of fully labeled satellite data.

The CUMULO data contain MODIS radiance measurements as well as retrieved physical cloud properties (Table **3.3**). With potential application to GCMs in mind, we decided to train the IResNet using the physical variables as features, these being more readily available from GCMs than the radiances at the particular MODIS spectral channels. This is another major modification to the approach of Z19. We found that the classes predicted by the model trained on the physical variable features were slightly more physically consistent. For example, we found that a number of high and thin clouds were given the Cumulus (Cu) label when using the radiances only, but this did not happen when using the physical cloud properties. However, the performance difference was marginal such that the classification step could be trained on either set of features. To be able to perform the training on physical variables, pixels containing missing values e.g. from failed MODIS retrievals are imputed, using the mean value for each 3×3 pixel tile. As the tiles are small, this is not expected to skew the values in the individual tiles significantly as neighboring pixels are expected to have similar properties.

The IResNet is trained on all available CUMULO granules for the year 2008 ($\sim$ 48000 multivariate images of 1354×2030 pixels) with standardized features. Instead of using a train/test

split, we used 2-fold cross-validation[2] to assess generalization to unseen data. The model used for final predictions is then trained on the complete year. Due to the high temporal resolution of the data, the variance in the features for the training data is comparable to that of longer periods typical for GCMs. This will compensate for the fact that only one year of training data is used.

4.2.3. Regression on Low Resolution Data

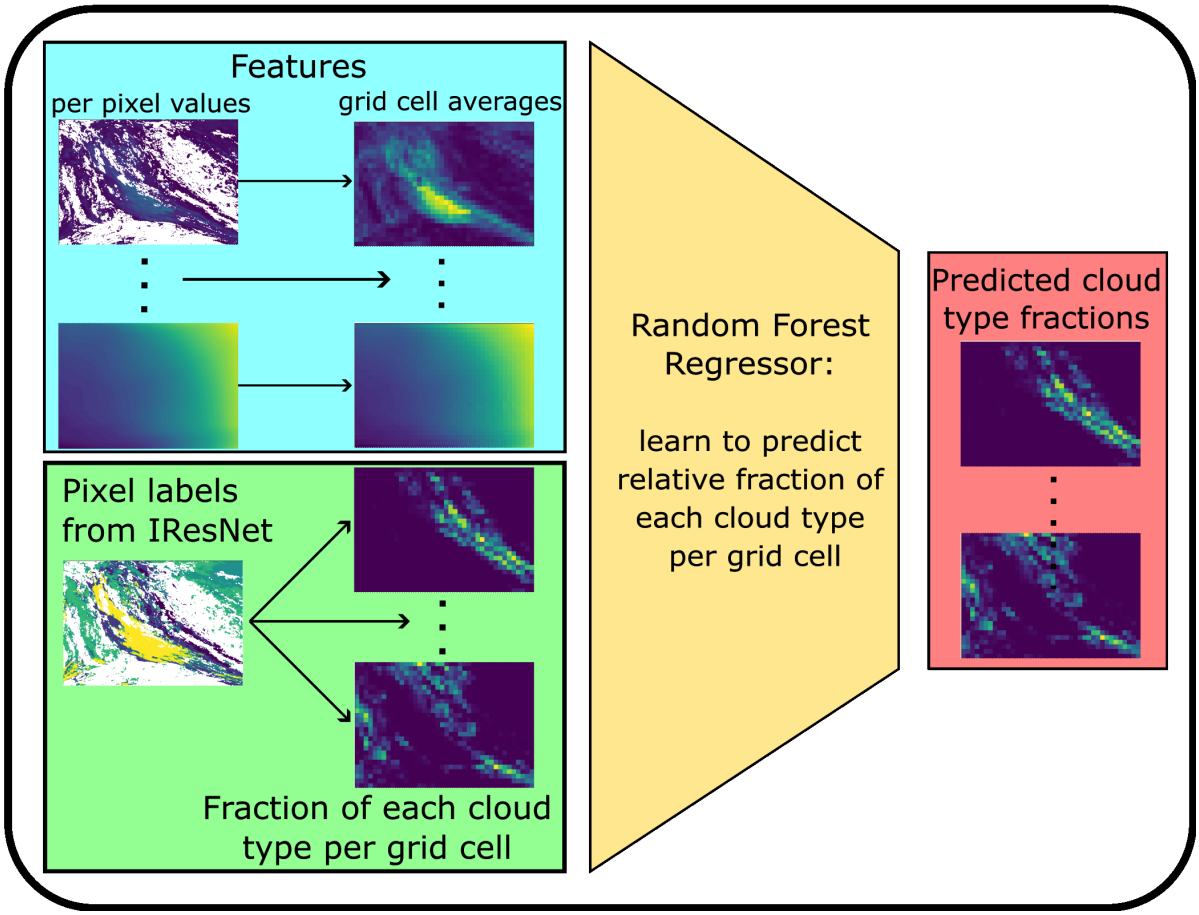


Figure 4.4.: Cloud type predictions for data with low horizontal resolution are obtained by coarse-graining high-resolution predictions as a basis to train a regression model predicting relative amounts of each cloud type for each coarse-resolution grid cell. From Kaps et al. (2023a).

The second stage of ML (see Fig. **4.4**) is designed to transfer the information contained in the (labeled) high-resolution satellite data to datasets of lower temporal and spatial resolution, like typical GCM output. For this, the labeled satellite dataset obtained from the pixel-wise classification is coarse-grained. All variables that are provided in both the CUMULO and the

[2]Kaps et al. (2023a) states that five folds were used but a time limit caused the computation to stop after two folds.

target dataset can be used as features, i.e. this is independent of the features used for the first stage.

The labeled data are provided on an evenly spaced metric grid, but many GCMs are provided on evenly spaced angular grids. The area covered by individual pixels will not match between these two grids and scale differently depending on their geographic location. For simplicity and for the purpose of a proof-of-concept, we determined the grid cell size that on average is most representative of the target grid and use the averages of each variable over these grid cells as features for our model. We later performed tests to determine the impact of the grid resolution and determined that for GCM-typical resolution ranges the results are barely affected (Section **4.3.5**). We assume that the remaining differences between the grids are mitigated by averaging. The output is the relative cloud class occurrence in the grid cell, i.e. the fractional amount of each of the eight cloud classes plus an additional "undetermined" class. The "undetermined" class contains all pixels for which the prediction of a label was not possible due to failed MODIS retrievals, which often indicate clear sky. Missing values for pixels with no cloud are processed accordingly when computing the grid-cell averages (see Section **4.2.5**), such that cloud class fractions are predicted consistently for all properties including those that are not defined for clear sky (e.g. *ptop*). Grid cells containing only "undetermined" pixels are discarded. Thus, we obtain a multivariate regression problem with a nine-dimensional output space, containing the eight classes plus "undetermined" pixels, and up to eleven features (i.e. the number of suitable physical variables provided by the CUMULO data, see Section **4.2.5**). For our model, we choose the RFRM (Breiman, 2001) regression method for reasons of simplicity, computational efficiency, as well as its inherent normalization of the predicted fractions (Section **2.3.1**). After training the RFRM on the coarse-grained classified images, it can be directly applied to the target data, i.e. GCM output, providing cloud class fraction predictions for each grid cell. In order to investigate the sensitivity to the resolution and choice of features used, we trained multiple RFs. The individual training samples are weighted with weights w_i given by the L_1-norm $w_i = ||\mathbf{y}_i - \overline{\mathbf{y}}||_1$, where $\mathbf{y}_i$ denotes the cloud class fractions for the i-th training sample and $\overline{\mathbf{y}}$ the average over all samples used in training. The weighting ensures that samples close to "the average sample" are given less weight in training, to reduce the effect of bias in the data. We have about 48 000 labeled CUMULO granules (multivariate images of 1354×2030 pixels) available. In order to limit the amount of memory required, the RFs are trained on roughly $50 \cdot 10^6$ random samples drawn from a training split of 10 000 labeled data images. The amount of samples varies because grid cells containing only "undetermined" pixels are excluded. The models are then evaluated on a test split from the last two months of 2008, containing about 8400 images. The hyperparameters of the RFRM models are chosen such that the depth of the individual regression trees is ≤ 17. We apply a bagging subsampling fraction of 0.7 and a minimum leaf size of 2, with 400 individual trees per forest. These hyperparameters showed an optimal trade-off between the model's performance and size.

4.2.4. Application to ESA-CCI Data

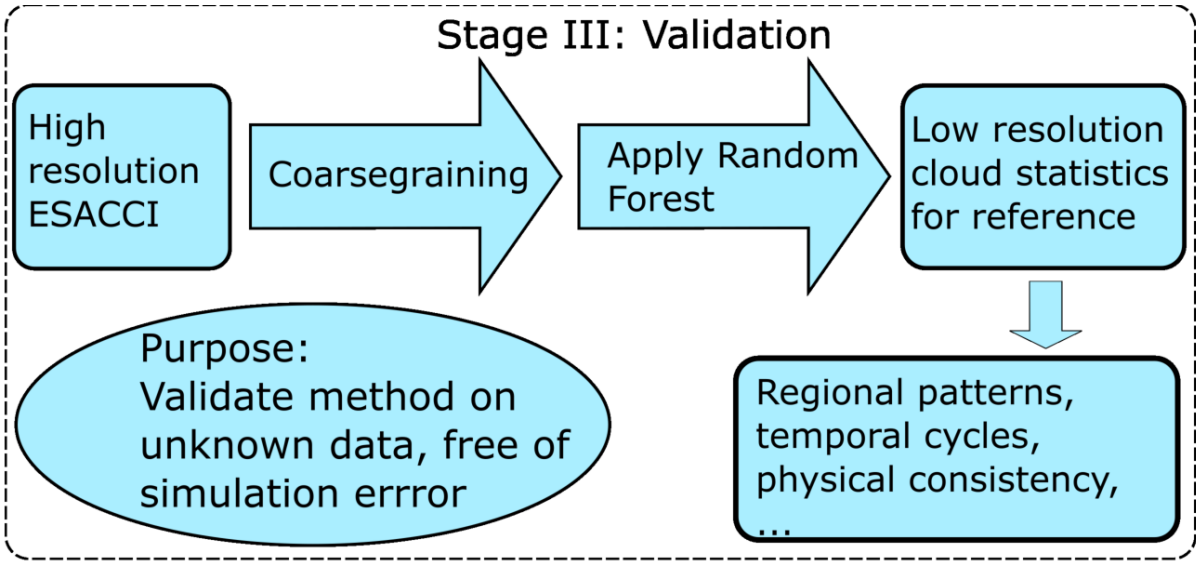


Figure 4.5.: The method is validated by applying the trained regression model to data the model has not seen before. The predictions are then analyzed for physical consistency. From Kaps et al. (2023a).

To validate the trained RF, we apply it to an independent satellite dataset, the ESA-CCI data (see Section **3.2**). Application to the output of current GCMs (e.g. those contributing to CMIP 6) is not yet feasible due to too coarse temporal resolution of the available output and/or key variables being unavailable in the standard output. Currently, GCMs often only contain monthly means in the standard output, which, as shown in Section **4.3.6**, is insufficient temporal resolution. Other models did not provide variables like *cer* or *cod*, which are required for sufficient performance. Therefore, we use this validation stage, as illustrated in Fig. **4.5**, to show that the method generalizes to coarse data obtained from different sources and it is thus expected to also be applicable to other datasets such as suitable GCM output. Compared to the MODIS product, the ESA-CCI data provide a similar representation of the observed cloud state and contain similar physical cloud variables. Inconsistencies might be introduced through the lack of LW-channels in AVHRR, which would mainly affect retrievals of multilayer clouds (Sus et al., 2018). However, the elaborate retrievals in CC4CL, such as using NNs for cloud phase and mask detection might compensate for this. The ESA-CCI dataset is, therefore, a prime candidate to test the RFRM on similar, but completely unseen data. This allows for an assessment of uncertainties and limitations of our method. Knowing possible errors introduced this way can then help distinguish them from simulation errors when the RFRM is applied to GCM output. For each validation experiment using the coarse-grained ESA-CCI data, we randomly sample 20% of the available grid cells.

4.2.5. Features and Preprocessing

The RFRM must be trained on features also available in the target dataset. For ESA-CCI, these include *ttop*, *htop*, *ptop*, *tsurf*, *cod*, *cwp* and *cer*. For the classification stage, the cloud top phase as a categorical variable was used as an input feature. However, categorical values are not straightforwardly coarse-grained. We therefore transferred the phase information contained in both MODIS and ESA-CCI to *cwp* and *cer* by introducing the new variables *lwp* and *iwp* as well as *cerl*/*ceri*. This means that for each pixel, the cloud top thermodynamic phase is applied to the complete vertical column. This procedure is an approximation using the assumption that in most cases, the phase flag provided by the satellite data is representative of the whole cloud column.
The grid box averages for the cloud liquid/ice water path, the radii, *cod* are computed over all pixels in each cell, i.e. replacing missing values with zero. This is useful as these values approach zero with decreasing cloud amount. In contrast, *ptop*, *htop* and *ttop* are only averaged over cloudy pixels ("in-cloud values").
The features used for the RFRM should ideally complement each other. As the features *ptop*, *htop* and *ttop* effectively contain the same physical information, only *ptop* is used. In addition to the cloud variables, we also use *tsurf* as it is readily available in many datasets. As a default, we therefore select *cwp*, *lwp*, *iwp*, *cerl*, *ceri*, *cod*, *ptop*, *tsurf*, which we call the *optimal set* of features in the following.

4.3. Results

4.3.1. Predicted Cloud Classes at Pixel Level

To assess the performance of the IResNet we use accuracy and F1-Score (see Table **4.1**). A qualitative analysis of the physical properties of the predicted cloud classes is used to evaluate the consistency of the results. This is important because the physical properties of the classes predicted by the IResNet model need to be consistent with those from the corresponding genera defined by the WMO.
The labels extracted from CC-L that are available in CUMULO display a strong class imbalance (Table **4.1**), which we also find in the predicted classes. Most classes occur with a similar frequency in the source data and predictions, with deviations being small enough to be attributable to real differences in the data. We would like to highlight two key properties of the class distributions: (1) there are very few stratus (St) and deep convective (Dc) clouds in both the source and the prediction and (2) cumulus (Cu) and cirrus (Ci) clouds are strongly underestimated in the predictions compared to the source data. For example, Cu has the smallest amount of all predicted cloud classes while this class is more common than St and Dc in the

source data. Further assessment of the representation of these four classes (St, Dc, Cu, Ci) is therefore of high importance. The mean accuracy of the classification in the validation splits of the cross-validation is larger than 0.8 for all classes but Sc, which seems to suggest considerable skill in the classification method. However, for a multi-class problem with a large class imbalance, the accuracy is not a suitable measure to assess the performance of the method, as it overvalues the true negatives. This is why we observe high accuracy for classes with few samples like St and Dc. We therefore additionally use the F1-Score, which is sensitive to the class imbalance and can help identify individual class biases in the predictions. The F1-Score is defined as the harmonic mean between precision P and recall R of the model.

$$\mathrm{F1} = 2\frac{P \cdot R}{P + R}. \tag{4.3}$$

The precision P is the ratio of the true positives of a class to the number of all samples assigned to that class. The recall R is the ratio of true positives to the number of samples with a positive ground truth value for that class. For our results, the F1-Score is at least 0.4 for all classes except for St, and especially high for Sc and Cu with values larger than 0.6. When considering the accuracy scores for all classes and high class imbalance, this suggests decent skill in representing the class imbalance, but still some class confusion. A negative outlier is the St class with an F1-Score of 0.21. As noted in Section **3.1.2**, the CC-L algorithm has trouble distinguishing between St and Sc, which is why this is also to be expected for the IResNet, resulting in this rather disappointing F1-Score for St. This effect can be seen in the confusion matrix of the model (Fig. **4.6**), where most of the false negatives of St are due to classification of Sc. The inverse does not happen as often. In general, Fig. **4.6** shows that most of the confusion is between classes that are physically similar or tend to occur together (Ns and As, Dc and Ci). To summarize the uncertainty of the classification, we compute the mean metrics with standard deviation for the validation splits in the cross-validation and obtain an accuracy of 0.886 ± 0.003 and F1-score of 0.472 ± 0.005. Interestingly, both of these metrics are better than what is reported in in Z19, suggesting that our modifications to the algorithm were indeed warranted. While neither accuracy nor F1-Score values are particularly impressive, we found that a random baseline, perfectly reproducing the ground truth class distribution, will achieve similar accuracy with considerably smaller per-class F1-Score between 0.01 and 0.3. The model thus easily outperforms a random baseline, and we show in Section **4.3.3** that the predicted classes are physically meaningful.

4.3.2. Cloud Class Distributions at Coarse Resolution

The RFRM is expensive to train on large datasets such as the year-long, high-resolution CUMULO dataset. Because of these computational constraints, we train the RFRMs on a subset of the labeled data of about 25% the size of the complete dataset, as we found that this is more

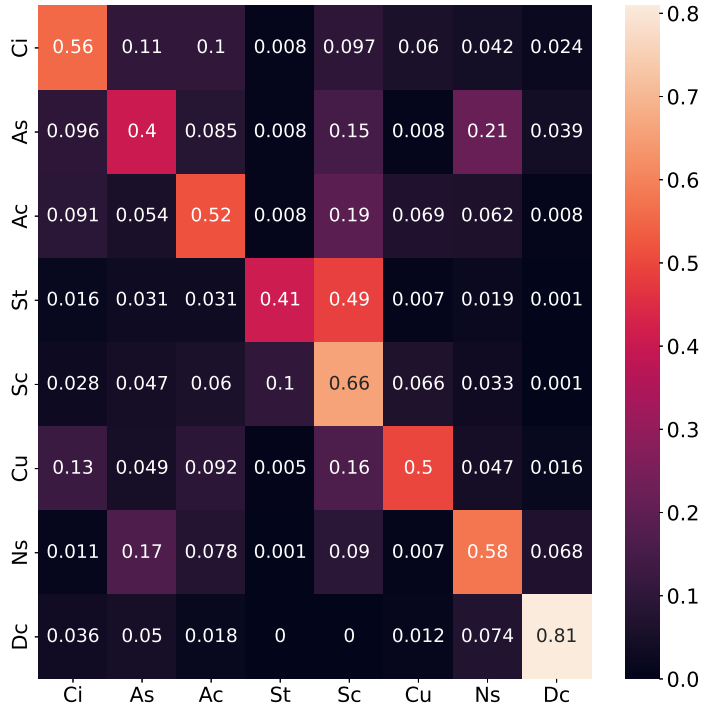


Figure 4.6.: Confusion matrix of the trained IResNet classifier on the validation split with the best metrics. The values are fractions with respect to the ground truth amounts for each class. The diagonal values denote the true positive rate for each class. For a given class, other values in a row denote false negatives, while the column shows the fraction of false positives. We found that the St samples erroneously classified as Sc occur for tiles with properties extremely similar to those of Sc.

Table 4.1.: Fractions of the cloud classes for pixel-wise classification with prediction accuracy and F1 score for the supervised part of the data. CC-L labels are for $21 \cdot 10^6$ labeled pixels included in CUMULO, predictions are for $800 \cdot 10^6$ pixels. Scores are averages over both validation splits. From Kaps et al. (2023a).

	Ci	As	Ac	St	Sc	Cu	Ns	Dc
CC-L fraction	0.259	0.132	0.112	0.021	0.313	0.065	0.082	0.015
Predicted fraction	0.154	0.10	0.180	0.027	0.353	0.014	0.134	0.041
Prediction accuracy	0.85	0.84	0.88	0.95	0.78	0.91	0.90	0.98
Prediction F1-Score	0.64	0.40	0.44	0.21	0.66	0.46	0.48	0.48

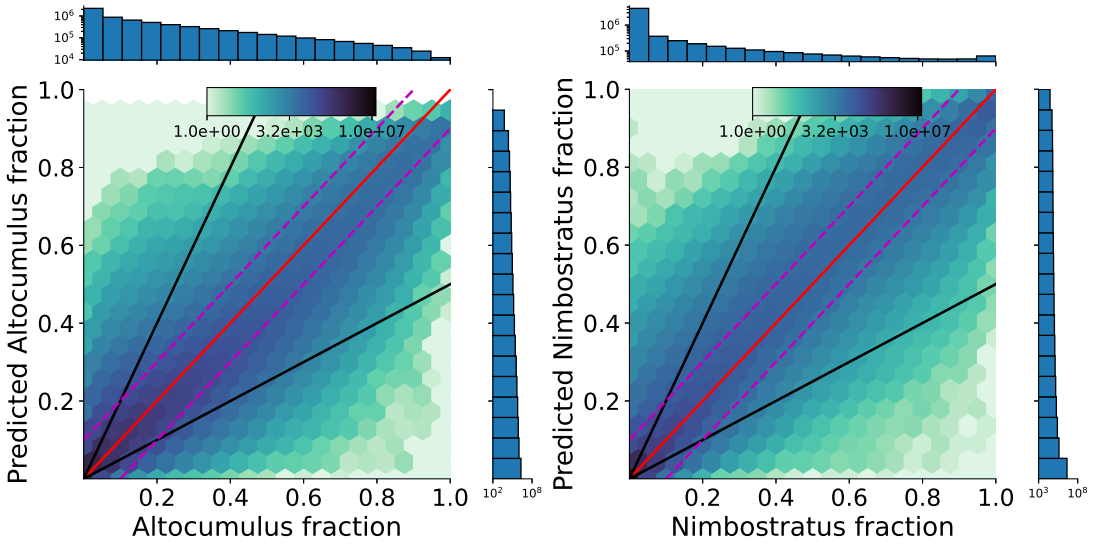
than sufficient for a stable error. The mean errors and R2 scores (Eq. 2.33) for the different settings are summarized in Table **4.2**. Since with smaller grid cell sizes, more cells containing only "undetermined" pixels are excluded, and the relative amount of cloudy pixels increases, such that we see larger mean absolute errors for small grid cells. Using the median, however,

we see better performance for smaller cells. The R2-score increases with grid cell size, most likely due to the decreasing variance caused by averaging over more pixels. The performance is therefore judged not to be strongly dependent on the grid box size. We also use joint densities of predicted and ground truth cloud-type fractions of the test split as a performance indicator. Figure **4.7** shows these for a grid cell size of 100 km × 100 km using the optimal set of features (see Section **4.2.5**). The joint density plot displays the concentration of samples in the truth/prediction space, and along the x- and y-axis the marginal distributions of the true and predicted fractions, respectively. For both cloud classes in Fig. **4.7**, there is a clear correlation between the ground truth and the predictions with a Pearson correlation of $c_P = 0.96$ for Ns and $c_P = 0.89$ for Ac. Many predictions are, however, far off the target: Fig. **4.7b** shows several hundred samples with a predicted Ns fraction of about 0.2 where the true fraction is close to 1. For this specific example, this is a small fraction ($\mathcal{O}(0.001\%)$) of the total number of grid cells, but it shows that the predictions can differ strongly from the true values in a non-negligible number of cases. This deviation is a manifestation of ambiguity between different cloud states, likely caused by noise generated by the averaging of the features. Furthermore, this is an example of the predictions favoring low cloud-type fractions, as the "undetermined" class is prevalent in the training data. As shown in Fig. **4.7a**, large altocumulus (Ac) fractions (> 0.9) are underestimated by the RF, but the deviation in this region remains largely below 0.1, as indicated by the magenta dashed lines. For fractions larger than 0.2, samples deviating by more than a factor of 2 (outside black lines) are rare. For fractions smaller than 0.2 (bottom left corner), deviations by a factor of more than 2 occur frequently, indicating difficulty in correctly predicting small fractions. Note that such predictions contribute significantly to the relative error, but have a negligible effect on the absolute error. We construct a random baseline by sampling from the class distributions in the IResNet predictions. We find that the mean absolute deviation is larger for the random baseline by roughly a factor of five, indicating that the regression model outperforms the random baseline. As the variables available

Table 4.2.: Results of the regression models for different grid box sizes. (1): Trained using a default set of features. (2): Using *cwp* and *cer*, not separated into ice and liquid, in addition to *cod, ptop, tsurf*. (3): Using (*cwp, cer, ptop, tsurf*) as features. From Kaps et al. (2023a).

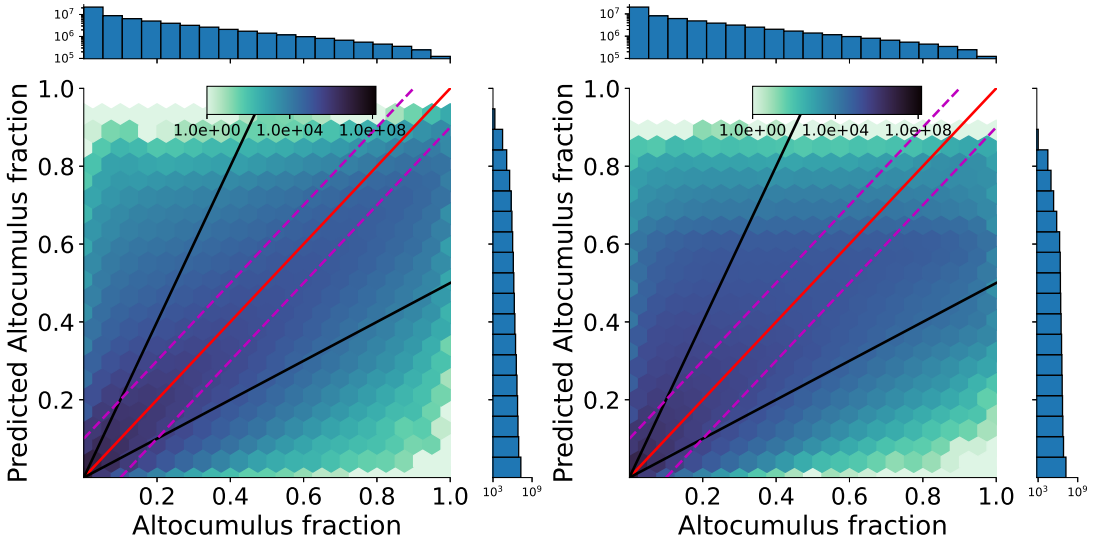
Grid cell size	Mean absolute error	Median absolute error	Median relative error	R2-Score	Random mean AE
3 km (1)	0.042	0.0018	18.9%	0.816	0.193
10 km (1)	0.036	0.002	33.9 %	0.836	0.178
20 km (1)	0.033	0.0025	41.3%	0.840	0.169
100 km (1)	0.027	0.0038	52.6%	0.845	0.151
200 km (1)	0.025	0.0045	54.2%	0.859	0.143
100 km (2)	0.028	0.0041	55.4%	0.837	0.151
100 km (3)	0.033	0.0043	60.4%	0.755	0.151

in typical GCM output can vary, not always matching our optimal set of features, we also determine which of these features are essential to achieve good performance. In addition to



(a) (b)

Figure 4.7.: Joint density of the predicted and true Ac (left) and Ns (right) fractions from the CUMULO test set for a grid cell size of 100 km × 100 km, using the optimal set of features (Section **4.2.5**). The color scale and the marginal histograms are logarithmic. The red line indicates the line of perfect correlation. The area between the dashed magenta lines indicate a deviation between ground truth and prediction of less than 0.1 and the area between the black lines indicates a deviation by less than a factor of two in either direction. From Kaps et al. (2023a).



(a) (b)

Figure 4.8.: Results for a model trained using features without liquid/ice distinction (*cwp, cer, cod, ptop, tsurf*) (Fig. **4.8a**) and a model trained also without *cod* (Fig. **4.8b**), for comparison with Fig. **4.7**. From Kaps et al. (2023a).

using the optimal set of features, we therefore also train the model using different alternative sets, containing fewer features. Using the cloud top phase flag to distinguish between ice and liquid for some of our features (*lwp*, *iwp*, *cerl*,*ceri*) produces a small performance increase as

the metrics indicate in Table **4.2**. Comparing Fig. **4.7a** and Fig. **4.8a**, shows that the correlation between the true and predicted values becomes less pronounced when the information about the thermodynamic phase is removed. Further ablation studies reveal, that using *cod* and *ptop* is critical for the RFRM performance, but these variables are infrequently contained in GCM standard output at sufficient temporal and spatial resolution. An example is shown using features without *cod*, where Table **4.2** shows a significant decrease in the R2-score. The effect on the joint density displayed in Fig. **4.8b** is visible as predicted fractions being skewed towards smaller values. Predicted fractions above 0.8 are very rare and the joint density seems to be shifted towards the lower black line, corresponding to half the true value.

4.3.3. Validation

To assess the generalization performance of the method, we compare predictions of the class distributions on ESA-CCI to the classes from the CC-L product for the year 2008, downloaded from `https://www.cloudsat.cira.colostate.edu/` (last accessed 7$^{\text{th}}$ Dec. 2023). A comparison of this kind is valuable because it simultaneously quantifies the effects of applying two ML models consecutively as well as changing the domain on which they are applied. As before, the 3-dim CC-L data are aggregated into two dimensions by using the most common cloud class within each vertical column as a representative cloud class. The labels provided by CC-L for individual orbits are sparse, but using a whole year of measurements provides sufficient samples to compare to the predicted distributions. Figure **4.9** shows the distributions of CC-L labels, which appear sparse, even though the data have been aggregated to grid cells of $2° \times 2°$. Consequently, not all cloud classes display clear regional patterns, but Ci, Sc, and Cu show distinct areas of frequent occurrence. For example, Sc clouds are frequently detected in the subtropical subsidence regions off the west coasts of the continents, Ci clouds are frequent in the deep Tropics, Cu is found frequently over the tropical and subtropical oceans away from the stratocumulus decks. In the following, for a better comparison of the CC-L ground truth and the predictions on ESA-CCI, we exclude the "undetermined" predictions such that the cumulative fraction of all eight cloud classes equals one in each cell. The reported fractions are therefore a relative measure and independent of the total cloud amount in each grid cell. Figure **4.10** shows the predictions on ESA-CCI using the RFRM trained on grid cells of $10\,\text{km} \times 10\,\text{km}$ and applied 10×10 pixel grid cells. Different classes occur in distinct patterns and the Sc class dominates in the predictions, while St and Dc occur very rarely.

Assessing the Method Uncertainty

When applying this framework to GCM output, deviations introduced by the data need to be separable from those caused by the evaluation method. As an uncertainty estimate for

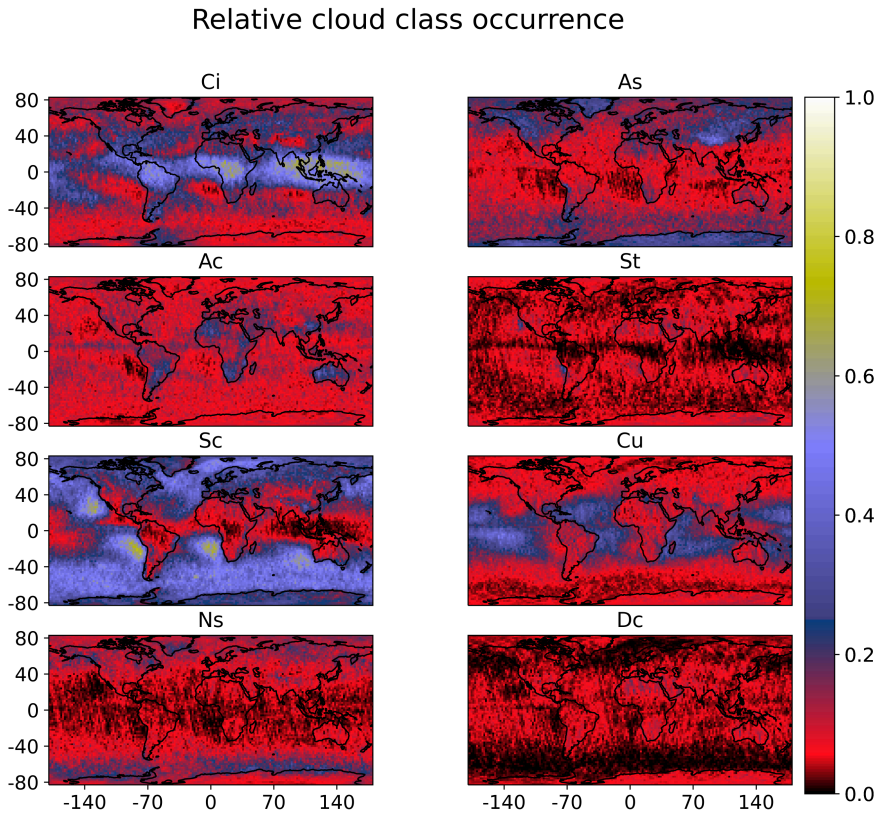


Figure 4.9.: Relative occurrence per class from the CC-L product for the year 2008 per $2° \times 2°$ grid cell. From Kaps et al. (2023a).

the consecutive application of both ML methods, we compute the Pearson correlation and difference between the predictions on ESA-CCI data and the CC-L labels. For this purpose, we bin the relative amount of each class to grid cells of $2° \times 2°$ size for both datasets. As an example, differences for the cloud types with the highest (Sc) and lowest (St) correlation are shown in Fig. **4.11**. The difference increases with the fraction of occurrence of each class (as displayed in Fig. **4.10**). Note that this is only a rough measure of accuracy as the two datasets differ in temporal and spatial resolution. Additionally, an exact match cannot be expected as the CC-L covers the year 2008, while ESA-CCI covers the period from June 2009 to the end of 2011. The mean within-class correlation is 0.65. Table **4.3** shows the mean fractions of the classes in the predictions and the CC-L data. The predictions here are comparable to the pixel-wise predictions obtained using the IResNet (Table **4.1**). The most notable difference in the distribution is again the under-representation of Ci in the predictions relative to the CC-L labels, which is caused by the under-representation in the predicted pixel-wise labels. Table **4.3** also shows the relative difference between the two distributions for grid cells showing a large class fraction in the predictions (90th percentile). For all classes, the magnitude of this deviation is below 50%. This is also the range of relative deviation we found on the test split, leading to an overall estimate of the uncertainty of 50%.

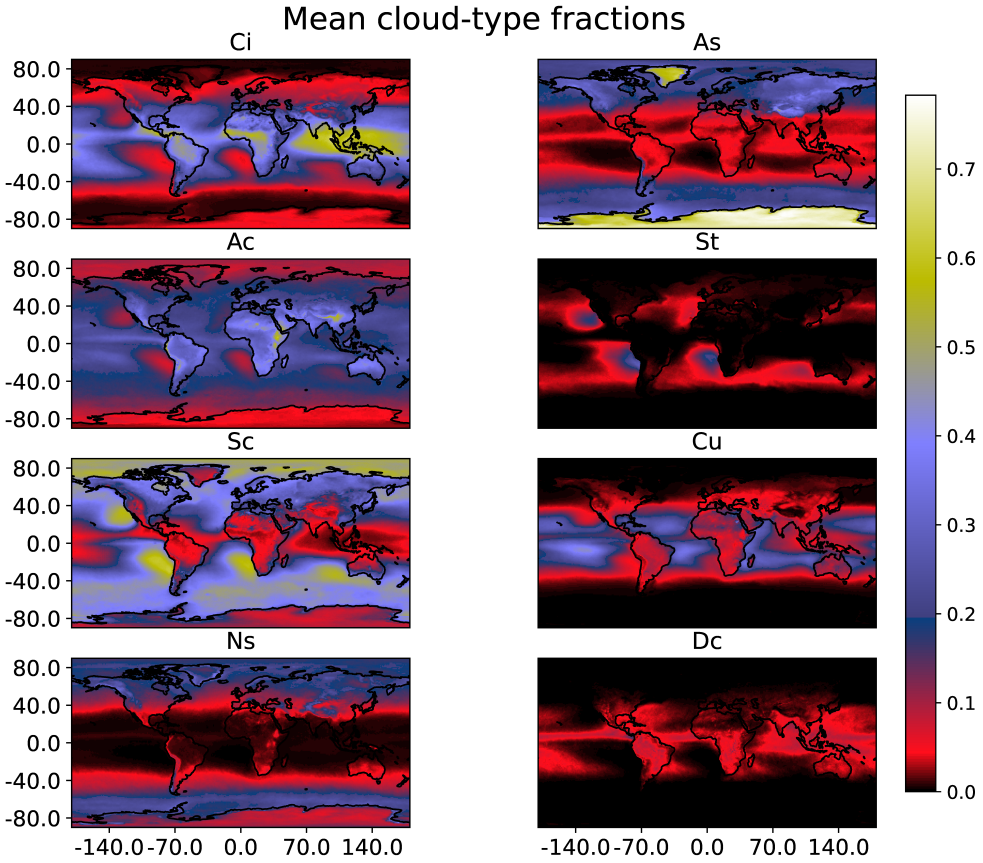
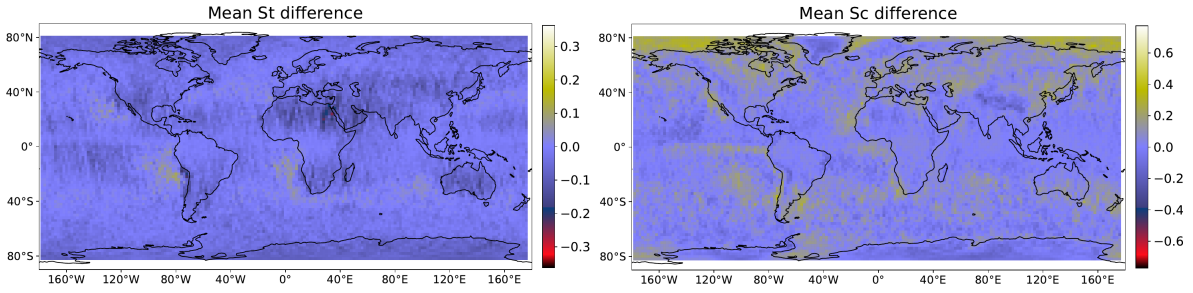


Figure 4.10.: Average class fractions for the predictions on coarse-grained ESA-CCI. The RFRM was trained on 10 km × 10 km grid cells and applied to 10 × 10 pixel grid cells. The results are projected onto a 1° × 1° grid. Many classes show similarities to the distributions in the CC-L data (Fig. **4.9**), even though the location is not used as a feature. From Kaps et al. (2023a).



(a) (b)

Figure 4.11.: Difference between the mean predicted fractions and CC-L per 2° × 2° grid cell for the relative amount of the classes with lowest/highest correlation, St (Fig. **4.11a**,$c_P = 0.18$)/Sc (Fig. **4.11b**,$c_P = 0.88$). The color map is normalized to the range $[-m, m]$, where m is the maximum value for the class across both (CC-L, predictions) distributions. Predictions obtained from RFRM trained on 100 km × 100 km data and applied on 100 × 100 ESA-CCI pixels. From Kaps et al. (2023a).

Table 4.3.: Mean fraction of the predicted classes compared with the relative amounts of the classes in CC-L. The last row shows the mean difference for pixels with predictions in the 90th percentile Δ_{90} relative to the mean μ_{90} of these predictions. Predictions are taken from a model trained on the default set of features using 100 km × 100 km and applied on 100 × 100 pixel ESA-CCI grid cells. From Kaps et al. (2023a).

	Ci	As	Ac	St	Sc	Cu	Ns	Dc
Predictions	0.13	0.14	0.19	0.01	0.31	0.10	0.10	0.02
CloudSat	0.20	0.13	0.11	0.05	0.27	0.12	0.09	0.03
c_P	0.87	0.80	0.60	0.18	0.88	0.84	0.83	0.36
Δ_{90}/μ_{90}	-29%	49%	49%	1%	18%	14%	30%	39%

Physical Consistency

By analyzing the features associated with each class we can show that the properties of the predicted classes are consistent with the expected properties from the meteorological definition of the respective cloud type (WMO Cloud Atlas). In Fig. **4.12**, we show the distribution of the features for the two most common classes Ci and Sc. The values for each feature are normalized to the range $[0, 1]$, with zero and one corresponding to the minimum and maximum value observed across all classes. We find that the Ci class predicted by the IResNet has a higher cloud top and is exclusively flagged as ice. In addition, Ci clouds show a smaller water content and optical thickness. Similarly, we find that clouds are identified as Sc when a tile contains low-level clouds with a higher liquid than ice content and a higher optical thickness than Ci. Most other cloud classes also show consistent feature characteristics. The exception are the classes Sc and St, which the IResNet predicts for very similar ranges of the feature values. This is, however, in itself in line with the definitions of the cloud classes, as stratus and stratocumulus clouds have similar physical properties, leading to potential difficulties in the identification between the two classes. The analysis of the physical consistency of the regression results requires additional steps. As most grid boxes include several classes, we cannot simply show the feature values associated with individual cloud classes like we did for the classification. Instead, we analyze the characteristic feature values for grid cells predicted to have an especially high fraction of a specific class. As an example, Figure **4.13** shows the feature value distribution for grid cells predicted to have a high fraction of the Ci or Sc classes, respectively. This allows for a comparison between the properties of the classes obtained in the pixel-wise classification (Fig. **4.12**) and the class fractions of the grid boxes (Fig. **4.13**). Most prominent are the differences in liquid/ice particle radii between the two sets of grid cells. The respective values are consistent with cirrus being comprised of ice and stratocumulus consisting mostly of liquid particles. These plots also show that grid cells with a high fraction of Sc typically have higher cloud top pressures (i.e. lower cloud top heights) and contain more condensed water than grid cells with a high fraction of Ci. As expected, we find that the predictions of the regression model for high fractions of a cloud class are based on similar feature values as the predictions on the pixel level. This shows that the regression model

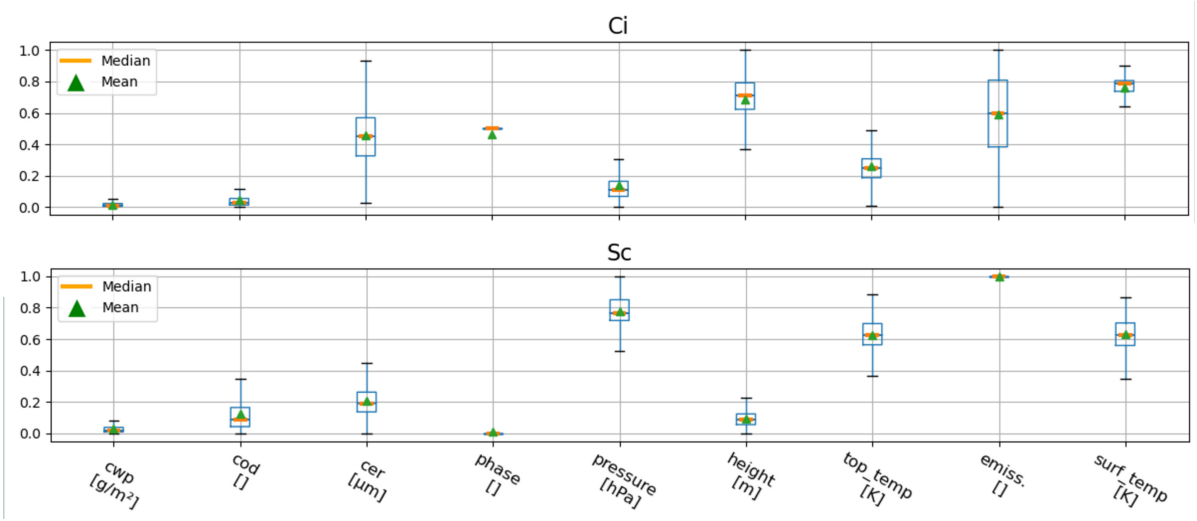
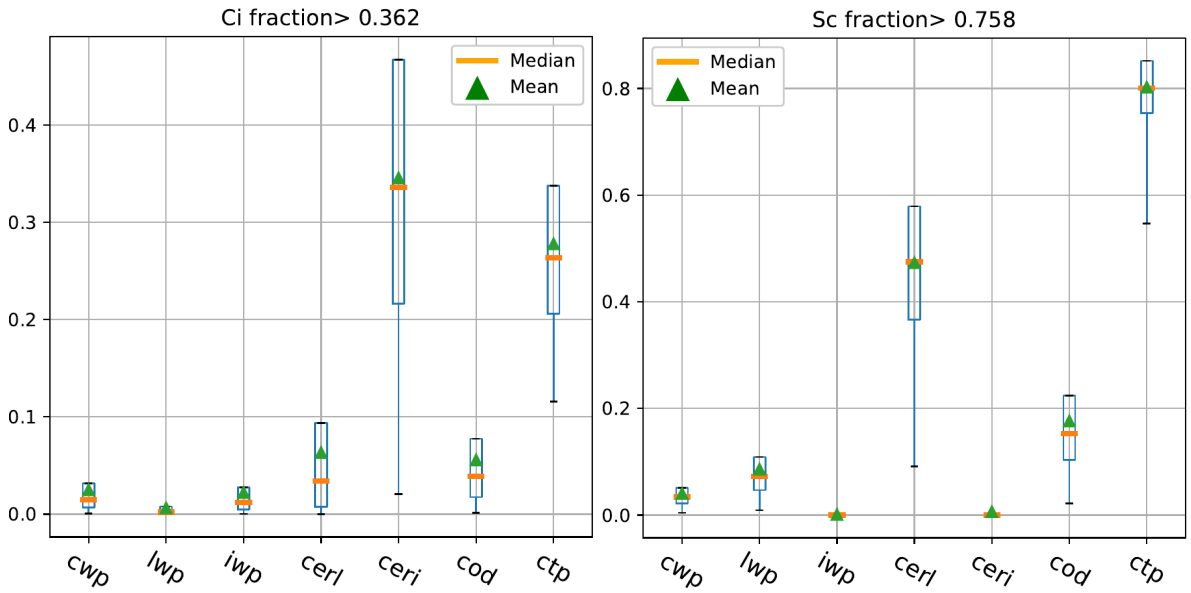


Figure 4.12.: Distributions of the feature values for IResNet predictions of the two most common classes in the CUMULO classification, Ci (top) and Sc (bottom). The values are min/max scaled such that the feature values for all classes lie in the same range. Phase is a categorical feature: 0 for liquid, 0.5 for ice and 1 for undetermined. Boxes extend from lower to upper quartile, and whiskers extend from 10th to 90th percentile. Comparing the relative locations of the boxes between the classes allows assessment of the physical properties the IResNet associates with the respective classes. For example, the cloud top pressure values for Ci are at the low end of the range, while the opposite is true for Sc. From Kaps et al. (2023a).

can indeed predict cloud classes from average states of large grid cells distinctly and that the physical basis used for the CC-L classification is propagated throughout the individual steps of our method. We can perform a similar analysis on the results obtained using ESA-CCI, also taking into account regional prevalence of certain features. Figure **4.14** shows global maps of predictions for the Ci and Cu classes and the features *iwp* and *ptop*. We find a high Cu fraction, particularly in the subtropical subsidence regions characterized by a high average cloud top pressure. These areas contain essentially no predicted Ci clouds. In contrast, Ci clouds are frequently predicted in low- and mid-latitude regions and are characterized by a large ice content. This is consistent with the expectation that cirrus clouds are characterized by high cloud tops and high relative ice content, whereas the opposite applies to shallow cumuli. Indeed, the fraction of Ci is slightly positively correlated with *ceri* (Pearson correlation $c_P = 0.18$) and anti-correlated with *ptop* ($c_P = -0.56$), while the opposite is true for Cu (*ceri*: $c_P = -0.4$, *ptop*: $c_P = 0.29$). This is another indicator for the physical consistency of the predicted class fractions with the corresponding WMO cloud genera.

4.3.4. Feature Importance

Beyond the physical relationship of the classes to the features, we can also determine which features are important for the model to provide good predictions. We did this for the regression



(a) (b)

Figure 4.13.: Distribution of the feature values of grid cells predicted by the RFRM to have a large fraction of Ci (left) or Sc (right) (90% quantile, i.e. thresholds 0.362 for Ci and 0.758 for Sc; see also Fig. **4.12**). The boxes indicate the ranges of the individual features that the RFRM associates with especially high occurrences of the respective class. Box ranges as in Fig. **4.12**; values scaled such that 0 and 1 correspond to the minimum and maximum values found across all classes, respectively. From Kaps et al. (2023a).

model because this will be the model that is applied on new datasets, where not all features might be available and a new RFRM might need to be trained. Note that the following analysis applies only to this specific model. A model trained on fewer features might rely on a different combination of these.

First, we analyzed the features of the regression model using the permutation importance method. This method quantifies the impact of shuffling individual features throughout the data while keeping all other features fixed. The importance is measured in terms of the decrease of chosen metrics, of which we use both the R2-score and the MSE. These metrics are displayed in Fig. **4.15** for an RFRM trained on 100 km × 100 km grid cells. The permutation importance can both be computed on the train (Fig. **4.15b**) and test data (Fig. **4.15a**). A high importance in the train split can indicate that the model overfits on the respective feature. The importance in the test split highlights the features important for generalization. However, in our case the permutation importance is virtually identical for both splits, indicating good generalization to the test set. The features *tsurf* and *ptop* seem to have the largest impact on both the R2-Score and the MSE when permuted, followed by *ceri* and *cod*.

The permutation importance can be skewed when features are correlated, as information about the permuted feature can be inferred from its correlated values. As Fig. **4.15d** shows, all features but *tsurf* are strongly correlated with at least one other feature ($|c_P| \geq 0.69$). This might lead to a comparatively high feature importance for *tsurf*. However, using correlated features

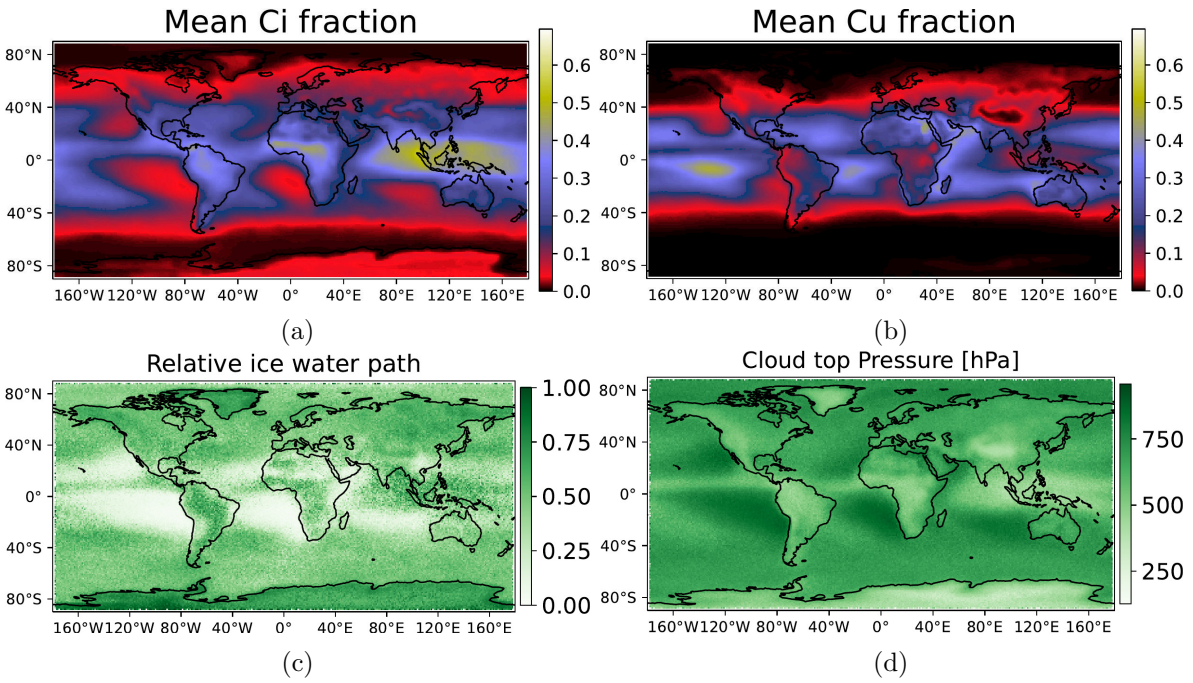


Figure 4.14.: Mean predicted class fractions for Ci (Fig. **4.14a**) and Cu (Fig. **4.14b**) compared to the mean feature values for *ptop* (Fig. **4.14d**) and the relative ice water path ($iwp/(lwp+iwp$, Fig. **4.14c**). Grid cells are 100×100 pixels and the RFRM was trained on $100\,\text{km} \times 100\,\text{km}$ grid cells using the optimal set of features. From Kaps et al. (2023a).

has still proven to be useful to the results. Even though most of the features are correlated with at least one other feature, these correlated features provide additional information that would otherwise be lost when spatially averaged. We additionally took into account the mean decrease in impurity (MDI) attributable to the individual features (Fig. **4.15c**). The importances produced this way are similar to those indicated by the permutation importance using MSE.

Taken together, these feature importance measures indicate that the model relies strongly on *ptop* for its predictions. The results for the other features are less clear but a strong dependence on *ceri* and *cod* is likely, since their impact ranks high across all measures. Due to the correlated features and the fact that the MDI can only be determined for the test set these results however only give a rough indication about which features are required for successful application of our method to GCMs. The high importance these measures attribute to *tsurf* is however unexpected, even though it is most likely related to *tsurf* being the least correlated to the other features.

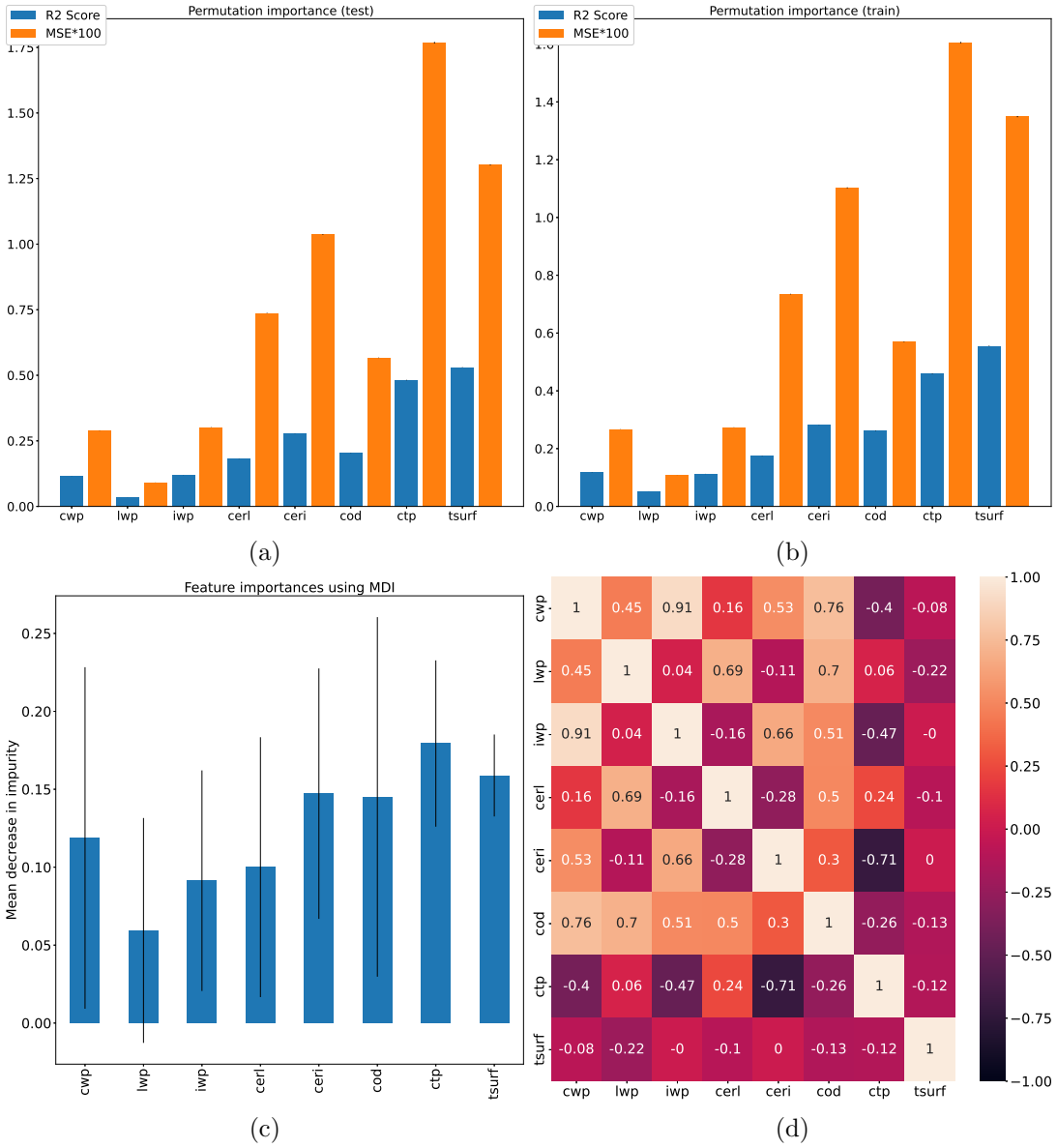


Figure 4.15.: Analyzing the impact the individual features used in the regression have on the predictions of the RF. Figure **4.15a**: Permutation importance, in terms of the R2-Score and the mean squared error (scaled for visibility), computed on the test split of the data. Standard deviation is shown but so small it is barely visible. Figure **4.15b**: same as Fig. **4.15a** but computed on the train split. In both cases the importance was computed for a dataset of one million samples, using 20 different permutations per feature. Figure **4.15c**: Mean decrease in impurity with standard deviation over all trees. Figure **4.15d**: Pearson correlation of the features . From the supplementary material of Kaps et al. (2023a).

4.3.5. Impact of Changing the Coarse-graining Resolution of ESA-CCI

Coarse-graining MODIS and ESA-CCI such that corresponding grid cells everywhere on the globe cover the same area cannot be achieved without interpolative regridding, which we want to avoid here. We instead aim to find a coarse-grained resolution that is most similar to the

others across the globe and assume that the grid cell averaging mitigates resulting differences. We have therefore evaluated the RFRM trained on a fixed grid cell size from coarse-grained MODIS on different resolutions of ESA-CCI. We applied an RFRM trained on 100 km × 100 km grid cells on 0.5° × 0.5°, 2.5° × 2.5° and 5° × 5° ESA-CCI grid cells. Figure **4.16** compares these results using the most common cloud type, binned to a 1° × 1° grid as well as the zonal average of the As class. Note that again, the cloud distributions per coarse-grained grid cell are interpreted as point values for the grid cell center. The representation using the most common cloud type, while being useful to represent the result in a single plot, hides the less prevalent cloud types, such that only Sc (teal), Ci (gold), As (dark green), Ac (blue) and Cu (light green) are displayed. The results show, that the global class distributions stay similar, but some features shift depending on the resolution. The other cloud types do however not disappear completely in the predictions but are being "hidden" in this representation. For example, we see that the Cu class gets hidden by Ci when the resolution is decreased. An interesting difference is the increase in As fraction in central Asia with decreasing resolution. Furthermore, the zonal averages of Ac show that with decreasing resolution the variability of the predictions decreases as well. The zonal mean of Ac however stays the same for all three resolutions.
We conclude that if trained on data coarse-grained to around 100 km × 100 km, the method can provide useful predictions on any resolution typical for GCMs. Even though some classes with similar regional occurrence might "switch places" in terms of their rank in the average occurrence, the zonal distributions stay essentially the same across resolutions indicating no fundamental changes.

4.3.6. Impact of Temporal Resolution

The results shown in the main paper were produced by training and applying the ML models to instantaneous data. GCMs, however, often provide output in the form of daily or monthly averages. This averaging process might introduce deviations from the feature distributions obtained with instantaneous data. The impact of using temporally averaged data is investigated in the following using averaged inputs for the regression model. Here, we use the two daily measurements from the ascending and descending orbits provided by ESA-CCI. We compare the results to the ones obtained with the same regression model applied to instantaneous data. Exemplary results for using features from the instantaneous and mean input data are shown in Fig. **4.17** and Fig. **4.18**, respectively. For all classes, the distributions look very similar, also to the results obtained using an RFRM trained and applied on higher resolution data (Fig. **4.10** in the main paper). The per class correlations range between $c_P^{Dc} = 0.714$ and $c_P^{As} = 0.893$, with unweighted mean $\overline{c_P} = 0.839$, giving an indication that using daily averages instead of instantaneous features for this method is possible. We also applied the RFRM trained on 100 km × 100 km grid cells to ESA-CCI monthly mean (L3C) data. Figure **4.19**

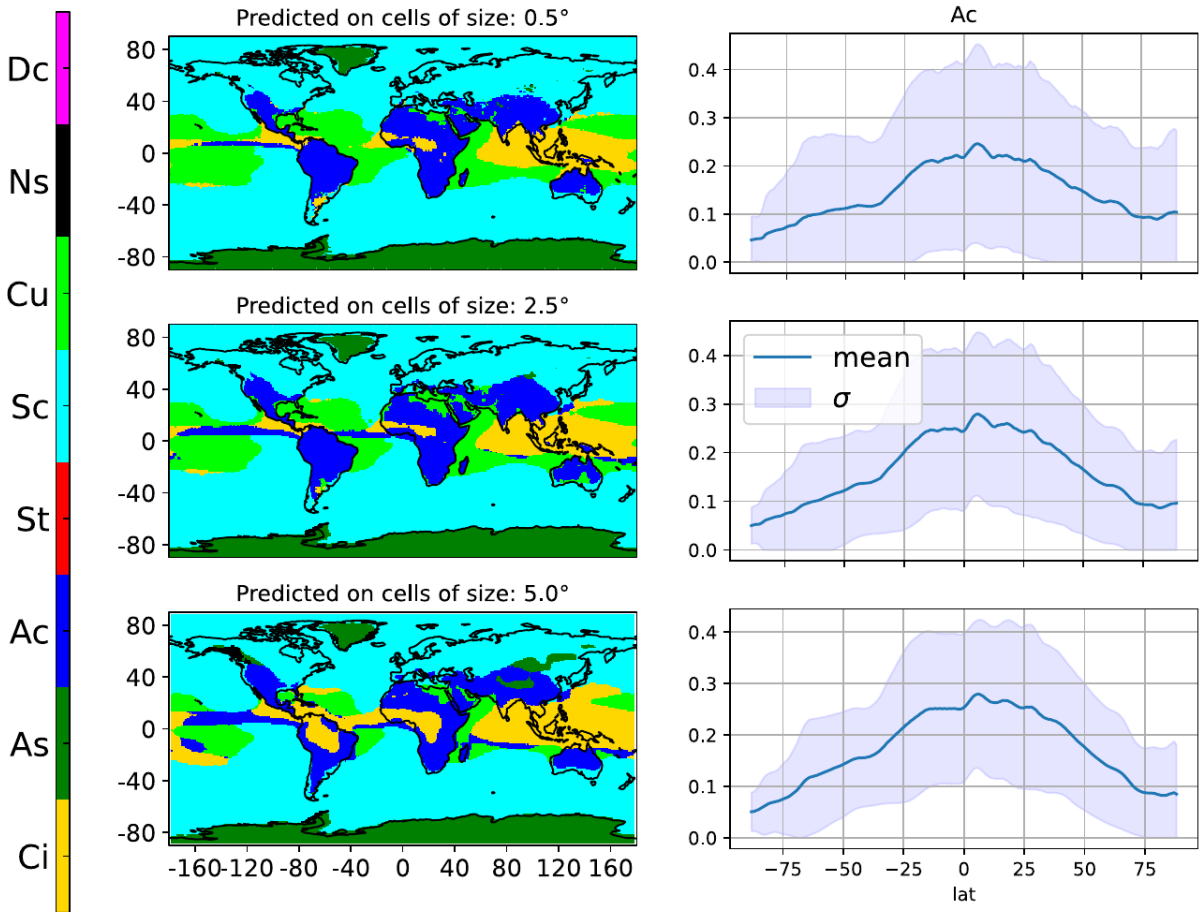


Figure 4.16.: Results for predictions performed on ESA-CCI with different coarse-graining resolutions. The panels on the left show the most common cloud type in each $1° \times 1°$ area, while those on the right show the zonal average for the Ac class. From the supplementary material of Kaps et al. (2023a).

shows that the average predicted fractions are similar to those of the instantaneous input data (Figs. **4.17** and **4.18**), but with less pronounced geographical features. As one would expect from using monthly means, the predicted fractions appear smoothed out and show rather similar magnitudes over large areas. The most notable difference in the representation of the individual classes with respect to using instantaneous data is the globally increased amount of the predicted Ns fraction. We see a further increase in Sc and a decrease in St compared to the distribution in CC-L, suggesting that monthly mean data are not well suited as an input to our method. The increase in both Ns and As further suggests that the monthly average data show much larger effective ice particle radii than the instantaneous data. In fact we find a 4 fold increase in the median ice particle radius and similarly large increases in the cloud water path. Since both, the L3C and L3U ESA-CCI data, are derived from the same instrument (AVHRR), the time averaging of the data must somehow cause this increase, which in turn causes the regression to produce unreliable results.

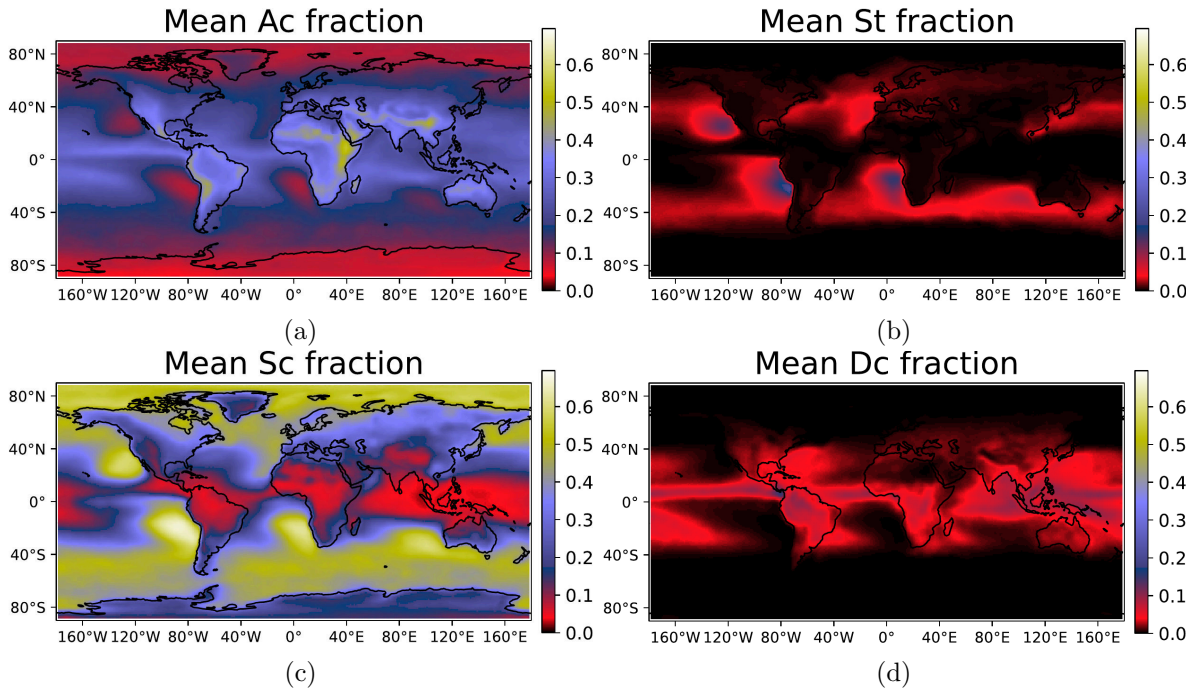


Figure 4.17.: Examples of predicted mean class fractions using feature values from instantaneous source data. Predictions were made using an RFRM trained on 100 km × 100 km grid cells applied on 100 × 100 ESA-CCI pixels. From the supplementary material of Kaps et al. (2023a).

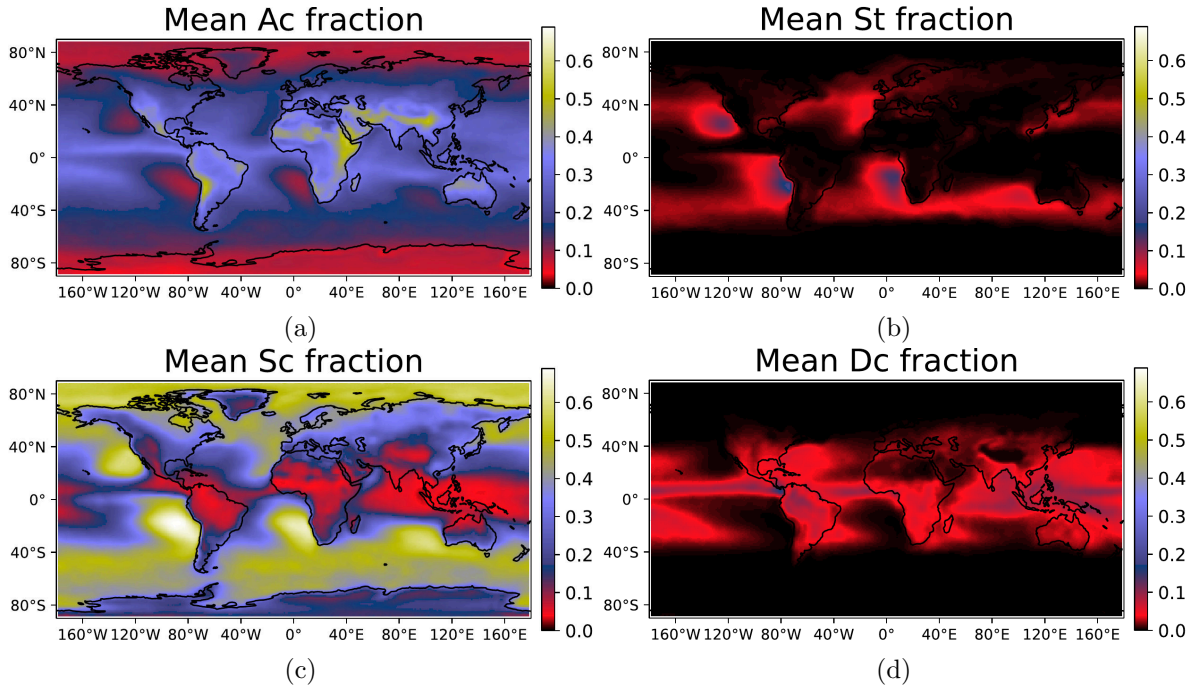


Figure 4.18.: Mean class fractions using feature values obtained by averaging over ascending and descending orbits. Predictions were made using the same RFRM as for Fig. **4.17**, using 100 × 100 ESA-CCI pixels as well . From the supplementary material of Kaps et al. (2023a)

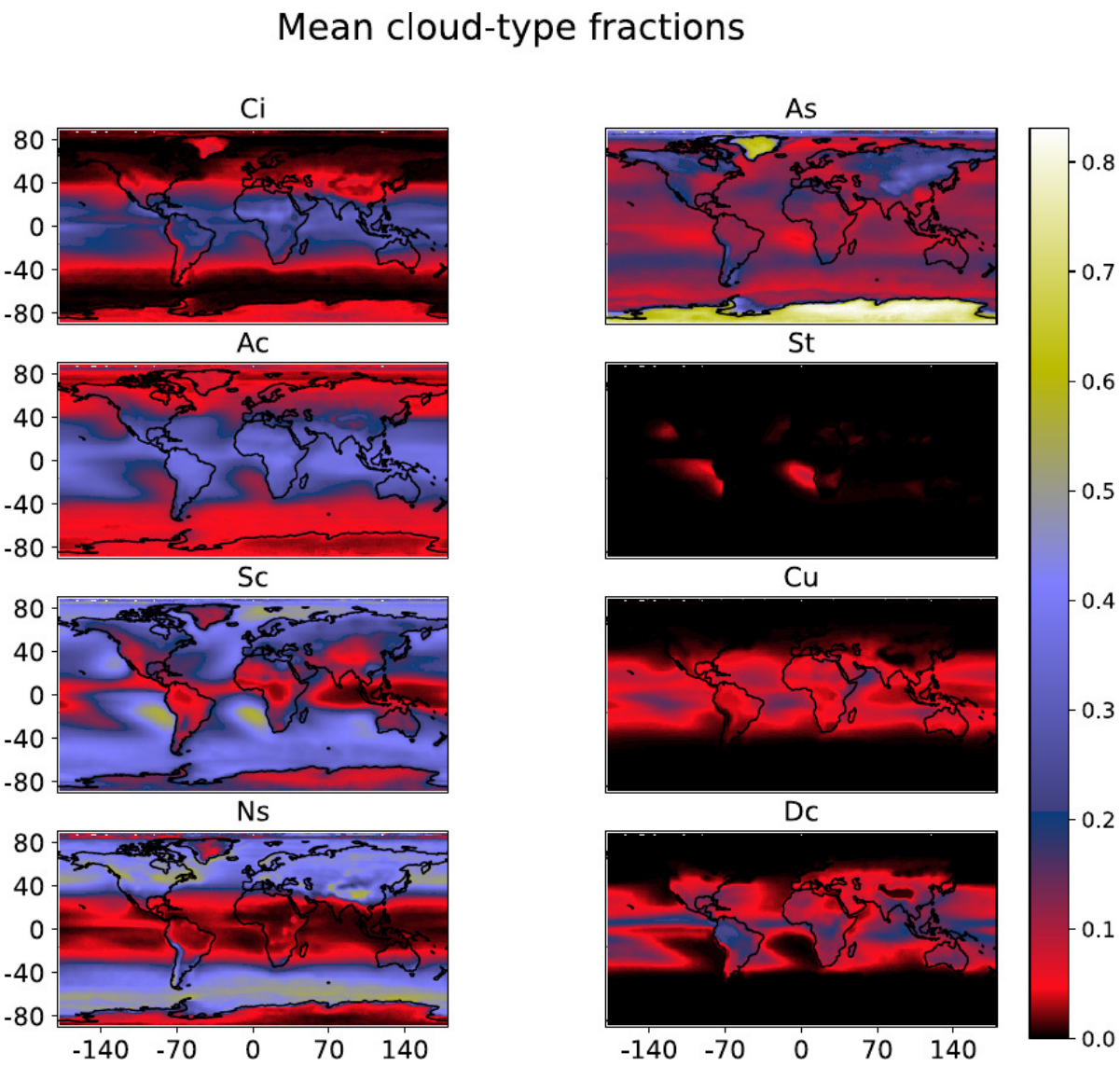


Figure 4.19.: Average predicted fractions using features obtained from monthly mean data (L3C), using the RFRM trained on 100 km × 100 km grid cells. The source data consists of 220 months sampled randomly between 1984 and 2016. The grid cells are constructed from 10 × 10 ESA-CCI pixels of 0.5° resolution. From the supplementary material of Kaps et al. (2023a)

4.4. Summary

We presented a method for the evaluation of clouds in coarse resolution data, employing the consecutive application of machine-learned classification and regression models. Using this method, information on clouds from high-resolution, three-dimensional CS and CALIPSO lidar products is first added to passive sensor data from MODIS by using the CUMULO framework and then transferred to coarse resolution data. This approach could offer a new perspective on statistical and process-oriented assessment of the performance of GCMs by being able to analyze the model output in terms of different cloud classes and thereby distinguishing the

driving mechanisms for the formation and evolution of different cloud types more clearly. This provides the potential to better understand and ultimately improve on existing model deficiencies.

The pixel-wise classification has a high accuracy of at least 0.8 for each class, with little variation across the validation splits. The relative amounts of predicted Cu and St, however, can differ by more than a factor of two compared to the CC-L data used as ground truth. While the performance metrics of the classifier leave room for improvement, the predicted cloud classes show distinct physical properties that are consistent with the expected properties of the corresponding WMO cloud classes, thus instilling some confidence in the validity of the predictions. The subsequent regression can reproduce consistent cloud class distributions on regional scales with mean errors being at least one order of magnitude smaller than the random baseline. Furthermore, the RFRM successfully generalizes to different data as could be shown using ESA-CCI. The predicted global distributions of the individual cloud classes compare well with the CC-L ground truth. This is evident when qualitatively comparing the distributions for each class as well as in the correlations and differences in areas of a high-class fraction. The correlation is larger than 0.6 for all but two classes (St/Dc) and the relative difference in areas of a high-class fraction is smaller than 50% for all classes. The spatio-temporal location of a sample is not used as an input for the ML algorithms. Therefore, any predictions are solely based on the physical properties represented by the features. Yet, even small-scale regional characteristics of the CC-L ground truth are similarly represented in the predictions using ESA-CCI. Notable examples are a peak of Ci in the tropical warm pool region or an increased As fraction in the Himalayas. Additionally, the geographical means for all classes correlate positively with the respective relative occurrence in the CC-L ground truth, with higher correlations for the classes with many available samples. We further showed that the regression model associates each class with specific feature values (see Section **4.3.3**). These values are consistent with the expected properties of the different WMO cloud types. Analysis of the effects of temporal averaging of the target data showed that the method works well with near-instantaneous data but cannot be applied to monthly averaged data. Tests with multiple sets of input features have shown that information about the cloud height, *cwp* and *cod* are essential for good performance, with information on the thermodynamic phase of the cloud providing additional robustness. In contrast, the horizontal resolution of the data the model is trained on seems to be less critical. Models trained on different grid cell sizes show differences but no clear optimal resolution can be defined from these initial results. It is recommended, however, that predictions be performed using data at their highest available horizontal resolution as more features can be resolved.

Predictions on test data show clear correlations of ground truth and predictions (average $\bar{c}_p = 0.92$). Predictions on ESA-CCI provide enough information to isolate individual features and processes. This suggests that this method can be successfully applied to any dataset of sufficient length and horizontal and temporal resolution to allow for statistically robust

predictions. When applied to the test data, the median relative deviation was about 50%. Comparing the predictions with the raw labels from CC-L we find similar values. Especially in regions where specific cloud types are predicted with a high frequency of occurrence, we find relative deviations mostly below 50%. Only the St class is consistently underestimated for which the pixel-wise classification already showed poor performance.

4.5. Discussion and Outlook

The considerable deviations in the amount of St predicted by the classifier are at least partly explainable by the relatively small amount of training samples and the similarity to the physical properties between St and Sc. In fact, the only metric by which St and Sc are distinguished in CC-L is their horizontal homogeneity, which, due to their small footprint, these instruments are not well suited to ascertain in the first place. Nevertheless, from the physical consistency of the predicted classes, we conclude that pixel-wise labeled data are suitable as a basis for training a regression model that learns cloud class distributions on datasets with a horizontal resolution typical for GCM scales. Generally, our results suggest that the method is, therefore, suitable for a process-oriented assessment of clouds simulated by GCM in the space of cloud classes. Using cloud classes obtained from a consistent and fairly objective source (CC-L, (Wang, 2019a)) instead of physical variables, removes a layer of subjective interpretation of the data. Because we are using cloud classes explicitly defined to be similar to the WMO cloud genera, the resulting cloud class distributions can then be analyzed and interpreted in terms of the key processes driving formation and evolution of the different cloud classes. This greatly simplifies process-oriented analysis and evaluation of clouds in GCMs. Secondly, as the deep learning algorithm learns from high-resolution 3-dimensional data, the GCMs are implicitly analyzed in a horizontally super-resolved manner which also takes into account information about the vertical structure of clouds, i.e. learning from a combination of 2- and 3-dimensional data can potentially take advantage of information from the vertical that would not be included in the cloud top view. That the method can resolve phenomena on regional and seasonal scales provides the opportunity to identify spatio-temporal regions in GCM output in which clouds are not correctly represented. This could for example be done for the low-level clouds found in the subtropics west of the continents, investigating their horizontal extent, their dependence on feature values, and their temporal evolution.
However, due to the nature of the multi-stage process, some limitations apply: by building the regression on 2-dimensional, spatially averaged source data it is hard to make correct predictions on individual grid cells. This results in several samples for which the predicted cloud-type fraction differs by a factor of two or more. Additionally, the under-representation of the Cu class and the limited accuracy of the St class show that this method can still be improved. This probably stems at least partly from the CS ground truth itself, as the CS algorithm has trouble distinguishing between St and Sc. Downstream analysis should therefore combine St

and Sc into a single "low-level stratiform" class. Also, some features of the predicted cloud distribution such as, for example, the high fractions of Ns along the Antarctic coast, might be amplified or hidden by noisy satellite retrievals. Especially in high latitudes, clouds can be challenging to characterize with passive sensors like MODIS and AVHRR.
Since our ML models are trained on instantaneous measurements, they do not provide satisfactory results when applied to temporally averaged data. Using geostationary data available e.g. every 30 minutes (Geostationary Operational Environmental Satellite (GOES) satellite, Walther et al. (2013)) for the pixel-wise classification instead of MODIS data available only twice a day might improve the results. Such an approach has been applied to other atmospheric variables like convection and rainfall (Gorooh et al., 2020; Lee et al., 2021). The physical properties of the predicted clouds could then be safely averaged over time due to the high and consistent temporal resolution of the data allowing the regression model to train on data more comparable to typical GCM output. However, the majority of the processes of interest here are not resolved at large temporal scales. This contributes to the poor RFRM performance for monthly mean data and will still be an issue when the RFRM is trained on temporally averaged data. In turn, this means that this method is suitable for detecting model deficiencies relatively quickly in contrast to using climatological means from long-term simulations. This is because we would expect an inaccurate representation of the global and regional cloud distributions to be already detectable with model output available for less than a year.

The consecutive application of two ML steps makes it difficult to quantitatively estimate the propagated errors. Even though the error of the classification and regression can be individually estimated using test splits, the combined impact of these errors is not clear. We do not, however, see any specific inconsistencies in the physical properties or the regional distributions of the predictions, suggesting that the propagated uncertainties are reasonable. We estimate an uncertainty of up to 50%, compared to what would be reported by CS for individual predictions. Due to the high correlation with the CC-L ground truth, however, we have high confidence in our method predicting the correct long-term statistics for most regions. The absence or underestimation of regional patterns in GCM cloud distribution would be a good indication of possible deficiencies. Even for classes for which limited training data are available (Cu, St, Dc), we find that the predictions are self-consistent: the characteristic feature values are distinct for each class and do not vary regionally. The regional distributions of the classes are attributable to the predominant atmospheric conditions. For example, the Dc class occurs more frequently near the equator, Cu predominantly in tropical and subtropical regions over the ocean and St mostly west of the continents in the subtropical subsidence regions. Both, Cu and St are predicted as low-level clouds with low cloud top heights, as is their WMO definition.

In terms of implementation, this method can be applied to new datasets quickly and does not require individual implementation for each model, in contrast to satellite simulators (Swales et al., 2018). While many of the variables needed for this method are typically part of the standard output of GCMs it would be important to provide instantaneous or near-instantaneous values

(i.e. model output not averaged over longer time intervals) for the important variable *cod*, *cer* and *ptop*. Chapter **5** show that, when these variables are all available, application of this framework to GCM is possible. The modeling community is therefore encouraged to provide such an output e.g. in future model intercomparison projects such as CMIP7. Adding *cod*, and *cerl*/*ceri* as instantaneous 2-dimensional variables to the CMIP7 data request would be a good step to make the presented framework more widely applicable.
Future improvement of the method could include replacing the RFRM as a regression model. The most significant advantage of the RFRM is the use of the bagging process during training, which helps to generalize well to unseen data. However, the size of the RFRM scaling with the size of the dataset while batched training is not straightforward demands a high degree of sub-sampling to make training computationally feasible. Therefore, as noted previously, this required us to disregard data. However, once trained, applying the RFRM to batched data for predictions is possible and fast, making it suitable for practical application. A CNN could be a reasonable replacement for the RFRM due to the image-like structure of the data. First attempts to replace the RFRM with a CNN, however, did not yield satisfactory results independent of specific architecture, with the network's loss not converging. Additionally, implementation, training and tuning of the RFRM are much simpler than that of a CNN, which makes the RFRM more suitable in practice. Also, in this study, the cloud classes in the CC-L source data are aggregated in the vertical dimension by assigning the most common class in the cloud column to the respective pixel. Even though this provides an implicit resolution of vertical features, a full classification in three dimensions would be a clear improvement. An improved representation of the vertical cloud structure might be obtained by using a more sophisticated aggregation algorithm. Taking into account the physical properties of the observed pixel in each vertical column might lead to more representative ground truth.

5. Characterizing Clouds with the CCClim Dataset, a Machine Learning Cloud Class Climatology

5.1. Overview

While observations of clouds continue to be taken at the surface or from aircraft, the global or near-global datasets that can be obtained from spaceborne remote sensing instruments play an important role in climate science. Satellite observations have been used to better understand the distribution of clouds on the global scale and to assess the quality of cloud representations in GCMs (Bodas-Salcedo et al., 2012; Ceppi et al., 2016; Lauer et al., 2023; Vignesh et al., 2020; Wall et al., 2022). However, uncertainties and differences among available satellite products make it difficult to interpret the data objectively and to evaluate GCM results (Dubovik et al., 2021; Evan et al., 2007; Zhang, 2005). Climate science would therefore significantly benefit from satellite products that are more accurate over large areas as well as easily comparable and interpretable. The previous Chapter **4** introduced a framework to make satellite observations of cloud-related data more objectively interpretable. This is achieved by obtaining cloud class distributions even at low resolutions. This chapter builds on the results of Chapter **4**, providing applications of the trained cloud-class prediction framework in terms of a novel cloud-class climatology and cloud-class predictions for a GCM. In creating a new dataset, these challenges are addressed by combining active and passive sensor data, providing a global and long-term time series of relative cloud-type amounts.

Text, tables and figures in this chapter are part of the first-author publication Kaps et al. (2023c) and have all been initially compiled by the author of this thesis, with co-authors contributing to the final submitted text. As this thesis is submitted, Kaps et al. (2023c) is still under review at "Earth System Science Data" and available as a preprint.

Current satellite products typically include physical variables retrieved from the measured radiation such as cloud top temperature, cloud water path or thermodynamic phase. These retrievals are, however, subject to sensor limitations. Passive sensors can typically only mea-

sure integrated properties of the complete atmospheric column and characterize the topmost cloud layer. For this reason, only limited insight may be gained into overlapping clouds with passive sensors. Optically thin clouds or clouds over snowy landscapes are also challenging for passive satellite instruments due to limited measurement contrast relative to the surface. In contrast, active instruments such as lidar and radar sensors can resolve certain properties vertically but provide small measurement footprints limited to a narrow swath, such that global coverage with short revisit periods is usually impossible. A common approach is, therefore, to combine measurements from multiple sensors to obtain a more complete assessment of the atmosphere (Haynes et al., 2011; Jiang et al., 2012; Stubenrauch et al., 2017; Wang et al., 2016). Additionally, measurements can be subject to calibration issues, which may introduce further deviation between individual instruments and/or simulations (Loeb et al., 2009). With improving machine learning (ML) capabilities, the opportunity arises to produce smart combinations of measurements from active and passive sensors to provide enhanced observational products benefiting from both types of sensors (Reichstein et al., 2019).

Complementing existing products, we present a new dataset for the climatology of cloud types, named CCClim (Kaps et al., 2023b), produced using ML algorithms trained with a combination of data from active and passive instruments. The cloud types contained in CCClim are physically consistent with most of the major cloud genera defined by the World Meteorological Organisation (WMO, 2023) making the dataset easier to compare to human observations than other definitions as used e.g. in Kurihana et al. (2022). A consistent, long-term cloud-type dataset like CCClim is required to empirically study the processes governing the interaction of clouds within the climate system, which in turn is a necessary component in improving current GCMs (Li et al., 2015).

Climatologies of the cloud type from remote sensing data have long been in use to categorize multi-dimensional large-scale cloud properties, like in the dataset created by the International Satellite Cloud Climatology Project (ISCCP) (Schiffer and Rossow, 1983). Now, with increasingly advanced measurements and computational capabilities such categorization has started incorporating other quantities beyond cloud top pressure (*ptop*) and optical depth (*cod*) as in Rossow and Schiffer (1999). More recent versions of the ISCCP dataset have been used in combination with ML methods to provide data categorized by cloud regime (e.g. Tselioudis et al., 2013; Tzallas et al., 2022; Young et al., 2018). These regimes are defined by unsupervised clustering algorithms on observational data of horizontal resolutions in the order of tens to hundreds of kilometers (Tselioudis et al., 2021). Using unsupervised methods is a popular strategy when labeled data are sparse or expensive to produce. However, clustering strategies, especially when based on low-resolution data, can struggle to produce meaningful cloud regimes, as such strategies are prone to compensating errors and uncertainties introduced by overlapping clouds (McDonald and Parsons, 2018). Creating manually cloud-labeled satellite data of sufficient quantity is possible, but comes with considerable cost and limitations (Rasp et al., 2020; Stevens et al., 2019). In contrast, Zantedeschi et al. (2019) (hereafter Z19) used sparse labels obtained from active instruments aboard CloudSat (Stephens et al., 2002) and the Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation (CALIPSO) (Winker

et al., 2003) to train a convolutional neural network (CNN) in a supervised framework to label images from the passive Moderate Resolution Imaging Spectroradiometer (MODIS) (Platnick et al., 2003; Platnick et al., 2017; Stephens et al., 2018). Another way of training a supervised model was employed by Kuma et al. (2023), who combined measurements from the Clouds and the Earth's Radiant Energy System (CERES) with cloud types observed from ground stations to create a labeled dataset. Using a supervised method with prescribed classes can arguably be restrictive, which is why some studies deliberately employ unsupervised methods to find more distinct classes. Kurihana et al. (2021) combined an autoencoder network with a clustering algorithm, finding twelve cloud-type clusters from 128×128 pixel patches of MODIS data. They used a refined version of this method to produce a dataset of 42 individual cloud-patch classes (Kurihana et al., 2022).
The CCClim dataset presented here has been created using a combination of supervised ML methods trained using passive as well as active satellite sensors, using a previously published framework (Kaps et al., 2023a), see also Chapter **4**.
We provide an overview of the datasets and methods used in training, application and evaluation in Section **5.2**. Section **5.3** deals with the contents of CCClim and shows examples of how it can be used to study clouds. The potential of using CCClim to evaluate GCMs is indicated in Section **5.4**. Finally, we discuss the scope of the dataset's capabilities in Section **5.5**.

5.2. Data and Methods

We trained two ML models that are applied consecutively to predict each cloud type's RFO in low-resolution grid cells similar to the resolution of most current GCMs. The reasoning behind our approach is explained in Section **5.2.3**. The first stage is trained on the CUMULO dataset (year 2008) created by Z19. CUMULO contains physical variables (see Table **5.1**) obtained from the MODIS Cloud Product MYD06 dataset from the Aqua satellite, which we use as input features (Platnick et al., 2003; Platnick et al., 2017). As Aqua is part of the A-Train constellation, its measurements can be aligned with measurements from other A-Train satellites, such as CloudSat and CALIPSO (Stephens et al., 2018). The target labels are cloud-type labels from CloudSat's CC-L dataset (Wang, 2019a), which are defined according to WMO cloud genera. In CUMULO, coinciding measurements from MODIS and CC-L are aligned at the pixel level. The second stage is a regression model trained on coarse-grained output from the first stage. Finally, CCClim is produced by applying the trained regression model to the daily product (L3U) of ESA-CCI (Stengel et al., 2019, 2020).

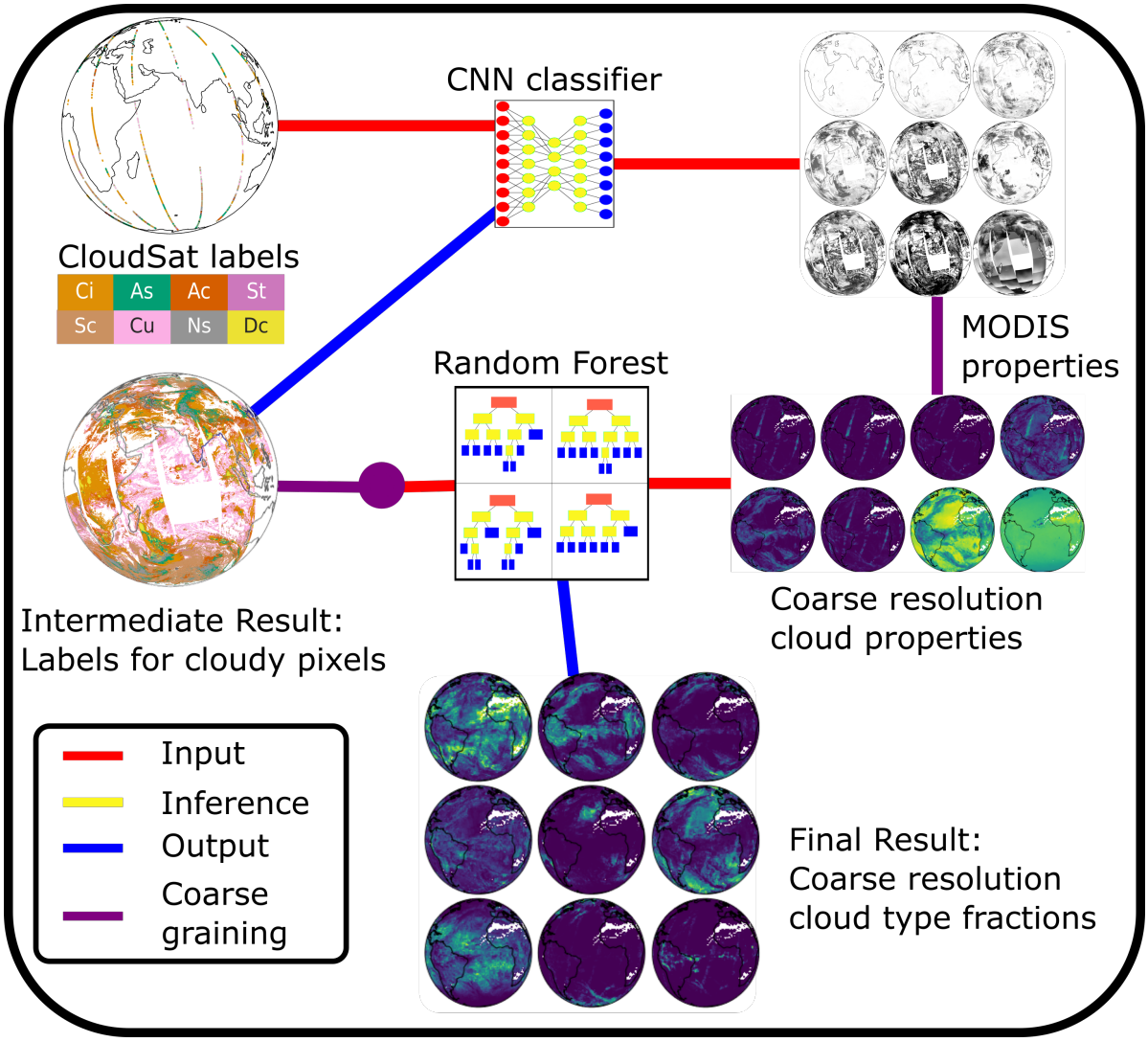


Figure 5.1.: Schematic of the training of the two machine learning models. The second stage is trained on coarse-grained output from the first stage. The trained RFRM is then applied to ESA-CCI to obtain CCClim. From Kaps et al. (2023c).

5.2.1. Data

The main three datasets relevant to this chapter - ESA-CCI, CC-L and MODIS - are described in Chapter **3**. Also, we use data from the ECMWF fifth-generation reanalysis ERA5 (Hersbach et al., 2020) to assess the plausibility of the derived cloud types in CCClim. Specifically, we use monthly means of the vertical velocity at 500 hPa (ω_{500}) and the sea-surface temperature *SST* as proxies for the dynamical and local thermodynamic conditions (Bony et al., 2004). These data are available for the complete period covered by CCClim, such that we can spatiotemporally co-locate them.

5.2.2. Method

The first of our two ML models is a pixel-wise classifier based on the Invertible Residual Network framework also used by Z19 (Behrmann et al., 2019). However, we use the retrieved physical quantities instead of radiances as inputs to our network for better interpretability. The pixel-wise CNN classifier trained to predict CC-L cloud type labels from MODIS observations is used to obtain a label for each cloudy 1×1 km^2 pixel in the MODIS data. The classifier achieves good accuracy for most classes and deals well with the significant class imbalance, except for the classes stratus (St) and stratocumulus (Sc). Both of these cloud types can occur under similar conditions, while Sc is much more frequent than St.
For the second stage, the CUMULO dataset, now fully labeled by our classifier, is coarse-grained to a horizontal resolution of 100×100 MODIS pixels, comparable with the size of a grid cell of a typical GCM. The coarse-graining entails averaging the input features and converting the pixel labels into RFOs for each cloud type, i.e. the relative amounts of each class per cell. Pixels without a label are treated as undetermined, such that clear-sky is also included in the "undetermined" class. As some features such as *ptop* are only defined if a cloud is present, we average these only over pixels with a cloud label. We then apply a Random Forest (RF (Breiman, 2001), which is used as a regression model to predict the RFO of each of the nine classes (see Table **3.2**). In Chapter **4**, the physical consistency of the predicted RFOs was validated by using 2.5 years of the independent ESA-CCI dataset as input to the RF, which showed good agreement with the cloud type distribution in CC-L.
The final product of this framework is the RF model, which can predict relative cloud-type amounts using low-resolution cloud properties as inputs. The details regarding RF training and performance evaluation are the subject of Chapter **4** (see also Kaps et al. (2023a) or the respective code via `DOI:10.5281/zenodo.7248773`). We created CCClim by using this RFRM model to predict cloud-type RFOs for the ESA-CCI dataset. cloud type predictions are made for grid boxes of 10×10 cells, i.e. $0.5° \times 0.5°$ and the output is averaged to daily values on a 1°-grid. Note that the RF model is not applied to the same grid resolution it was trained on. This is possible since the method depends little on horizontal resolution in this range, as shown in Table **4.2** and Section **4.3.5**.
The RFOs are normalized by the sum of the eight cloud types, i.e. they are independent of the total cloud amount. The "undetermined" class consists largely of clear-sky but also pixels for which one or more retrievals failed for other reasons. This class is therefore not a suitable indicator of the total cloud amount. Furthermore, cloud-free grid cells are explicitly excluded from our analysis, as the RF could not process them. The following samples were excluded from CCClim because of faulty retrievals:

- All of July 2010 because of faulty surface temperature retrievals
- Two days at the turn of the year 1985/1986

- Retrievals at the end of 1994 were faulty due to the orbital drift of the NOAA-11 satellite, leading to the removal of the last 115 days of 1994 (November and December are already not included in ESA-CCI)

- Extreme outliers and fully cloud-free cells, detected by $tsurf < 10$ K or $ptop < 10$ hPa (amounts to $\sim 11\%$ of data, ~ 727 Mio. grid cells)

5.2.3. Concept Rationale

The framework used to produce CCClim has been developed primarily to facilitate GCM evaluation: it is meant to produce meaningful and self-consistent cloud-type amounts on horizontal and temporal scales similar to the output of current GCMs. The consistency of the derived cloud-type RFOs was validated against their related physical variables and those of the classes in the CC-L dataset. This showed that the cloud types in CCClim are consistent with the classes obtained with the CloudSat algorithm (Kaps et al., 2023a).
Since RFOs are obtained for low-resolution grid cells instead of using a classifier for high-resolution image pixels, details are inadvertently smoothed out. Conversely, this approach compensates for having multiple cloud types in a single pixel, e.g. via overlapping, a cause of ambiguity in a 2D analysis. Neighboring pixels can reinforce information on the dominating overlapping cloud type or provide complementary single-layer information. The coarser approach, therefore, leads to predictions that are more relevant for large-scale systems, while implicitly retaining information from the underlying high-resolution 3D observations.

Training on MODIS data and application to ESA-CCI data might at first glance seem like an unnecessary error source. However, training the classifier on ESA-CCI data is far more difficult, as the satellites providing the AVHRR data used in ESA-CCI are not part of the A-Train constellation (L'Ecuyer and Jiang, 2010). In contrast to CC-L and MODIS (onboard Aqua), easy co-location of the CloudSat/CALIPSO and AVHRR data is therefore not possible. Furthermore, MODIS measurements are only available for the shorter measurement period ($\sim$ 20 years), which is significantly less than the 35 years covered by AVHRR data. Furthermore, omitting the second ML step and using the pixel-level results directly would increase storage requirements, not enhance comparability with GCM output and also be subject to the previously mentioned issues with overlapping or ambiguous clouds.

Abbreviation	Variable	Unit	Classif. Input	Regr. Input
cwp	total cloud water path (ice + liquid)	$\frac{g}{m^2}$	✓	✓
lwp	liquid water path	$\frac{g}{m^2}$		✓
iwp	ice water path	$\frac{g}{m^2}$		✓
cph	cloud top phase	categorical	✓	
cer	effective cloud particle radius	µm	✓	
cerl	eff. radius liquid cloud droplets	µm		✓
ceri	eff. radius cloud ice particles	µm		✓
cod	cloud optical depth	1	✓	✓
ptop	cloud top pressure	Pa	✓	✓
htop	cloud top height	m	✓	
ttop	cloud top temperature	K	✓	
tsurf	surface temperature	K	✓	✓
ceff	cloud effective emissivity	1	✓	

Table 5.1.: Retrieved physical input variables for the ML methods. The classification inputs correspond to the variables available from the MODIS Cloud Product. The regression inputs are computed from the ESA-CCI and MODIS data. The published CCClim dataset contains the corresponding ESA-CCI regression inputs for each cell. From Kaps et al. (2023c).

5.3. Structure and Features

Each year of the dataset is saved to a separate `netcdf` file. Each CCClim sample contains the daily mean fractional amounts of the nine classes and the eight input features (Table **5.1**), as well as *clt* from ESA-CCI. The samples are identified by date and location (latitude-longitude) on a 1° regular grid. The total amounts of each cloud type in CCClim are shown in Fig. **5.2**(a). Additionally, CCClim contains on average 22.6% of the "undetermined" class, i.e. at least 77.4% cloud, which is roughly 10% and 7% more than reported by ISCCP and CC-L, respectively (L'Ecuyer et al., 2019; Young et al., 2018). Some of this overestimation is due to the removal of cells that are (almost) totally comprised of clear-sky ($\sim$ 11% of the data). The inclusion of *clt* in CCClim enables easier assessment of the impact of this filtering. Taking into account the filtering, the CCClim contains an average $clt \approx 68\%$ due to the low base cloud amount in ESA-CCI of $\sim$ 64%, resulting in a significant overestimation of cloud cover in CCClim compared to ESA-CCI (discussed in Section **5.5**).
The classes are subject to a large class imbalance with the most common class (Sc) occurring more than 27 times more often in the global mean than the least common class (Dc). This class imbalance is largely learned from the CC-L data (see Fig. **5.2**(b)). The regional distribution is shown in Fig. **5.3**. All cloud types show distinct and physically plausible regions of high occurrence. Since the color scale was chosen to show that Sc is the most common type almost all over the globe, regions with an increased probability of forming a specific cloud type are not easily visible. We solve this by showing in Fig. **5.6** the type with the highest value of $\frac{RFO_{cell}}{RFO_{mean}}$ in each grid cell, i.e. the highest cloud type fraction relative to its respective global

mean value. This highlights for example the regions of increased St or Dc occurrence, even though these are the least common types in CCClim.

As another example for analysis on climatological timescales, we show the full time series for all eight cloud types averaged over the southern hemispheric oceans (defined as all ocean grid cells in the latitude belt from 0°S-90°S) is shown in Fig. **5.4**. All eight cloud types show a distinct seasonal cycle. The spatial variability given by the shading in Fig. **5.4** does not change noticeably over the years. Figure **5.5** shows the average seasonal cycles calculated as the mean of each calendar day averaged over the full 35-year period. Comparing the mean seasonal cycle to the seasonal cycle from individual years shows that the relative deviation from this average cycle is typically smaller than 20% for a given calendar week.

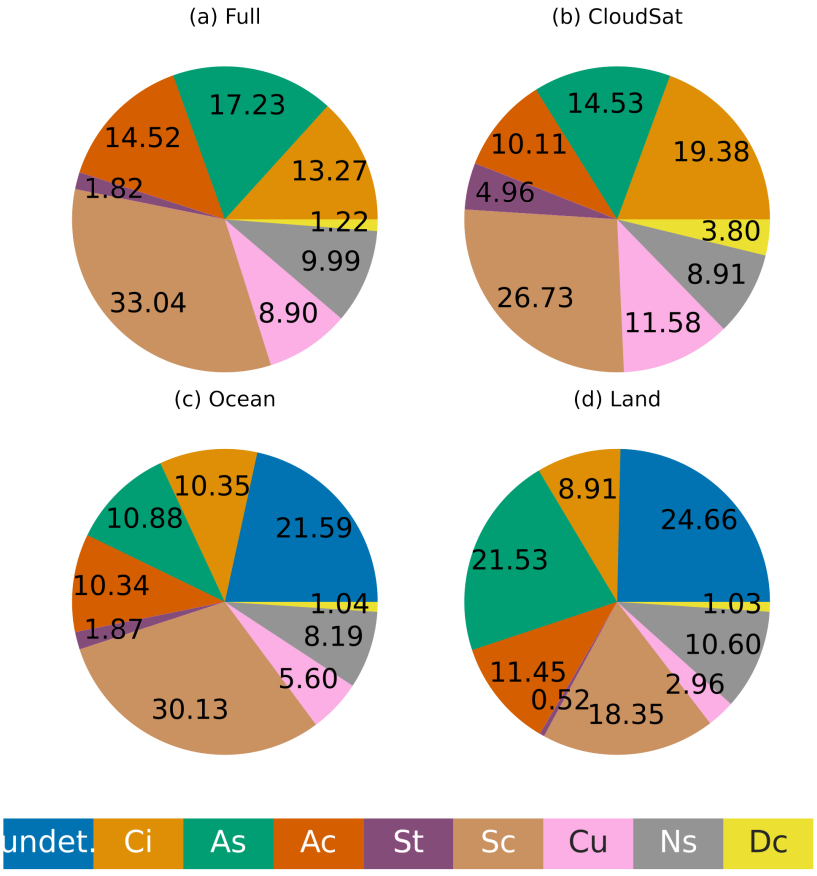


Figure 5.2.: Relative occurrence of each cloud type in (a) CCClim and (b) CC-L for the year 2008 if clouds are present. CCClim results for (c) ocean-only and (d) land-only grid cells are shown in the lower row. The "undetermined" class includes clear-sky and is not included in (a) and (b) for comparability reasons. From Kaps et al. (2023c).

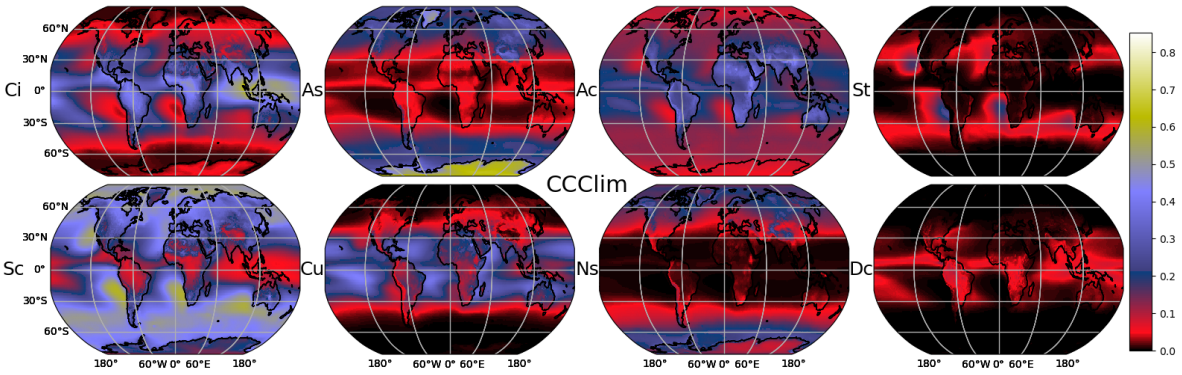


Figure 5.3.: Average geographical distribution of the RFOs for all cloud types in CCClim. From Kaps et al. (2023c).

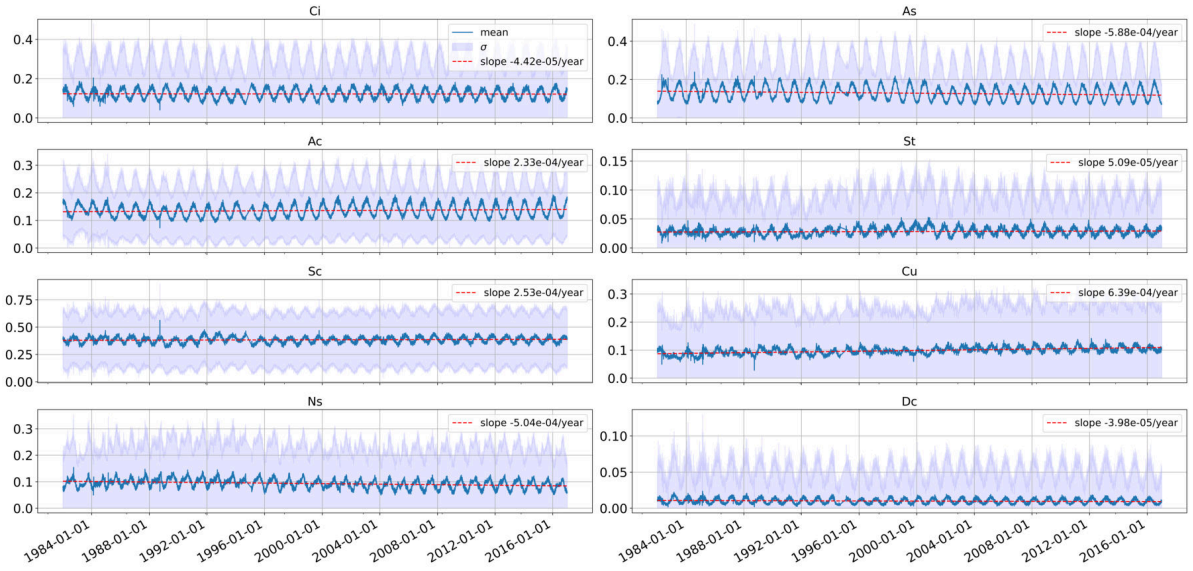


Figure 5.4.: Time series of daily mean RFO with the spatial standard deviation shown as shading for all cloud types averaged over the ocean in the southern hemisphere. All types show a consistent seasonal cycle and little anomalies and drift, as shown by the slope of the linear fit over the full period. Grid cells with a maximum RFO close to zero (1% quantile) are filtered out. From Kaps et al. (2023c).

5.3.1. CCClim classes

While an exhaustive analysis of the classes and related processes is beyond the scope of this paper, we highlight some characteristics of each cloud type in CCClim.

The Sc type is the dominating class in most regions, with a median global RFO of 0.31. Relative to its total amount, Sc shows little seasonal variability. Sc is subject to confusion with the less common St type at all stages of the cloud classification: the CC-L dataset used as ground truth has trouble distinguishing between St and Sc due to the small footprint of the active sensors. The pixel-wise classifier and RF model propagate this uncertainty to CCClim, where St and Sc are predicted for similar conditions. Ci, the fourth most common class, appears

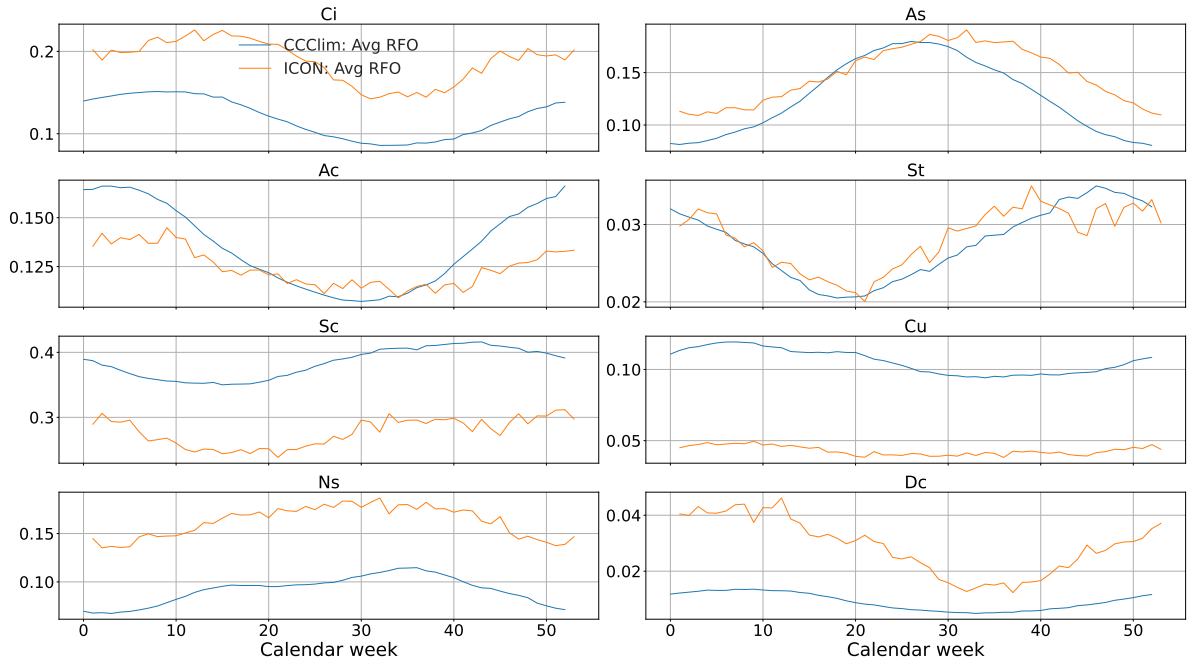


Figure 5.5.: Example comparison of the mean cloud type RFOs by calendar week over the southern hemispheric oceans between CCClim (blue) averaged over the full 35-year period and an ICON-A simulation (orange, see Section **5.4**) averaged over two years. From Kaps et al. (2023c).

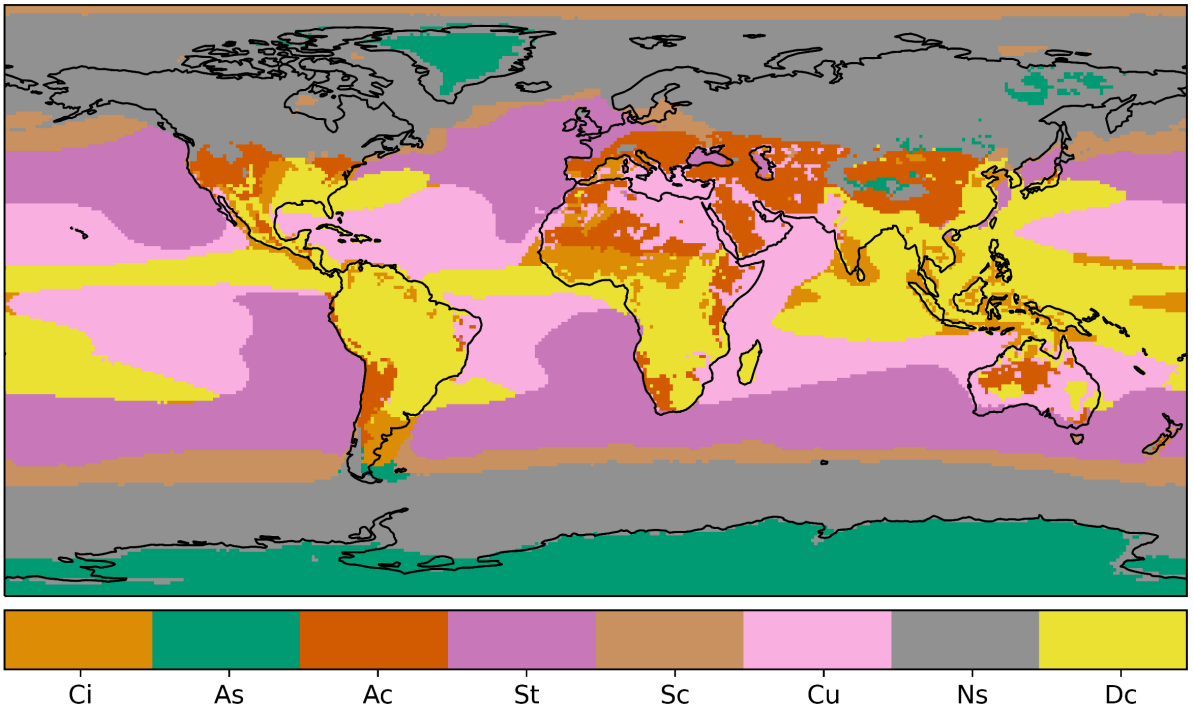


Figure 5.6.: Cloud type with the highest RFO relative to its global mean value. This shows which regions have the most favorable conditions for the occurrence of a cloud type, independent of its total amount. From Kaps et al. (2023c).

most frequently in the tropics and subtropics, peaking in Southeast Asia. This geographical distribution is in line with expectations from previous studies (e.g. Sassen et al., 2008). The As type is prevalent in middle to high latitudes, especially over high-latitude land masses, with a high correlation with Ns. Although the CCClim As distribution is similar to what is reported in literature (e.g. Sassen and Wang, 2011), results in polar latitudes might be unreliable due to limitations of passive sensor retrievals in these regions. The amount of As is strongly modulated by the seasons in both hemispheres, peaking in winter. In contrast, the Ac amounts peak in summer, with a slightly larger seasonal amplitude in the Northern Hemisphere. For cumulus (Cu), the amplitude of the seasonal variation is smaller than for most other cloud types. Interestingly, the Cu amount increases over the 35 years covered by CCClim, with its mean RFO over the SH ocean increasing by $\sim 0.02 \approx 20\%$. Dc is the least common cloud type in CCClim because CC-L distinguishes sharply between deep convective and multilayered cloud systems (L'Ecuyer et al., 2019). Even though Dc is rarely the dominant cloud type in a coarse grid cell, Fig. **5.6** shows that there are distinct regions in which deep convection can occur more frequently (relative to the global mean RFO of this type) such as the intertropical convergence zone (ITCZ) over Southeast Asia or tropical landmasses. Dc is subject to significant seasonal variations with values in summer about three times larger than in winter.

5.3.2. Process-based Approaches

As an example for analyzing the impact of certain cloud types on climate, Fig. **5.7** shows the joint distribution of the short- and LW cloud-radiative effect (CRE) for each cloud type over the ice-free oceans. The cloud radiative effects are calculated from the top of the atmosphere (TOA) radiative fluxes provided as part of the ESA-CCI dataset by calculating the differences between the clear-sky estimate and the corresponding all-sky value for short and LW fluxes, respectively. For this analysis (Fig. **5.7**), only pixels with a sea surface temperature $SST > 275$ K are taken into account to reduce spurious effects introduced by sea ice. Also, only cells with a cloud type RFO larger than the 84% percentile are considered to minimize "contamination" with other cloud types. Note that despite this pre-selection of pixels every sample contains multiple cloud types and thus derived values cannot be interpreted as absolute for the "pure" cloud type. We would like to note this is the case for every sample in CCClim, i.e. in an approach like this, where cells with a relatively high amount of a cloud type are analyzed, other types can still significantly influence the values of the physical variables. For Fig. **5.7**, using a percentile instead of an absolute threshold means that for types with globally low occurrence (St, Cu, Dc), contamination will be more of a factor.
To illustrate this we will use the example of the Cu type, which displays one of the lowest LW CRE in Fig. **5.7**. Cu is strongly correlated with the "undetermined" class (not shown), which contains many clear-sky cases, partly explaining the small CRE values. Consequently,

the CRE shown in Fig. **5.7** is not that of a cell filled only with Cu, but rather of cloud regimes associated with high amounts of Cu. Since these regimes tend also to contain clear-sky or low-top clouds other than Cu, the resulting CRE shown here is small. Another low-top cloud type, Sc, displays a larger CRE both in the short- and LW spectrum. Again, these are influenced by co-occurring clouds resulting in the CRE of Sc varying with latitude. While Sc and St co-occuring in the Tropics and subtropics have a small LW CRE, the correlation of Sc with Ac leads to a stronger LW CRE in higher latitudes in the Sc-dominated cases. The stronger SW CRE of Sc in comparison to Cu can be attributed to Sc decks being more horizontally dense than the more cellular nature of shallow Cu. These relationships between the cloud types and their physical properties, as displayed with CCClim, might at first glance seem counter-intuitive. However, the interdependence of the cloud-type RFOs and their physical properties can also be beneficial. CCClim provides the context required to not only disambiguate the otherwise counter-intuitive relationships but also to highlight important cloud interactions. In Section **5.4** we exploit this interdependence to better characterize the cloud types (see Fig. **5.10**).

Mid-level clouds (Ac, As, Ns) display a more complex relationship between long- and SW CREs. Ac mainly differs from Cu and St through a larger LW CRE, which is due to the higher and thus colder cloud tops. Ns and As, however, are distributed along an almost constant LW CRE of approximately 32 ± 5 Wm^{-2}, with a peak at very small SW CRE values. An analysis of the associated cloud properties reveals that the clouds responsible for the weak SW CRE are located particularly at higher latitudes where As, Ns and Sc dominate cloud cover. Here, CCClim typically displays *iwp*>*lwp*, with small *cod*. This suggests optically relatively thin mixed-phase or ice clouds, which is qualitatively consistent with the observed CRE. The strong LW CREs of the high-top clouds Dc and Ci are in line with expectations and very similar due to the high geographical correlation of the two types.

Another example for a process-based analysis of the CCClim data employs additionally SST and ω_{500} from co-located ERA5 reanalysis data (Hersbach et al., 2020). ω_{500} is used as a proxy for the dynamical regime (large-scale circulation) and SST as a proxy for the local thermodynamic conditions (Bony et al., 2004). The results are shown in Fig. **5.8**. The maximum occurrence of a cloud type in this phase space quantifies the conditions favorable for the formation of this type: while it is known that Sc occurs frequently in the large-scale subsidence regions over the subtropical oceans, we can narrow down the dynamical regime of maximum Sc occurrence to $SST^{Sc} \approx 299$ K and $\omega_{500}^{Sc} \approx 0.008$ $\frac{\mathrm{Pa}}{\mathrm{s}}$. The mid-latitude regions (30° to 60°) also show a significant amount of Sc for ascending motions over cold ocean surfaces ($SST^{Sc} \approx 275.4$ K, $\omega_{500}^{Sc} \approx -0.024$ $\frac{\mathrm{Pa}}{\mathrm{s}}$). This Sc distribution is consistent with ISCCP data (Young et al., 2018), but the dynamical regime is not typical for Sc, such that mixed cloud regimes (Sc + As/Ns) can be inferred for this region. In contrast, Dc clouds are prevalent particularly in tropical regions with $SST^{Dc} \approx 302$ K and display large ascending motions with $\omega_{500}^{Dc} \approx -0.056$ $\frac{\mathrm{Pa}}{\mathrm{s}}$. Cirrus clouds often appear adjacent to deep convection in this phase

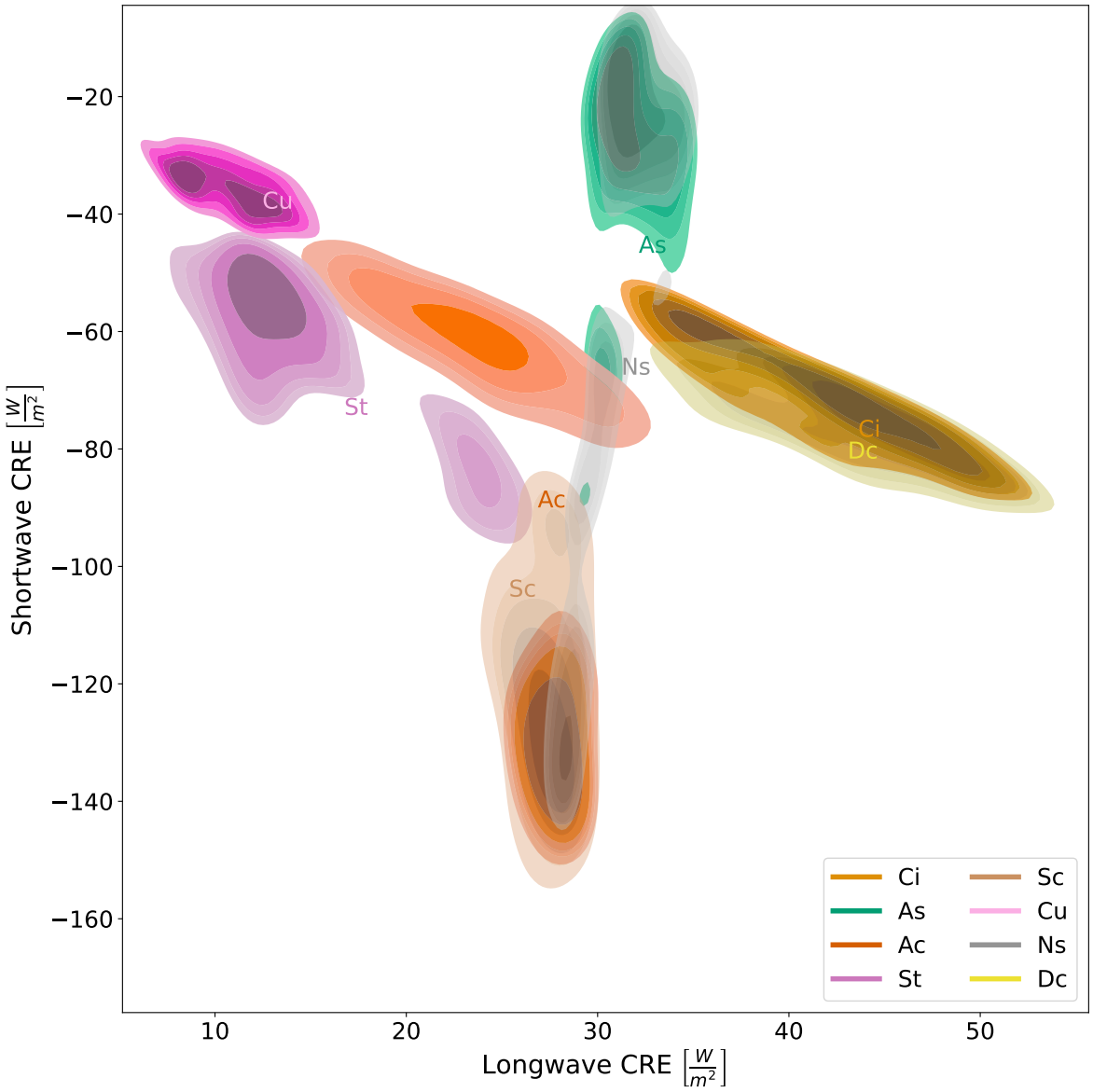


Figure 5.7.: LW and SW cloud radiative effect (CRE) of the cloud types averaged by calendar month as kernel density estimates. The outer-most density level contains 30% of the probability mass of the samples per cloud type. Since each sample contains fractional amounts of multiple cloud types, the CREs are influenced by other clouds in the same cell. This influence is stronger if the type appears less often overall. We only include pixels over ocean with a sea surface temperature above 275 K to reduce possible spurious effects of sea ice. Due to filtering and spatial/temporal averaging, roughly $333,000$ samples are available, of which for each cloud type the $5 \cdot 10^4$ largest are sampled, corresponding roughly to the 84% percentile. The cloud type abbreviations placed in in the plot denote the median CRE values for this type. From Kaps et al. (2023c).

space, presumably representing anvils or remnants of strong convection, as they show similarly warm sea surface temperatures ($SST^{Ci} \approx 301$ K) but rather descending air associated with the outflow regions of deep convective cells ($\omega_{500}^{Ci} \approx 0.004$ $\frac{\text{Pa}}{\text{s}}$). Ac clouds show a distribution

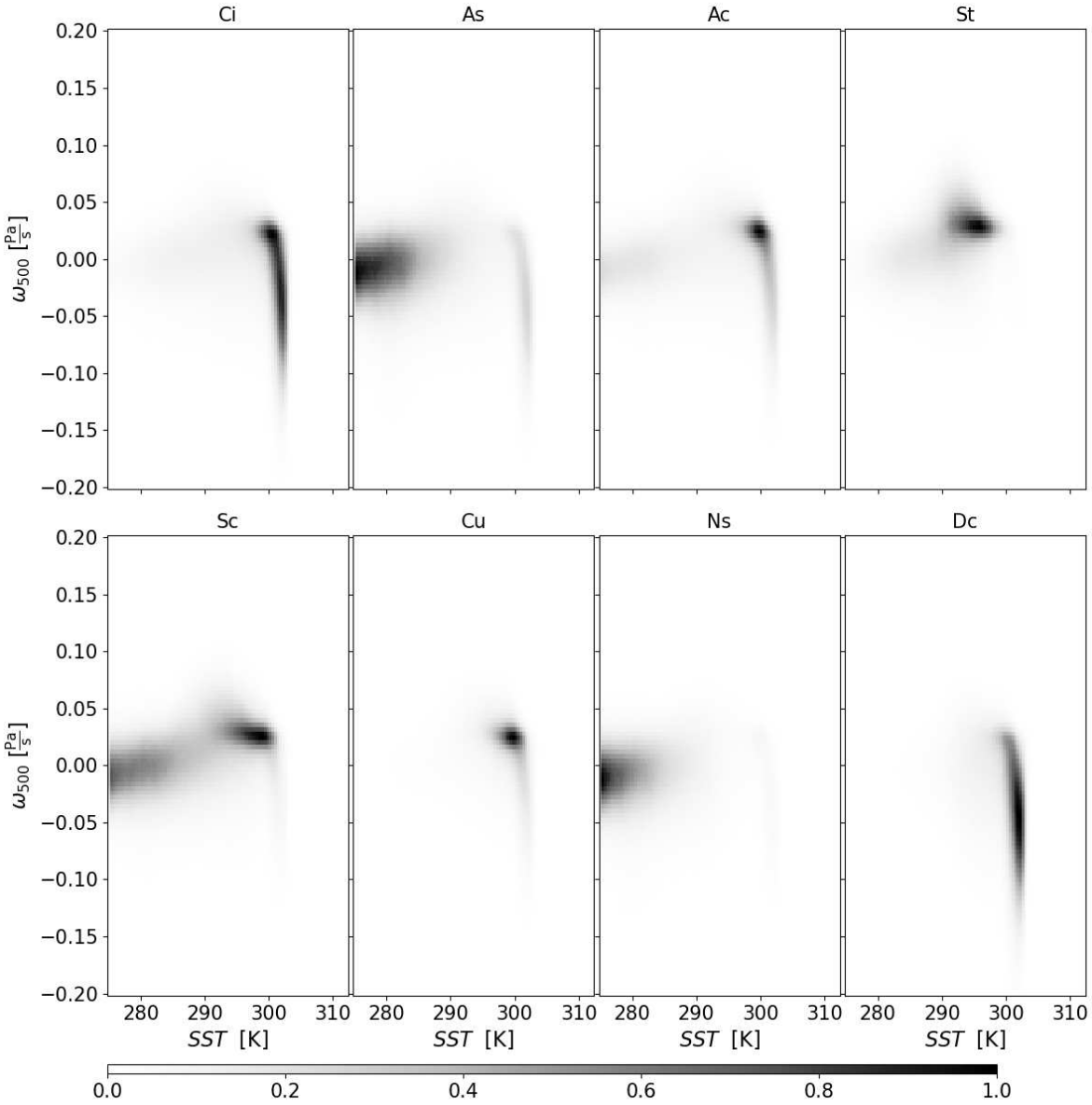


Figure 5.8.: Distribution of each CCClim cloud type in SST - ω_{500} space. For each grid point and type the RFOs are summed up and scaled by the maximum of this sum. The resulting distribution is therefore independent of the global amount of the respective type. These distributions can be used to infer the dynamical regimes favorable for the occurrence of each cloud type. As in Fig. **5.7**, samples are averaged by calendar month with cells with $SST < 275$ K excluded. The largest $5 \cdot 10^7$ samples are used in terms of each cloud type RFO. From Kaps et al. (2023c).

similar to Dc and Ci in the SST^{Ac} - ω_{500}^{Ac} space, corresponding to the Ac known to develop in the Tropics from air detraining from deep convective systems.

5.4. Evaluation of Global Climate Models

Comparison with CCClim can be a new avenue to evaluate clouds in GCMs. To compare GCM data to CCClim, cloud class distributions can be obtained from the GCM output using the same trained RF (Fig. **5.1**) used to create CCClim from ESA-CCI. As an example, we applied the RF to output from the atmosphere component of the icosahedral nonhydrostatic model (ICON-A, version 2.6.1, untuned) (Giorgetta et al., 2018), see also Section **3.3**. Figure **5.9** shows the geographical distribution of the relative fraction of each cloud type averaged over the 2-year ICON-A simulation compared with CCClim. Geographically, the cloud type distributions in both datasets are largely in agreement, but it is evident that low-top clouds (Sc, Cu, St) are underestimated in the simulation, in favor of Ci, Ns and Dc. This is consistent with the known underestimation of low-level marine clouds in ICON-A and many other GCMs (Crueger et al., 2018). The deviations will partly be due to erroneous cloud representation in ICON and partly due to inherent differences between model and satellite data. The contributions of either have yet to be determined. No conclusions should therefore be drawn from our analysis regarding the general performance of ICON-A, especially since we are using an untuned version. We can however provide an example of how to use CCClim for GCM evaluation:
A high cirrus fraction is usually associated with a strong long-wave warming effect. However, both aspects of the CRE have been reported to be smaller in ICON than in observations (Crueger et al., 2018; Gettelman et al., 2020). We also observe smaller CREs in the output of this specific simulation. With respect to CCClim the simulation also exhibits an increase in higher cloud RFO (Ci, Ns) and a decrease in Sc and Cu. Figure **5.9** does show an increase in deep convection and we noted an almost binary distribution of either a very high or very low Dc fraction per cell (not shown). This could indicate that convection in the simulation only develops under specific conditions, but is strong and stable if it does get started. The fact that we have a higher average cloud top means that non-convective processes play a significant role in forming high clouds here. The resulting Ci clouds are thin as most cells contained $> 90\%$ Ci have an ice water path $iwp < 10\frac{g}{m^2}$ (Fig. **5.10**), corresponding to a cirrus-attributable optical depth of $cod < 0.1$ (Heymsfield et al., 2003). Similarly, an overestimation of high, thin clouds has been found in other GCMs (Kodama et al., 2012). The Ci-property analysis resolves the apparent contradiction with the decreased LW CRE, as subvisible Ci have a negligible radiative impact (Spreitzer et al., 2017; Turbeville et al., 2022). The CRE analysis performed for CCClim can be equivalently performed for the cloud-type distributions in ICON-A (Fig. **5.11**), which support the conclusion of a high frequency of optically very thin Ci. Also, Fig. **5.11** emphasizes the strong reduction of SW CRE attributable to the reduction of low-top cloud amount (Sc, St, Cu) in ICON-A. Also, Fig. **5.11** only shows one peak in the distributions of Ns and As, while Fig. **5.7** shows most of the density at low SW CRE, but a significant fraction of Ns and As are associated with a strong SW CRE in CCClim. Dc clouds are the only type for which the SW CRE is not decreased, but in fact, increased compared to CCClim. Since the fraction of Dc in ICON-A is more than double the value form

CCClim, and the grid cells contributing to the densities therefore on average contain more Dc clouds, this higher CRE might be largely attributable to this increase. Using the same logic and accounting for the relative increase of Ns and As, their smaller SW CRE is possibly related to incorrect radiative transfer for ice clouds or more generally an incorrect ice particle size distribution. Conditioning cloud property distributions on cloud types like in Fig. **5.10** can also facilitate the analysis of distinct cloud processes as similarly demonstrated with the cloud regime/weather states methodology (Oreopoulos et al., 2016, 2014; Tselioudis et al., 2013). Using the conditional distributions, we find that in ICON-A both, Ci and Sc, tend to be much thinner than in CCClim, indicated by the ice/liquid water path values, respectively. Furthermore, even in cells containing mostly Sc, Ac is often present. This is not evident from the regional distributions (Fig. **5.3**), which show that the Ac amount decreases over the oceans. We can deduce that this mixture of Ac and Sc occurs at the interface of more convective regions near the equator and regions of large-scale subsidence in the subtropics.

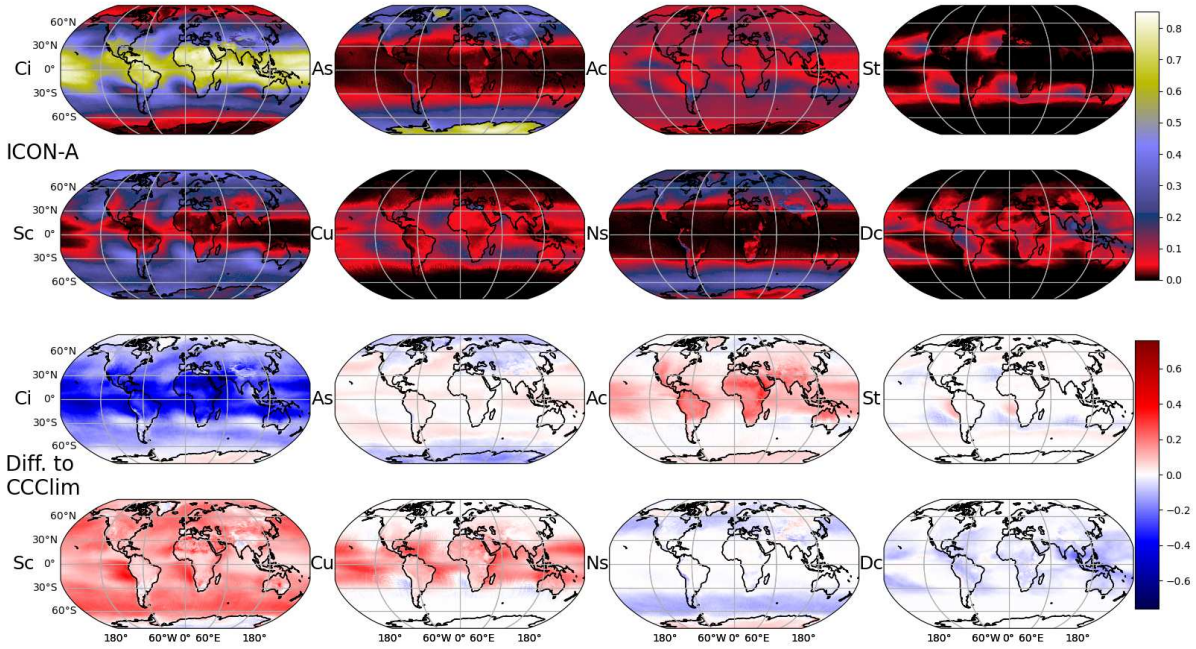


Figure 5.9.: Average cloud type distributions obtained with our RF model from two years of ICON-A output (top rows) and the differences to the CCClim distributions (bottom rows), where positive (red) values denote a higher value in CCClim and negative (blue) a higher fraction in ICON-A. Note that these values show the fractions of existing clouds, irrespective of cloud cover, thus the high fraction of Ci over northern Africa in ICON, which cause the extreme deviations at the end of the color scale. From Kaps et al. (2023c).

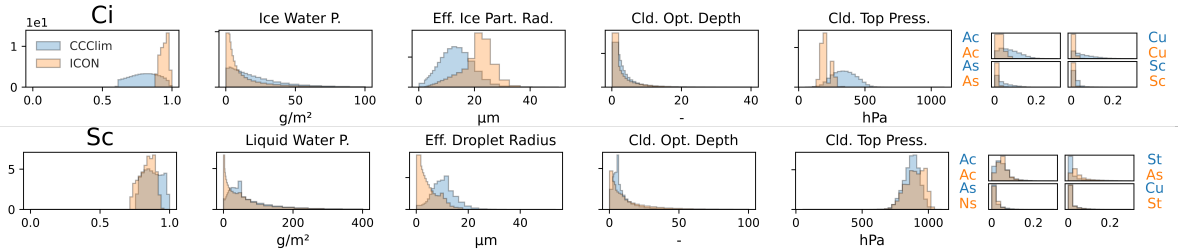


Figure 5.10.: Analysis of cells "characteristic" for Ci/Sc in ICON and CCClim in terms of probability densities. Characteristic cells are defined to have at least 85% of the chosen cloud type (Ci or Sc) and/or the "undetermined" class and a higher cloud RFO than the respective global median RFO. The resulting distribution of the class is shown in the leftmost panel. The four middle panels show cloud-relevant properties in the characteristic cells. The rightmost panel shows the distribution of the four cloud types that coincide most often with Ci/Sc in the respective dataset. This allows characterization of the physical properties of the cloud types while being able to take into account the contributions of other classes in the cells. From Kaps et al. (2023c).

5.5. Capabilities and Limitations of CCClim

The CCClim dataset enables investigation of clouds by WMO-like cloud type with a long coverage period and high spatial resolution as daily samples. Furthermore, using multiple cloud properties to define the classes makes the cloud-type predictions physically very consistent and reliable. We showed that categorizing complex atmospheric data into types defined similarly to established WMO types is more expressive and interpretable than the cloud properties by themselves. This was made evident by an example-analysis of the properties of different cloud types for given atmospheric conditions, increasing insight into important processes driving cloud development. As an example of how to evaluate GCMs with CCClim, we showed a comparison with cloud types obtained using the same method from the ICON-A model.

Our results show that the cloud types in CCClim have consistent seasonal variations, sensible regional distributions and little drift over the complete period. This makes CCClim suitable for statistical analyses of clouds and enables quantification of seasonal cycles of WMO-like cloud types on a global scale. As the cloud types have been learned from the CC-L dataset, its errors have been propagated. This applies to the distinction between St and Sc, such that the CNN classified many St clouds as Sc, leading to a small fraction of St in CCClim. For some applications, it might therefore be better to combine St and Sc into one new class. Furthermore, since CCClim contains a very low absolute amount of the Dc type, we recommend focusing on relative changes when studying this cloud type, like in Fig. **5.6**. With this approach, the prevalence of deep convective clouds in, for example, the ITCZ becomes more apparent.
Comparison with other cloud-type statistics has shown that our method likely underestimates the amount of cirrus clouds. This underestimation already occurs in the pixel-wise classification step, even though plenty of samples with the Ci label are available in the training data. The

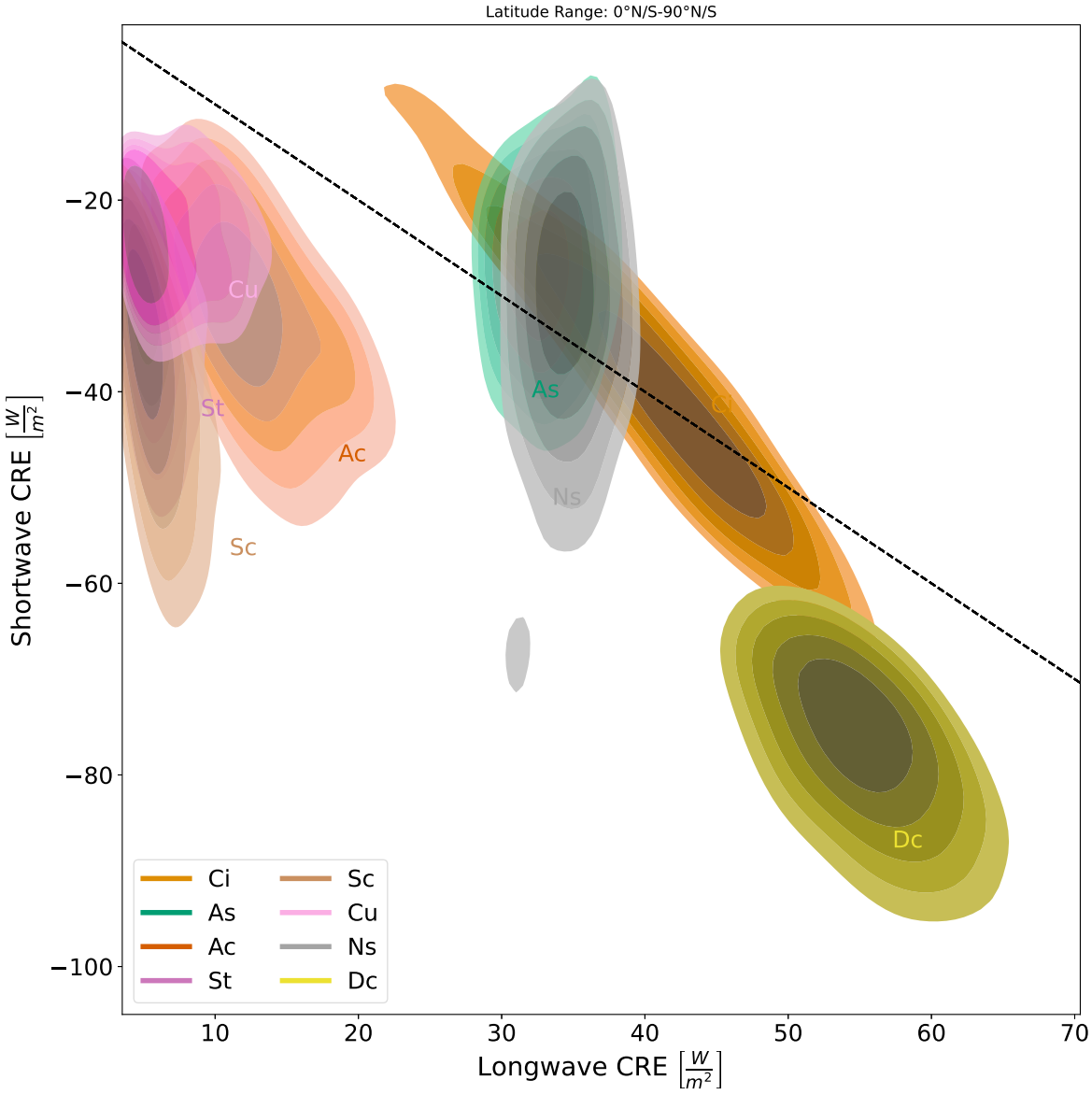


Figure 5.11.: Like Fig. **5.7**, but for the cloud types predicted for the ICON-A output. The dashed line indicates where LW and SW CRE cancel each other out. The most striking difference to Fig. **5.7** is the mean decrease in SW CRE, which is largely caused by both CREs attributable to Sc being much smaller. For Ci, the median LW CRE (as indicated by the "Ci" in the figure) is similar to CCClim but the possible range extends to much smaller values. This indicates a high occurrence of very thin Ci clouds here, which is also supported by the simultaneously very small SW CRE.

underestimation is therefore not caused by the difficulty of passive sensors to detect thin cirrus clouds, but rather by how the retrievals deal with multi-layer clouds. Wang et al. (2016) found, that when a cirrus cloud overlaps with a low-level cloud, the resulting MODIS retrievals indicate a high cloud of medium optical depth. If the corresponding CC-L label is Ci, the classifier learns that high-top, medium-thickness clouds are cirrus, leading to a combined confusion with As and Ac of 21%, i.e. 21% of Ci are falsely labeled as As or Ac (see Fig. **4.6**). A similar

phenomenon was found when the properties of ISCCP regimes were compared with active sensor measurements, indicating that this is a common problem when using passive sensor data for cloud classification (Haynes et al., 2011).
Grid cells with many unrealistic cloud property retrievals, which are often a result of clear-sky, have been removed as the RF can not process them. This reduces the amount of clear-sky contained in CCClim (implicit via the "undetermined" class). These cells can also not simply be labeled as clear-sky as we do not know why the retrievals failed. Furthermore, the sum of the cloud-type RFOs in CCClim is higher than the cloud fraction obtained from ESA-CCI. In combination with the high-latitude clouds with vanishing short-wave CRE, this suggests a bias induced by applying the RF on a different domain than it was trained on. We suspect that some of the As clouds in high latitudes might be clear-sky and differences between the retrievals of MODIS and AVHRR lead to different results. Therefore, the ESA-CCI cloud fraction (*clt*) is provided in CCClim to enable analysis of the cloud types in the appropriate cloud cover context. We would also like to note that CCClim is not suitable for detecting trends in the cloud-type distribution in response to climate change effects, as the underlying ESA-CCI data are deliberately adjusted to produce a long-term stable dataset (Sus et al., 2018).

CCClim is less affected by common problems of the few similar datasets dealing with objective cloud types and which disagree on important details (Li et al., 2015). For example, Stubenrauch et al. (2006) are in qualitative agreement with the ISCCP cloud distributions but for example find a much higher cirrus fraction, especially in the Tropics. Similarly, datasets produced with unsupervised methods are difficult to compare in the first place as the derived clusters do not have a common physical basis. The comparisons made between such clusters and classical ISCCP regimes show disagreements that can be attributed to differences between (active and passive) sensors, definitions of cloud types and distinguishing between individual cloud layers (Kurihana et al., 2022; Li et al., 2015; Marchant et al., 2020). Therefore, a product such as CCClim covering a long time period can take advantage of synergy effects to address some of the limitations inherent to passive and active satellite sensors and provide a more consistent dataset than available from individual instruments. While subjective thresholds are being used in the fuzzy logic classifier of CC-L, there is arguably no better way to obtain cloud observations from space than combining radar, lidar and passive instruments. Using ML for this combination avoids introducing any further subjective biases. The resulting dataset is easy to use in terms of physical interpretation of the cloud types, data volume, horizontal grid and common metadata.

6. Synthetic Observations of Climate Models from Generative Domain Adaptation

6.1. Overview

In the context of global climate model (GCM) evaluation, simulated physical quantities are compared to retrievals obtained from observed radiances. Since these data are obtained in fundamentally different ways, such comparisons are not straightforward. This chapter presents an approach that is complementary to that used in the previous two chapters as it aims to directly improve this comparability. While in the GCM domain physical variables are computed by solving differential equations describing physical processes (forward problem), in the observational domain retrieval algorithms attempt to solve inverse problems. This leads to a systematic inconsistency between the data obtained by any sensor and any GCM. As discussed in Section **2.2**, the sensitivity and resolution achievable for different atmospherical variables vary between satellite instruments, depending on available wavebands and viewing geometry. Further deviations are introduced because satellite instruments can not observe all locations at all times or with equal precision. Thus only limited comparisons can be made between the two domains. So-called *satellite simulators* attempt to mitigate these differences between the domains by computing *synthetic observations* from GCM simulations, in a way that mimics the observations obtained by a satellite instrument. The Cloud Feedback Model Intercomparison Project Observation Simulator Package (COSP) combines several satellite simulators, providing simulated observations for five active and passive satellite sensors to improve the assessment of cloud-related variables similar to Level 2 satellite products (Swales et al., 2018). COSP addresses both the mismatch in the distributions of the physical quantities and, if necessary, the resolutions between observations and GCM.

Satellite simulators have proved useful when evaluating GCMs and have enabled the computation of cloud regimes in GCMs (Schuddeboom and McDonald, 2021; Tselioudis et al., 2021). However, they are somewhat expensive to use, as they typically require application at runtime and individual implementation for each GCM to produce synthetic observations corresponding to the native GCM output. This chapter introduces a new approach that aims to produce synthetic observations without the need for online coupling of the model with a satellite simu-

lator. For this, a method to perform alignment between pairs of GCM and satellite instrument was designed, applicable to existing GCM output data. This is achieved via generative domain adaptation, where synthetic satellite observations are generated by an NN from GCM output on a scene-by-scene basis. The method reduces statistical deviations between the two domains while enforcing the physical consistency of the atmospheric state via a combination of several loss functions. The method focuses on cloud-relevant variables to be applicable with the framework described in Chapters **4** and **5**. In principle, this DA method can be extended to a more comprehensive set of GCM output variables, as long as the same quantities are available from the chosen satellite sensor. The synthetic satellite scenes can then be used as a basis for comparisons between the simulated atmosphere and observations, with biases caused by sampling, viewing geometry or instrument characteristics being reduced. Several metrics are employed to assess if the GCM quality can still be assessed from the adapted scenes, or if the DA introduces unwanted shifts in the data. While the method performs well on all these metrics measuring the difference between distributions as well as individual input and output scenes, it could not be demonstrated that the characteristics of the simulation are preserved, which is a requirement to make the method suitable for GCM evaluation. This chapter documents the design of the method and the challenges encountered in producing the synthetic observations. Despite these issues, this DA approach might still prove useful in climate science, which will be explored in Chapter **7**.

6.2. Domain Adaptation

domain adaptation (DA) has become an expansive topic for different ML applications. It was proposed as a solution to the problem of having too few samples to train an ML model in a *target* domain, by leveraging an abundance of samples in a different *source* domain (Daume III and Marcu, 2006; Zhou et al., 2022). More generally, the goal of DA is to enable a specific application to be just as reliable in the target domain as it is in the source domain. The definition of "application" here is very broad and can extend from ML models over statistical analysis methods to human perception, but for clarity, it will just be referred to as "model" in this thesis. The two domains are differentiated by a *domain shift*, i.e. they can differ in terms of distribution or number of features and labels. The DA can be implemented through modification of the model, modification of the data or possibly both. If labels are only available in the source domain, we speak of unsupervised domain adaptation. In the field of computer vision, domain adaptation has been employed as a way to improve classifier models for different image categories (Fernando et al., 2013; Saenko et al., 2010; Wang and Deng, 2018). Similarly, it was applied in natural language processing to apply the same models to different text types (Blitzer et al., 2007, 2006; Ramponi and Plank, 2020). Different approaches include learning a shared latent representation (Fernando et al., 2013) of the two domains or minimizing the distance between the data distributions (Tzeng et al., 2014).

Generative DA methods are designed to simulate additional samples in the target domain. Here, conditioning on the source domain samples enables the use of its label information, i.e. samples resembling a class with label y can be generated in the target domain even if only the source domain contains the y class. Models could then be trained on the synthetic data and applied in the target domain without errors caused by domain shift. In computer vision, a prominent approach to generate such synthetic data is called image-to-image translation (I2I) (e.g. Isola et al., 2017; Kim et al., 2017; Liu and Tuzel, 2016; Zhu et al., 2017). Most methods in this field are either adversarial- or latent-based (Pang et al., 2022). Latent methods include UNIT (Liu et al., 2017), and MUNIT (Huang et al., 2018), where autoencoder-style architectures are used to encode both domains in a shared latent space while using separate decoders. Adversarial methods based on GANs (Goodfellow et al., 2016a) are frameworks in which two or more models are trained to perform opposite tasks (see Section **2.3.2**). A significant amount of research on GANs in I2I has been published since the release of pix2pix (Isola et al., 2017), increasing scope and stability.
Most relevant to this chapter are methods that can be trained without supervision, such as DiscoGAN (Kim et al., 2017) and CycleGAN (Zhu et al., 2017), where a pair of GAN generators are trained simultaneously to perform inverse translations between two domains. Even though these methods do not require paired samples, they do require a significant amount of samples from both domains and are domain-specific, i.e. can not be applied to a new pair of domains without retraining. Methods like StarGAN (Choi et al., 2018) and StarGANv2 (Choi et al., 2020) aim to make the translation applicable to multiple domains by providing an encoding of the target domains to the generator along with the input image and not requiring a second GAN for the inverse translation. For improved training stability, StarGAN also employs the WGAN architecture (Arjovsky et al., 2017) with gradient penalty (Gulrajani et al., 2017). The CycADA framework (Hoffman et al., 2018) is an evolution of CycleGAN for task-specific domain adaptation, in which the two GANs are trained with additional losses obtained by applying auxiliary task models before and after translation. This way the task models can be adapted to perform optimally on the target domain. A more recently developed family of methods for I2I are denoising-diffusion models such as UNIT-DDPM (Sasaki et al., 2021), which can generate realistic images from a mixture of Gaussian noise and images from the source domain.

Applications of such DA methods in climate science include making the output of different satellite instruments more consistent (Tuia et al., 2016; Xu et al., 2022). Mateo-García et al. (2021) use the CycADA framework to adapt a cloud detection algorithm between two satellite instruments. François et al. (2021) use CycleGAN to help correct temperature and precipitation biases in a small region of GCM output with respect to reanalysis data. Pan et al. (2021) adapt CycleGAN to similarly adjust precipitation output from a GCM in the continental United States with respect to observational data. Fulton et al. (2023) combine UNIT with a classical bias correction method to reduce the statistical deviation of five variables between a single GCM and reanalysis data.

6.3. Methods

6.3.1. Data

The domains between which adaptation is performed are GCM output from the ICON-A (Giorgetta et al., 2018) and satellite observations from ESA-CCI (Stengel et al., 2020). The ESA-CCI (see Section **3.2**) data are instantaneous daily composites (L3U) from June 2009 to June 2011, coarse-grained to 1°. The preparation of the ICON-A data is described in Section **3.3**. In the following the ICON-A domain is denoted as $\mathcal{T}$, as it denotes the target domain, with the source domain ESA-CCI being represented by $\mathcal{S}$. The data are presented to the DA model as pseudo-simultaneous pairs of patches sampled from each domain (comp. Mateo-García et al., 2021). The pairs are pseudo-simultaneous as they are scenes of the same geographical location and calendar week, but not necessarily the same year. This means that on average, both scenes will have similar atmospheric conditions without the requirement for coinciding samples from $\mathcal{S}$ and $\mathcal{T}$. The patch size is set to 32 and the features included in the DA are the same as those used for coarse-resolution regression used in the previous chapters (Table **5.1**). Furthermore, the 2D *clt* as well as the upwelling radiative fluxes at the top of the atmosphere are used as ancillary variables, which are included in both datasets. Figure **6.1** shows the distribution of the eight physical variables in $\mathcal{S}$ and $\mathcal{T}$ domains. Because the shortwave radiative fluxes are only available during daytime and some ESA-CCI retrievals are less accurate during the night (see Section **3.2**), patches with less than 80% daytime pixels are discarded, where "day" is being approximated by requiring all TOA fluxes to be non-vanishing. The upwelling TOA fluxes are then used to compute the CRE (Eq. 2.7). All variables, including CRE and *clt*, are first *log-scaled* and then *minmax-scaled*, such that the distributions are less skewed towards small values and the patches only contain values in the range $[0, 1]$. The minimum/maximum values used for the minmax-scaling are obtained for each variable across both distributions and reduced/increased by 1%. This takes into account slightly out-of-bounds values that are physically possible.

Since the DA algorithm acts on fairly large 32×32 pixel patches, many contain missing values for *ptop* or other variables and they can not all be excluded from training. Missing values for *ptop* are very common because it is only defined for cloudy pixels. Unlike for the other variables, imputing them from surrounding values or replacing them with a fixed value is unphysical, but no physically sensible way of dealing with missing *ptop* values could be constructed. For cloud-free samples, *ptop* is therefore replaced with 1100 hPa as a placeholder, which does not significantly skew the distribution and can easily be replaced by the maximum measured *ptop* after DA has concluded.

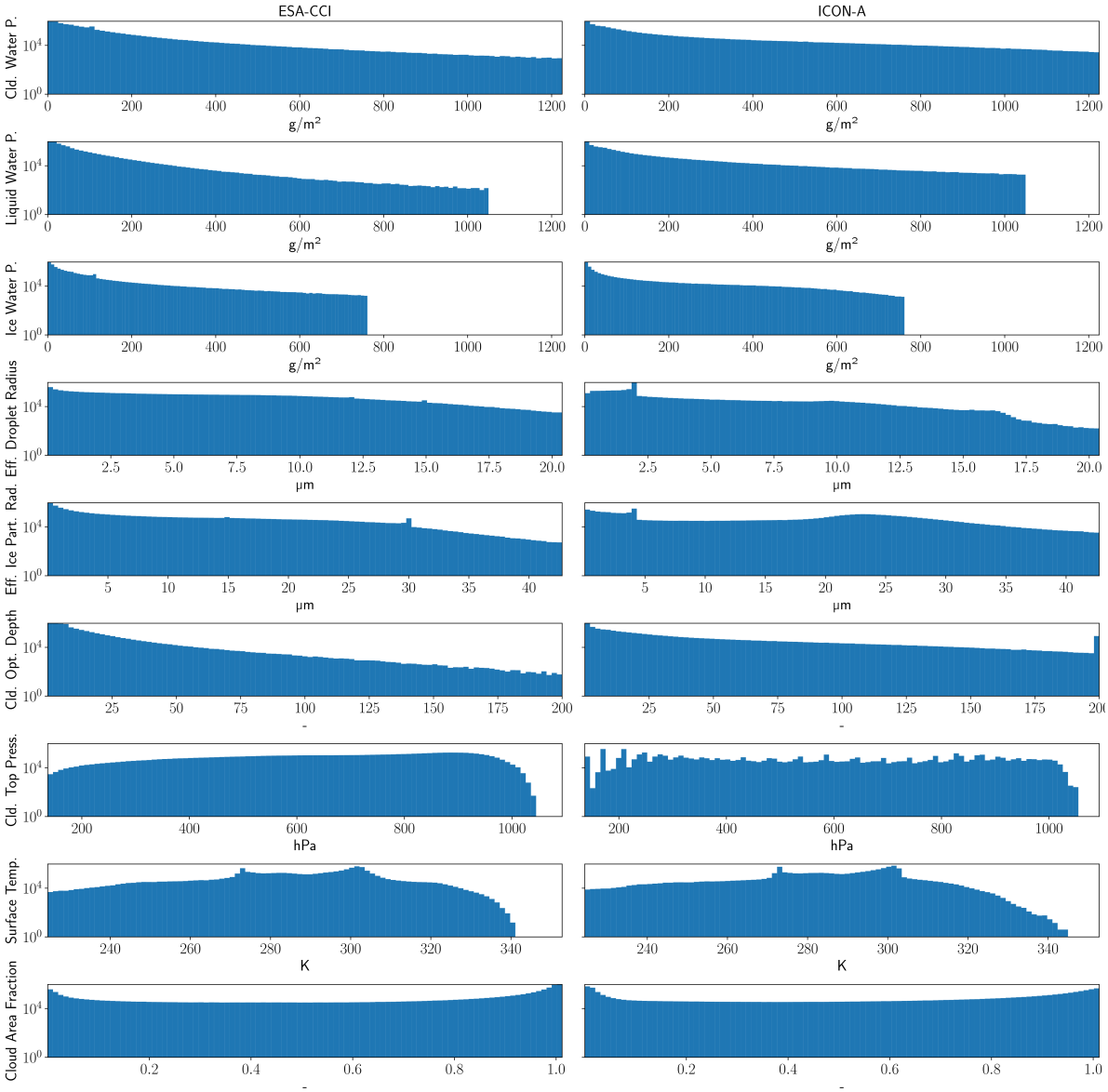


Figure 6.1.: Histograms of the variables in the two domains before domain adaptation. The differences between the histograms are indicative of the domain shift that is to be bridged.

6.3.2. Domain Adaptation Algorithm

The architecture used here is an implementation of the CyCADA framework (Hoffman et al., 2018), which is a DA specific variation of the CycleGAN method (Zhu et al., 2017). The implementation is based on the code [1] by Mateo-García et al. (2021). The CyCADA architecture uses two distinct GAN frameworks (Section **2.3.2**), in which the two generators are trained to transform a sample of domain $\mathcal{S}$ into a semantically similar sample of domain $\mathcal{T}$ and vice versa. This type of architecture works well for images and physical data, like the raw satellite radiances in Mateo-García et al. (2021). However, it seems like less well-behaved distributions with long tails or sharp peaks are less easily reproduced. The presented approach therefore replaces the standard GAN losses with WGAN losses, with which better performance was obtained. Using the building blocks introduced in Section **2.3.2**, a Wasserstein CycleGAN framework can be constructed, consisting of four neural networks, two generators $\mathbf{G_\theta}$, $\mathbf{G_\rho}$ and two critics $\mathbf{F_w}$, $\mathbf{F_v}$, with separate gradients Eq. 6.1 for the critics and Eq. 6.2 for the generators.

$$\nabla(\mathbf{F}_i) = \nabla_i \left(\mathbb{E}_{\mathbf{X}\sim P_{s,t}} \left[\mathbf{F}_i(\mathbf{X}) \right] - \mathbb{E}_{\mathbf{Y}\sim P_{t,s}} \left[\mathbf{F}_i(\mathbf{G}_j(\mathbf{Y})) \right] \right), \tag{6.1}$$

$$\nabla(\mathbf{G}_j) = \nabla_j \mathbb{E}_{\mathbf{Y}\sim P_{t,s}} \left[\mathbf{F}_i(\mathbf{G}_j(\mathbf{Y})) \right], \tag{6.2}$$

$$(i,j) \in \{(\mathbf{w},\boldsymbol{\theta}),(\mathbf{v},\boldsymbol{\rho})\}.$$

To enforce consistency of the synthetic scenes with the input scenes, additional loss functions help to optimize the generators. The cycle-consistency loss $\mathcal{L}_{cc}$ ensures that $G_i(G_j(\mathbf{x})) \approx \mathbf{x}$ and the identity loss $\mathcal{L}_{id}$ ensures $\mathbf{G}(\mathbf{x}) \approx \mathbf{x}$. $\mathcal{L}_{cc}$ is implemented as the MAE and $\mathcal{L}_{id}$ additionally uses a set of weights $\mathbf{w}$ that ensures that features that are harder to optimize for contribute more to the gradient:

$$\mathcal{L}_{cc,i} = \overline{|\mathbf{X} - \mathbf{G}_j(\mathbf{G}_i(\mathbf{X}))|}, \tag{6.3}$$

$$\mathcal{L}_{id,i} = \overline{\mathbf{w}\,|\mathbf{X} - \mathbf{G}_i(\mathbf{X})|}. \tag{6.4}$$

Even though $\mathbf{G}(\mathbf{X}) \equiv \mathbf{X}$ is not a desired outcome, optimizing for a small $\mathcal{L}_{id}$ allows for faster and more stable training. A third consistency loss $\mathcal{L}_{toa}$ acts as a physical soft constraint, ensuring that after adaptation the radiative fluxes at the top of the atmosphere and the total *clt* remain largely unchanged. To that end an auxiliary neural network $\mathbf{C}_{toa}$ is trained to predict *clt* and the long- and shortwave cloud-radiative effects from the DA features. The ground-truth cloud-radiative effects are available for both the ICON-A and ESA-CCI datasets as the top of the atmosphere flux difference between the fluxes including clouds (*all-sky*) and assuming clear sky (Eq. 2.7). Computing CRE and *clt* diagnostically from the cloud properties this way ensures a unidirectional dependence, approximating real physics. Allowing these variables to be directly changed by the DA could result in an unphysical interdependence or a complete

[1] `https://github.com/IPL-UV/pvl8dagans`, last accessed 31st Oct. 2023

lack thereof. $\mathcal{L}_{toa}$ (Eq. 6.5) is equivalent to the segmentation consistency loss in Mateo-García et al. (2021).

$$\mathcal{L}_{toa}^{\rho,\theta} = \mathbb{E}\left[\mathbf{C}_{toa}(\mathbf{X}) - \mathbf{C}_{toa}\left(\mathbf{G}_{\rho,\theta}(\mathbf{X})\right)\right]^2. \tag{6.5}$$

During development of the methods, several functional losses similar to Eq. 6.5 were implemented, with the objective that an NN-emulated function is invariant to the DA. A binary cloud-mask classifier and a multi-class cloud-type classifier, producing results similar to the RFRM in Chapter **4** were investigated as possible options. However, the approach of predicting the two CREs and *clt* appeared to be the most physically founded and accurate.
Because *lwp* and *iwp* by definition sum to *cwp* a physical constraint is implemented as another loss term $\mathcal{L}_{wp}$ (Eq. 6.6) optimizing for this relationship in the synthetic scenes.

$$\mathcal{L}_{wp} = \sqrt{\overline{(cwp - lwp - iwp)^2}}. \tag{6.6}$$

All additional losses contribute to the generator gradient (Eq. 6.8), with the critic only being regularized by a gradient penalty $\mathcal{L}_{GP}$ that depends on the gradient of the critic.(Eq. 6.7) (Gulrajani et al., 2017) [2]. Each loss function has a distinct purpose. Optimizing for Eqs. 6.1 and 6.2 ensures that the distributions of the individual variables are aligned to the new domain while minimizing Eqs. 6.4 and 6.5 and other functional losses conserves the original atmospheric state (e.g. *clt*). Equation 6.3 is helpful for training stability by ensuring that the synthetic scenes can be interpreted by the other generator as scenes from the real domain $\mathcal{S}$ or $\mathcal{T}$.

$$\nabla_{\mathbf{F}_i} = \nabla\mathbf{F}_i + \lambda_{GP}\mathcal{L}_{GP}, \tag{6.7}$$

$$\nabla_{\mathbf{G}_i} = \lambda_{GAN}\nabla\mathbf{G}_i + \lambda_{id}\nabla\mathcal{L}_{id} + \lambda_{cc}\nabla\mathcal{L}_{cc} + \lambda_{toa}\nabla\mathcal{L}_{toa} + \lambda_{wp}\nabla\mathcal{L}_{wp}. \tag{6.8}$$

6.3.3. Evaluation Metrics

Evaluating the output of the CycADA framework is difficult because no ground truth is available for the adapted scenes. While the WGAN minimizes the distance between the generated scene and its pseudosimultaneous counterpart, the distance itself is parametrized by the critic. The WGAN losses alone are therefore insufficient to assess the quality of the DA. All other losses employed here are not designed to quantify the difference between the two domains. A number of performance measures called metrics in the following are therefore applied to the WGAN output, quantifying both the differences between individual scenes as well as the overall distribution of the variables:

- structural similarity index measure (SSIM) (Wang et al., 2004)

[2] Since optimal configuration was found with $\lambda_{GP} = 0$, definition of $\mathcal{L}_{GP}$ is omitted here and referred to the original paper.

- Jensen-Shannon-Divergence (JSD) of univariate distributions
- Pixel-wise MSE of joint distributions
- Wasserstein distance (earth-movers distance (EMD)) of joint distributions

The SSIM can be computed during training and serves as validation metric that evaluates visually perceived similarity of images (see Appendix **A**). The other metrics are computed after training is finished, as the (joint) distributions of all variables are required to be able to compute them. The joint distributions (2D histograms) are computed for all pairs of variables in the DA process for the three relevant domains: real ESA-CCI observations $\mathcal{S}$, real ICON-A output $\mathcal{T}$ and synthetic observations $\mathcal{O}$ generated from the model output. The respective metric is then computed for the three pairs of domains for a quantification of the domain shift. The relative improvement I_r in metric M for the joint distribution of variables i and j is then computed as:

$$I_r(M, i, j) = \frac{M_{\mathcal{TS}}(i, j) - M_{\mathcal{SO}}(i, j)}{M_{\mathcal{TS}}(i, j)}. \tag{6.9}$$

The subscript of M denotes the domains between which the domain shift is quantified. For the computation of the JSD, only the univariate distributions are used for Eq. 6.9. By definition, Eq. 6.9 is positive if the distributions of the synthetic observations are closer to the real observations than they are for the model, measured by the metric M.

Jensen-Shannon Divergence

The Jensen-Shannon divergence D_{JS} is a measure of the distance between two probability distributions P and Q defined on the same sample space. It provides a symmetric ($D_{JS}(x, y) = D_{JS}(y, x)$) alternative to the Kullback-Leibler divergence D_{KL} via the mixture distribution PQ.

$$D_{JS}\left[P, Q\right] = \frac{1}{2}\left(D_{KL}\left[P, PQ\right] + D_{KL}\left[Q, PQ\right]\right), \tag{6.10}$$

$$D_{KL}\left[P, Q\right] = \sum_{x \in [0,1]} P(x) \log \frac{P(x)}{Q(x)}, \tag{6.11}$$

$$PQ \equiv \frac{P + Q}{2}.$$

Because the data are already minmax- and log-scaled for easier training of the NNs, the sample space is, therefore, $[0, 1]$, discretized to 100 bins for computational reasons. D_{JS} is non-negative, bounded by $\log(2)$ and becomes 0 as P approaches Q. Because D_{JS} applies to distributions, a high number of samples is required to provide a robust estimate. It is therefore

only used to evaluate the results of the fully trained models applied to the full validation set. Because D_{JS} is computed from individual distributions of the atmospheric variables the adaptation might lead to physical inconsistencies. A possible example would be a strongly increased *lwp* together with a decreased *cod* in the same sample. This would not show up as an increase in D_{JS} as long as the underlying distributions are still matched.

Joint Distribution Distance

The domain shift between the joint distributions of the physical variables is quantified using the average MSE and EMD for the 2D joint histograms of all variable pairs between the respective domains. These are equivalent to grayscale images and the MSE is easily computed from the pixel-wise differences. For the EMD, the `POT`[3] package is used, which finds the optimal transport distance between the joint histograms, using the L_2-Norm as distance measure. This EMD is not necessarily optimized by the WGAN, even though they operate on the same principles of optimal transport. Firstly, this EMD is computed in the space of joint histograms, not individual samples. Secondly, while in a WGAN the distance measure is parametrized through the critic, `POT` requires a predefined measure (L_2-Norm) to solve the optimal transport problem via iterative optimization (Bonneel et al., 2011).
Combining the four metrics provides measures of how well the synthetic observations reproduce the target domain cloud structure, and how well the statistics match real observations while preserving relationships between the variables.

6.4. Results

The modified CycADA framework implemented here has a large number of hyperparameters, which have been tuned using automatic hyperparameter optimization (HPO), maximizing the relative improvement (Eq. 6.9) of D_{JS} averaged over all variables. An explanation of the tuning procedure and final hyperparameters are included in Appendix **C**. The model that performed best during HPO is used to produce the following results.
The distributions of the bivariate and univariate distributions are shown for all three domains in Fig. **6.2** for three (*iwp*, *cerl*, *ceri*) of the variables as an example. Figure **E.1** in the Appendix shows this for all variables. Judging by the univariate distributions, the synthetic observations are more similar to the source data than the target data. As the HPO algorithm optimizes for a large improvement I_r in the univariate distributions, this is not surprising. Also, it is important to note that some characteristics the of original ICON-A distributions are conserved, e.g. the peak at ~ 0.65 in the distribution of *ceri*. This suggests that the

[3] `https://pythonot.github.io/`, last accessed 27$^{\text{th}}$ Nov. 2023

distributions are not over-corrected, which is one of the main requirements for this WGAN-DA to be useful for GCM evaluation. The performance metrics obtained for the WGAN are displayed in Table **6.1** and show that the optimal set of parameters leads to an improvement in seven of the eight distributions. The outlier is *tsurf*, which significantly reduces the average improvement. This lack of improvement here is most likely due to the "bleeding" of the other variables into the structure of *tsurf*, which can be seen in Fig. **6.4**. Some of the less performant setups did not show this behavior so it should be possible to avoid it with a small adaptation of the architecture. All joint distribution metrics improve, which include the *clt* diagnosed using the $\mathbf{C}_{toa}$ NN. Figure **6.3** shows an improvement for both metrics for all variable pairs. Distinct minima in the distance metrics occur for pairs including *clt*, indicating that *clt* is not as consistent with the DA adapted features as they are with each other. A possible reason for this could be a smoothing effect of $\mathbf{C}_{toa}$ (see Appendix **B**). Individual pseudosimultaneous samples of $\mathcal{S}$, $\mathcal{T}$ and $\mathcal{O}$ are shown in Fig. **6.4**. They show that the structure of the original scene is well represented by the synthetic observations but some overall bias is introduced, evident as a shift in the average color. This bias could possibly be exactly the domain shift that the DA aims to correct for, and is therefore not necessarily detrimental. Also, the structure of some of the channels seems to affect the other channels, most evident in the synthetic observation of *tsurf*, which shows a similar structure to that of the cloud properties. As this correlation is not evident in the source data (bottom row in Fig. **6.4**), it seems to be an artifact introduced by the architecture. Considering all distribution-based metrics in combination with the conservation

Table 6.1.: Metric scores achieved with the optimal WGAN on held-out validation data. Relative improvement is calculated for JSD for 8 univariate distributions, and for MSE and EMD for the resulting 36 2D distributions (see Fig. **6.3**). JSD and SSIM are obtained during HPO and full training, respectively. The 2D distributions are computed afterward and include *clt*, obtained from the other eight variables via $\mathbf{C}_{toa}$. The number of improved distributions are those with positive relative improvement, which are all metrics except JSD for *tsurf*.

SSIM		JSD		EMD		MSE	
$\mathcal{T} \to \mathcal{S}$	$\mathcal{S} \to \mathcal{T} \to \mathcal{S}$	Mean impr. I_r	Num. impr.	Mean impr. I_r	Num. impr.	Mean impr. I_r	Num. impr.
0.64	0.89	0.35	7/8	0.62	36/36	0.72	36/36

of the cloud structure in individual samples, the DA seems to achieve the intended purpose of making GCM output more comparable to observations. However, for this comparison to be useful in GCM evaluation, the modifications introduced by DA are not allowed to significantly affect the physics represented by GCM output, such as the relationships between individual variables. The losses Eqs. 6.5 and 6.6 ensuring correct reproduction of CREs, *clt* and *cwp* are in place to achieve this, but are only soft constraints that could be insufficient. The modifications induced by the DA are therefore tested heuristically for physical consistency by comparing them to modifications induced by the popular satellite simulator COSP (Bodas-Salcedo et al., 2012; Swales et al., 2018). Since COSP accounts for physics more explicitly than a "black-box"-NN and has been widely used and validated, the goal of a DA method should be to modify the

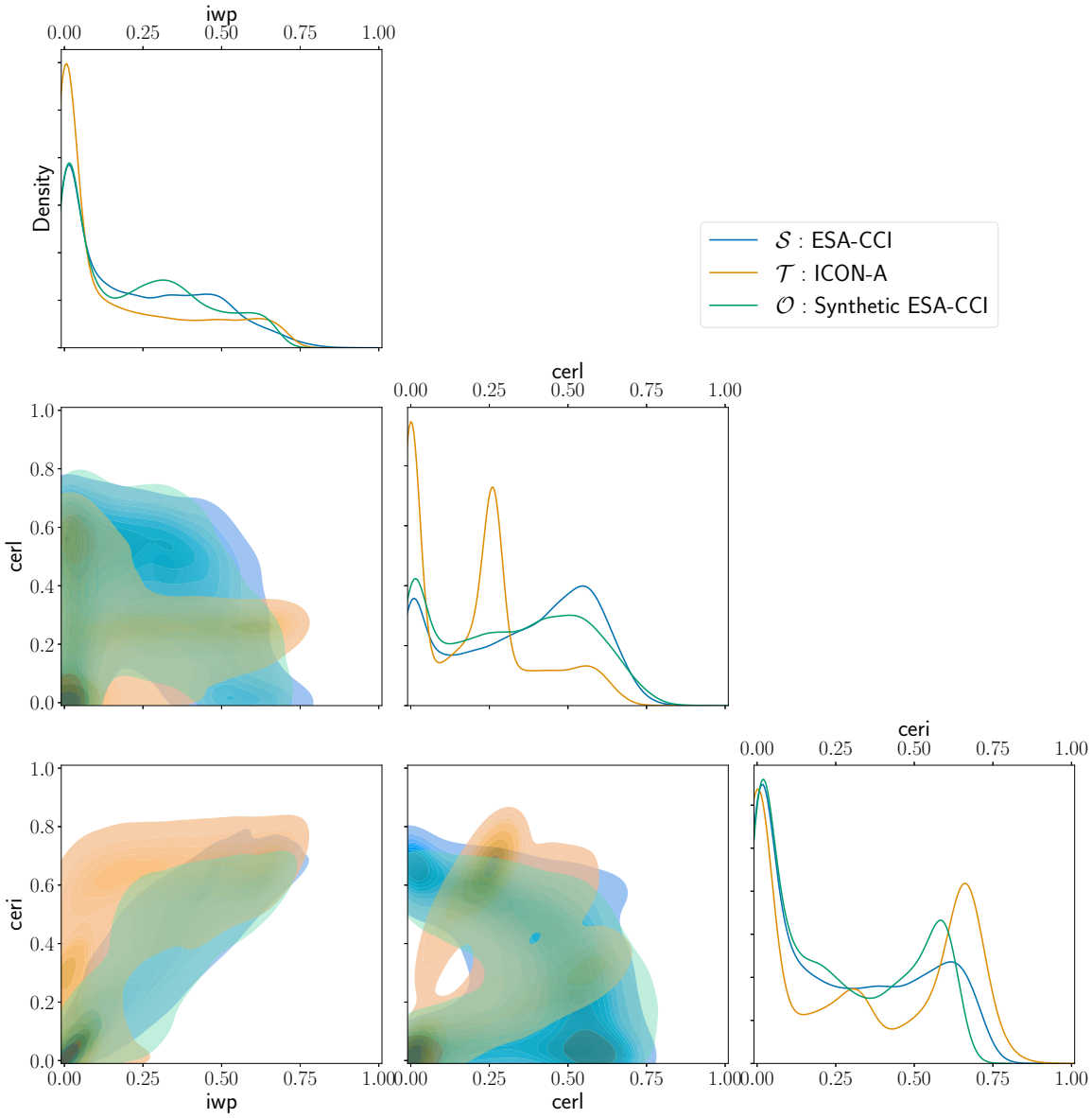


Figure 6.2.: Visual representation of the (joint) distributions in all domains for three exemplary variables. Diagonal: marginal distributions in minmax-log-scaled space. Lower triangle: joint distributions of 2-variable pairs as kernel-density estimates. Generally a higher overlap of $\mathcal{S}/\mathcal{O}$ (blue/green) than $\mathcal{S}/\mathcal{T}$ (blue/orange) is desired. This is clearly the case for the marginals, but harder to tell for the joint distributions. This necessitates using the MSE and EMD metrics to quantify the similarities (Fig. **6.3**).

GCM output in a way that is at least somewhat similar to COSP. Figure **6.5** shows the histograms of *clt* changes in Δ_{clt} from COSP for different GCMs compared to those induced by the WGAN-DA in ICON-A. For COSP, Δ_{clt} is computed by comparing output from CMIP6 simulations applying COSP to the native output of the same models and for each grid cell and time step computing $\Delta_{clt} = clt_{COSP} - clt_{nat}$. The specific COSP and GCM configurations are given in Appendix **F**. The histogram for COSP (blue) is averaged over all GCM runs and shows two major peaks at 0 and -1, corresponding to the instrument simulated detecting the same

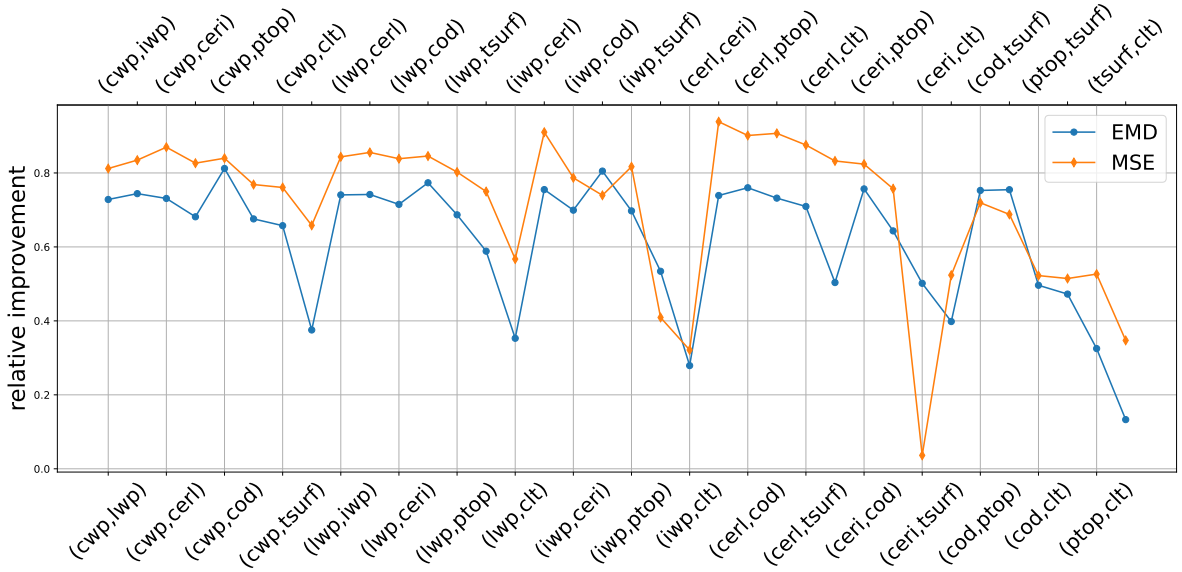


Figure 6.3.: Relative improvement (Eq. 6.9) obtained via DA, of the two metrics EMD and MSE applied to the joint distributions. Both metrics show positive values for all variable pairs, indicating improved statistical similarity between synthetic and real observations. Minima occur for joint distributions with *clt*, which is obtained from $\mathbf{C}_{toa}$. The relatively small improvement here is possibly caused by a small uncertainty introduced by diagnosing *clt* this way.

clt and not detecting any cloud where the GCM simulated a fully cloudy cell, respectively. This highlights the fact that the satellite instrument is not able to detect optically very thin clouds. For the synthetic observations, the *clt* produced with the WGAN-DA is obtained from the modified physical variables using the auxiliary network $\mathbf{C}_{toa}$. Since $\mathbf{C}_{toa}$ displayed strong performance during testing (Appendix **B**), it is assumed that the predicted *clt* distribution is sufficiently accurate. The WGAN-DA *clt* displays the opposite effect of COSP, as the corresponding histogram (orange) shows only a single peak at 0 and more increases in *clt* than decreases. To some extent, this is unsurprising because Eq. 6.5 enforces a small change in *clt*. However, the change in the *clt* statistics (see Fig. **E.1**), rather indicates that the distribution of *clt* is the root cause of the problem. A distribution with peaks at both tails of the image of the target function $\mathbf{f}_{true}$ is hard to learn for a DL model, as the activation functions either do not constrain the output to the image of $\mathbf{f}_{true}$ or have plateauing gradients in the tails. The problem manifests in the univariate distributions as both tail-end peaks being moved slightly towards the center of the distribution. In any case, including *clt* in the set of variables to be directly modified by the WGAN-generators and avoiding the problem of enforcing cloud-cover consistency did not improve this result. The RFRM trained for Chapter **4** can be used to assess how much more comparable the synthetic observations from the WGAN-DA are to the real observations by producing cloud-type predictions and comparing them to CCClim and the cloud types obtained from native ICON-A output. The geographical distributions of the average cloud type RFOs obtained from the synthetic observations are shown in Fig. **6.6**. These

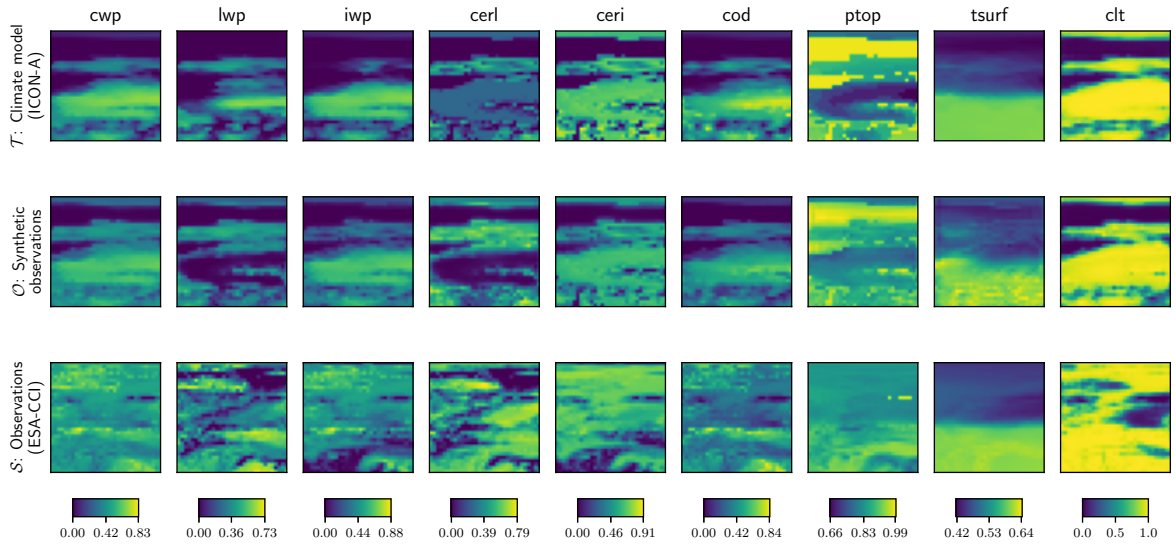


Figure 6.4.: Illustration of the DA process for an individual sample. The upper row shows the different channels from the input domain $\mathcal{T}$, which are used by the generator to produce synthetic observations ($\mathcal{O}$, middle row). The bottom row shows the corresponding observations from $\mathcal{S}$, which show different cloud states than $\mathcal{T}$ and $\mathcal{O}$ by design. The synthetic observations capture the structure of the scene well but are slightly smoothed out. The cloud structure visible in *tsurf* suggests that the connection between individual channels is too high in the network but an architecture that does not display this effect could not be found.

distributions show more similarities to the average distributions in CCClim (Fig. **5.3**), than to the native ICON-A distributions (Fig. **5.9**). Consistent with the analysis of the ICON-A cloud types in Section **5.4** which concluded that many of the Ci in ICON-A are subvisible, the average Ci RFO for synthetic observations is substantially reduced to 0.15, compared to 0.2 for the native output. At the same time, however, the Sc fraction (0.333) is higher than in both native ICON-A (0.259) and CCClim (0.330). Taking into account known biases of ICON-A (Crueger et al., 2018; Giorgetta et al., 2018), the synthetic observations should rather show a relative increase in mid-level cloudiness (Ac, As, Ns), as ICON-A tends to be biased towards these types and rather underestimates low and shallow clouds, especially in the tropics.

6.5. Discussion

The presented method for generative DA uses a WGAN to reduce the domain shift between observational data and GCM output to make them more comparable. The WGAN-DA model turns scenes from ICON-A output into synthetic observations that are similar to measurements from the AVHRR instrument as represented in the ESA-CCI dataset. Almost all metrics measuring similarity between real and synthetic observations show that the synthetic observations

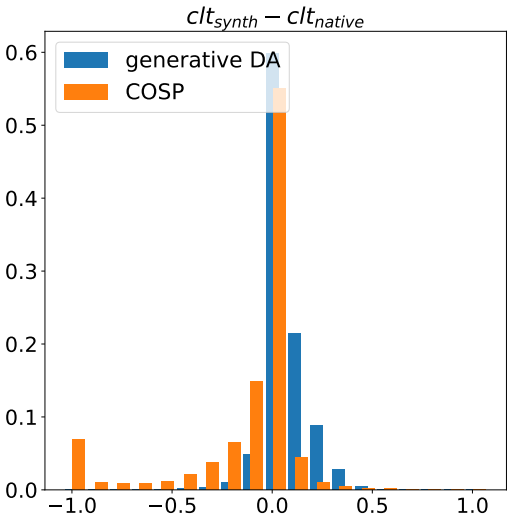


Figure 6.5.: Distribution of differences in *clt* between a native GCM simulation and corresponding synthetic observations. Blue bars show the frequency of a Δclt-value for COSP, simulating synthetic *clt* observations, orange bars show Δ_{clt} for the WGAN-DA method applied to ICON-A.

are more similar to the real observations than the original ICON scenes. Visual inspection of the marginal and joint distributions in the respective domains shows that while the distributions are moved closer to the ESA-CCI data, some characteristics of the original ICON-A distribution remain. This suggests that features of the target domain (ICON-A) are preserved and that the DA does not produce an overcorrection towards the source domain (ESA-CCI). Even though the metrics indicate a successful and physically consistent DA, the counter-intuitive change in *clt* demonstrates that some of the changes do not correspond to characteristics of the AVHRR sensor that is to be simulated here. The cloud type distributions produced from synthetic observations also indicate that relevant ICON-A characteristics, such as the underestimation of low clouds have been removed by the WGAN-DA, rendering the synthetic observations unsuitable for evaluation of ICON-A. Since the WGAN operates as a black box, an explanation of why specific modifications are induced can not be easily obtained. The DA modifications can therefore not be physically justified. Analyzing the changes of all variables as has been done for *clt* (Fig. **6.5**), shows near-Gaussian histograms centered around 0. This indicates, that despite the extensive HPO that was performed to optimize the architecture, the model remains fairly simple. This is likely a consequence of the extremely complex loss hyperplane formed by the seven loss functions and four different gradients. The consequence of a complex loss surface is that even if a configuration of the model parameters exists that is (near-)optimal in terms of the losses while simultaneously preserving physical consistency, it is hard to find through SGD. The issue of not being able to perform physically reasonable and consistent DA therefore does not stem from the architecture or its (hyper)parameters, but rather from the definition of the losses and metrics. It is conceivable that replacing the soft constraints

designed to ensure physical consistency ($\mathcal{L}_{toa}, \mathcal{L}_{cwp}$) with harder constraints would lead to more fit-for-purpose results (Beucler et al., 2021; Harder et al., 2022). The challenge here is to find losses that allow for stable training without making the loss hyperplane prohibitively complex and a physical constraint or metric that can be optimized for during validation and HPO.

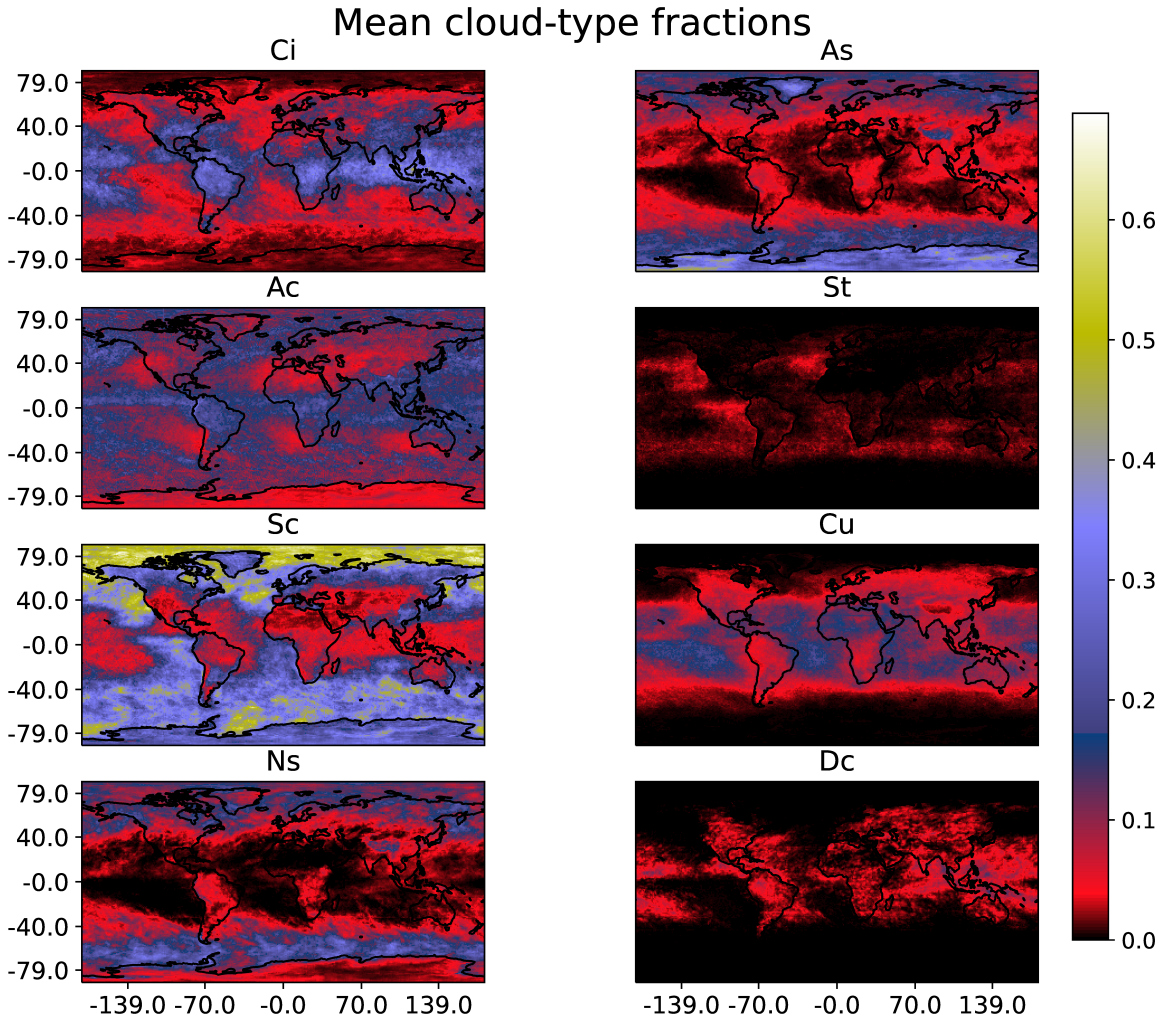


Figure 6.6.: Cloud type distributions obtained with the trained RFRM from Chapter **4** using the synthetic observations of ICON-A from the WGAN-DA as inputs. In comparison to the cloud type distributions obtained from the native ICON-A output (Fig. **5.9**), the synthetic observations show much higher Sc and much lower Ci RFOs. The geographical distributions are much closer to those in CCClim (Fig. **5.3**), but with less prominent subtropical Sc decks.

7. Conclusion

7.1. Summary

Global climate models are important tools to assess the effects of anthropogenic climate change on the Earth system (IPCC, 2021) and satellite observations are a crucial part of their evaluation (Bock et al., 2020; Eyring et al., 2021). Clouds have repeatedly been identified as the primary source of uncertainty in GCM projections (Bony, 2005; Schneider et al., 2019; Sherwood et al., 2014; Stevens and Bony, 2013; Zelinka et al., 2017), owing to their complex interaction with many dynamical, thermodynamical and chemical processes from the microscopic to the global scale. There is therefore a need for observational products that highlight cloud-related physical processes and for analysis techniques that enable meaningful comparison between these cloud products and GCM output. In this dissertation, novel concepts for using satellite observations to evaluate clouds in GCMs are proposed, aiming to make this evaluation easier, more objective and effective. This goal is approached from two sides, firstly to make cloud representations in both observations and GCMs more interpretable through the use of cloud classes and secondly to make these two data domains more directly comparable. Both approaches are based on ML methods, enabling efficient processing of large amounts of data with minimal subjective biases.
The work conducted for this dissertation led to the submission of two articles to scientific journals and the publication of a new dataset. The framework to predict interpretable cloud classes from coarse resolution data presented in Chapter **4** was published as a peer-reviewed paper (Kaps et al., 2023a) and used to create a novel satellite cloud-class climatology, which is now publicly available (Kaps et al., 2023b). The corresponding description paper is under review at the time of submission of this thesis (Kaps et al., 2023c) (see Chapter **5**).

The cloud classification framework presented in Chapter **4** uses a combination of data from active and passive satellite sensors to produce a global climatology of cloud-type distributions called CCClim (Chapter **5**). With CCClim, which is based on the ESA-CCI dataset, clouds can be studied globally and for a long period in a more interpretable and process-oriented way than with only the observed cloud properties. Unlike most previous products, CCClim is based on data from active sensors, which were used to train ML models to predict the cloud classes. This way, information on the vertical cloud structures is implicitly contained in the

predictions of the two-dimensional distributions. Specific sampling criteria for the cloud-type distributions and strategies for presenting these data are proposed. This aims to unlock the potential of CCClim and similar datasets for a more in-depth, process-based analysis.
The framework used to create CCClim is designed to be also applicable to GCM output, but appropriate output providing the required variables at sufficient temporal resolutions is not readily available at this stage. An application to GCM output from a custom simulation is used to illustrate the GCM evaluation possibilities provided by this framework. The cloud types predicted from GCM output can provide additional information by themselves or through comparison to CCClim. Since such comparison between GCM and satellite data is subject to uncertainties coming from the systematic differences between the two, a novel method to mitigate those differences is investigated in Chapter **6**. This method uses generative domain adaptation with NNs to produce synthetic observations of GCM scenes, to provide output similar to satellite simulators, but without the need to be specifically implemented for each GCM and run on top of the GCM itself in each simulation. However, simultaneously optimizing the NN for both improved domain similarity and physical consistency is an issue that remains unresolved. Nevertheless, the positive and negative results of this study can inform the development of similar methods, or at least help with the interpretation of results from other DL methods used in this field, for example for bias correction (e.g. Fulton et al., 2023).

7.2. Discussion

The results for Chapters **4** to **6** are discussed in the context of the three central scientific questions posed in Chapter **1**:

- **Question 1**: "Can physically robust and self-consistent cloud-type distributions be obtained from data at resolutions typical for global climate models?"

- **Question 2**: "Does the explicit addition of cloud-type labels benefit the analysis and understanding of cloud-related processes to improve climate model evaluation?"

- **Question 3**: "Can the systematic bias (domain shift) between satellite data and climate model output be quantified or possibly reduced using generative domain adaptation"

To present the capabilities of the cloud-type prediction framework in the context of current research, two secondary questions are discussed first:

- **Question 1.1** : "Can combining active and passive sensor data improve cloud classification with respect to methods using only passive sensors?"

- **Question 1.2** : "Can interpretability and physical accuracy be increased by using supervised methods with morphological cloud classes in contrast to unsupervised clustering?"

This work significantly relies on the CC-L dataset, which enables the supervised training of ML models to predict morphological cloud types. The underlying fuzzy-logic classifier is more efficient and arguably more objective than having humans manually assign new labels as in e.g. Stevens et al., 2019. Using CC-L as a label ground truth and MODIS data as input is therefore a natural choice that leverages the synergy and consistency of the two A-Train sensors. The strengths and weaknesses of both the CC-L and the MODIS Cloud Product dataset are combined in the CUMULO dataset by Zantedeschi et al. (2019). Strengths include better characterization of ice and mixed-phase clouds via the radar/lidar combination, which seems to extend to the ML predictions as both Ns and As predictions show a high correlation with the CC-L ground truth and there is little confusion between As and Ac. A weakness of passive sensors that could not be fully overcome with the CUMULO based classification is the handling of overlapping clouds, as evident in the $\sim$ 20 % confusion between Ci and Cu. While the physical variable distributions associated with each of the eight cloud types are as expected, the large fraction of St mistakenly classified as Sc suggests that the physical distinction between Sc and St is not correctly determined already at the classification stage. As mentioned before, this is an issue the classifier inherits from the CC-L retrieval, which distinguishes these types through differences in mesoscale cloud homogeneity, which cannot be reliably detected by these active sensors. St and Sc should therefore not be distinguished by this method but combined into a single class. The capability of this framework to produce cloud-type distributions suitable to downstream tasks therefore only extends to seven cloud classes:

- High Ice Clouds / Cirrus + Cirrostratus (see Wang (2019a))
- Altostratus
- Altocumulus
- Low Stratiform / Stratus + Stratocumulus
- Cumulus
- Nimbostratus
- Deep Convective

Considering that for example the traditional ISCCP classification (Rossow and Schiffer, 1999) does not account for homogeneity, removing the distinction between St and Sc would not

introduce a relevant disadvantage. The answer to **Question 1.1** is, therefore, a "yes" but not without limitations: combining active and passive sensor data with ML introduces benefits for cloud classification and should therefore be considered in the future even though some issues remain at this stage.

Besides the use of active sensors, another key feature of this new framework is the use of established morphological cloud classes, contrary to existing methods and datasets that provide *cloud regimes* or *weather states*. These unsupervised approaches have several advantages: they are easy to implement, usually very interpretable due to the use of only two input features and applicable to most GCMs at coarse horizontal and temporal resolution. These aspects, however, stem from the simplicity of the clustering and result in an arguably coarse categorization. The approach developed here instead leverages high-quality, high-resolution labels to obtain interpretable cloud classes even at coarse resolution. The resulting cloud-type distributions provide reliable long-term statistics for each of the classes, which can be associated with physical processes and are more directly related to established cloud classes than clusters. Also, the CCClim approach is fundamentally different from that of Rasp et al. (2020) and Stevens et al. (2019), where *new* morphological cloud regimes are found from manual categorization of satellite images. New cloud regimes found this way are not necessarily related to physical processes in a way that is easy to interpret when used for GCM evaluation. CCClim in contrast incorporates information from more detailed cloud types whose relationships to physical processes have been studied for more than a century.
The process-based characterization of the cloud type distributions used here inherently deals with mixtures of several types and is therefore not dissimilar to an analysis using weather states, e.g. in Oreopoulos and Rossow (2011) and Tselioudis et al. (2013). In contrast, by including the other cloud types contained in a grid cell, CCClim provides direct access to all possible confounders contributing to the state of the cell. This is important as especially at low resolutions typical for GCMs a grid cell is rarely defined by only one cloud type. Comparing for example to the results of Schuddeboom et al. (2018), where cloud clusters were found to vary in composition regionally, these variations are explicitly accounted for here by using subgrid-scale distributions of cloud types. Furthermore, presenting the data in terms of cloud-type fractions allows for a large variety of analysis methods, some of which are demonstrated in Chapter **5**.
A disadvantage of the presented method compared to unsupervised approaches, like weather states, is the increased complexity. This is shown by the cloud type distributions obtained from ICON-A output, which show significant deviations from the CCClim data, but attributing them to specific causes requires detailed investigation. Nevertheless, the cloud type distributions found for ICON-A using the RFRM could hint at possible problems in the depth of convection and the dissipation of Ci clouds. In the context of this work, a more comprehensive evaluation of ICON-A is not possible since no tuned simulation is available with the required variables. In contrast, methods only using *cod* and *ptop* as inputs can be applied to GCMs relatively bias-free as long as a corresponding output from a satellite simulator is available (e.g. Tselioudis et al., 2021).

While the synthesis of multiple sensors contributes to the quality of CC-L and CUMULO, the extent to which this translates to the two-dimensional distributions in CCClim is less obvious. This needs to be evaluated in the context of CCClim's capabilities and similar datasets since no consensus "ground truth" classification is available: with CCClim the co-occurrence of certain amounts of cloud types can be quantified, giving insight into the relationships - such as transitions - between different cloud types. While similar analyses would be possible with the ISCCP-H dataset (Young et al., 2018), CCClim is closer to the cloud-type distributions of the CC-L dataset, that ensures high fidelity through the use of active sensors. Good examples are the Ns and As types, which in ISCCP-H differ strongly from the CC-L distributions in midlatitudes, polar regions and over mountain ranges. This overestimation of mid-level cloudiness is a known bias of ISCCP (Haynes et al., 2011), and the absence of this bias in CCClim indicates a significant advantage afforded by the use of active sensor data from CC-L. If the CC-L dataset were adopted as ground truth because of its use of active sensors, CCClim could be considered more accurate than datasets based on passive sensors only. However other factors than the probable accuracy of the cloud classes need to be taken into account. For example, CCClim not only differs from other datasets through the way the cloud classes are obtained but also in resolution and included cloud properties. For example, the ISCCP data have higher temporal resolution and provide a different set of physical variables alongside the cloud types. These additional cloud properties form the physical context in which process-oriented analyses are performed. Therefore employing multiple complementary cloud class datasets like CCClim, ISCCP and CC-L will prove most useful for GCM evaluation.

To summarize, this new ML method conveys an increased amount of information associated with better interpretability at the cost of some convenience compared to unsupervised clustering. **Question 1.2** can therefore be answered with "yes": the results show that the synthesis of measurements from active and passive sensors can serve as a basis to train ML models to be able to predict physically robust, consistent and interpretable cloud classes. The validity of this holds even for data at typical horizontal resolutions of GCMs, thus affirmatively answering **Question 1**.

Because application of this framework to GCMs is still limited to a single custom simulation that outputs all required variables, examples for process-based analyses are largely based on the CCClim dataset. Because the included cloud types are defined similarly to the well-established WMO types, they enable a more direct qualitative assessment of the atmospheric state than what is possible with only the corresponding physical properties. While any cloud classification scheme will introduce uncertainties and inaccuracies, in CCClim these uncertainties can be reduced through careful sampling of the data, which also accounts for mixtures of cloud types. This way, CCClim can provide insights into co-occurring cloud types or regimes dominated by a certain type even without access to discrete high-resolution cloud-type labels. Therefore, with suitable statistical analysis, CCClim also provides quantitative benefits over datasets containing only cloud property retrievals. Regarding **Question 2** it can therefore be concluded

that cloud type analysis is a beneficial addition to the conventional analysis of cloud properties with the potential to aid GCM evaluation by highlighting processes of interest more directly. The extent to which this potential can be realized is less clear. The application to the custom ICON-A simulation provides some initial results, but this aspect of **Question 2** can not be answered conclusively with the current results. In general, the cloud-type distributions found for ICON-A are mostly as expected from its known characteristics, such as a smaller low-cloud amount in the subtropics. Cloud characteristics that are not as evident from the simulated cloud properties include a more binary and strongly seasonally modulated distribution of deep convection and an abundance of very thin Ci near the tropopause. This type of analysis is therefore a valuable addition to the growing selection of methods evaluating GCMs via cloud types or regimes.

Cloud regime evaluations of GCMs usually use satellite simulators to reduce systematic deviations between models and retrievals (e.g. Tselioudis et al., 2021; Williams and Webb, 2008). The DA method presented in Chapter **6** is designed to solve the same issue with a DL approach that is applicable offline to existing GCM output. The difficulty in analyzing the ICON-A cloud-type distributions in Chapter **5** shows why removing the systematic biases between the two datasets using DA could be helpful. Since the method is applicable to existing GCM output, the contributions of systematic bias and GCM errors to the predictions from Section **5.4** could possibly be quantified. This was attempted by applying the RFRM before and after DA, such that ideally, remaining deviations would then be almost completely attributable to GCM errors. However, the WGAN-DA seems to overcorrect in favor of the observational data and the RFRM predictions on the synthetic observations show little similarity to those from native ICON-A output. This lack of physical consistency means that the WGAN-DA does not allow for such an attribution of deviations only to ICON-A errors. Regarding **Question 3** it has to therefore be concluded that as of now, the WGAN-DA cannot reduce or quantify the domain shift in a physically reasonable way. Two main reasons for this have been identified. Firstly, the physical constraints used in training the method do not seem to be suitable since the results appear to be physically inconsistent even though the constraints are being observed. Secondly, the loss hyperplane is too complex to find a model that satisfies all requirements, because too many gradients are involved. Both problems could possibly be solved at once through implementation of physical constraints not as loss functions but as hard constraints (Beucler et al., 2021; Harder et al., 2022).

7.3. Outlook

The most important aspect of a ML cloud classification is physical consistency, which for the method presented in Chapters **4** and **5** could be shown across all stages, including classification, regression and transfer to other datasets. However, the performance metrics such as F1-Score

in classification and R2-Score in the regression stage are less impressive. A possible remedy would be changes in the architectures, which for both models can be physically motivated: since a major issue of the classification is the confusion between St and Sc and these cloud types mostly differ in horizontal homogeneity, a better handling of this feature might improve results. It is therefore possible that an architecture that classifies all pixels in larger patches of data instead of individual 3×3 tiles could account for this. However, designing such an architecture is difficult due to the sparse labels in CC-L and requires that the St and Sc labels from CC-L are distinguishable. While it is known that CC-L has issues with St and Sc, the extent of this problem is unclear (Wang, 2019a). A CNN that accounts for spatial structures of clouds would also be an obvious candidate to replace the RFRM in the regression stage. However, as stated in Chapter **4** this strategy was attempted without success. Nevertheless, a DL solution can be expected to be found that is more accurate and efficient than the RFRM.

Analysis of the CCClim dataset has shown that there are benefits in including active sensor measurements in cloud classification that might prove useful in downstream tasks. Assuming that future GCMs provide the required variables, i.e. by including them in the standard output required for CMIP Phase 7, the RFRM can easily be used to obtain cloud-type distributions from GCMs for comparison with CCClim, which is a strategy with the potential to significantly improve process-oriented GCM evaluation, as possible deficiencies would be much more accessible. Using synthetic observations for this task is recommended but not required, as the RFRM responds little and predictably to small changes in the data, due to the properties provided by the bagging and boosting methods. Finding a way to produce reliable synthetic observations with the WGAN-DA relies on effective restrictions of the DA-induced changes to be physically founded, preferably by directly embedding them in the architecture. A possibility might be to supplement or completely replace aspects of the NN with physical relationships between variables instead of implementing additional loss functions to optimize for these relationships. Two possibilities are the sum of liquid and ice water path (Eq. 6.6) or the relationship between *cwp*, *cer* and *cod* (Eq. 2.27). It also might be possible to use generative NNs for applications other than GCM evaluation. The WGAN-DA produces synthetic observations that are statistically close to real observations while keeping structure, CRE and cloud cover close to the original scene. As satellite simulators are not designed to do so, this could be the key feature the WGAN-DA can exploit in a different application. For example, assuming that retrievals from satellite observations produce a more realistic estimate of the atmospheric state than typical GCMs, a generative DL method correcting towards these observations might be useful for correcting errors and biases of GCMs. Such bias corrections could then also be applied to projections from these models potentially leading to more accurate projections of future climate. A possible source of problems here would be "correction" of the projections towards the observed, i.e. current, climate. The WGAN-DA would therefore have to be updated and employ explainability methods to make bias correction explainable and therefore justifiable (McGovern et al., 2019).

While direct evaluation of cloud types in GCMs still requires a few more steps, ML and specifically DL methods already enable sophisticated characterization of clouds in observations. A combination of sensors and variations of ML model architectures has many benefits for GCM development. Observational products like ISCCP and CCClim can be compared with each other, leading to a more complete and less ambiguous definition of cloud types. New satellite sensors and platforms like EarthCARE (Wehr et al., 2023) can contribute high-quality measurements to improve the accuracy of retrievals. More powerful DL methods will transform the way this data is used and interpreted (Bazi et al., 2021). Multiple studies have already shown that employing cloud types or regimes will advance the understanding of cloud processes (Oreopoulos et al., 2016; Williams and Webb, 2008; Zelinka et al., 2022a). This advance is expected to enable the development of more accurate and reliable GCMs with the ultimate goal of a better understanding of climate change and how it can be mitigated and prepared for.

A. Appendix

A. Invertible Residual Network

The IResNet architecture is adaptable in the number and scale of blocks and residual bottlenecks. The IResNet used here consists of five blocks, each containing three bottlenecks, with $Z_k \in (4, 16, 32, 32, 32)$ filters per convolutional layer, $k = 1...5$ being the consecutive blocks. All convolutional layers use the exponential linear unit (ELU) with $\alpha = 1$ as activation function:

$$z_{elu}(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha \cdot (e^x - 1), & \text{if } x \leq 0 \end{cases} \tag{A.1}$$

Each bottleneck is followed up with activation normalization (Kingma and Dhariwal, 2018). The final output is produced using a classifier head, that is a dense NN with input batch normalization, RELU activation and averaging over each tile.

B. Auxiliary Network for Domain Adaptation Physical Consistency

Matching of *clt*, LW and SW CREs in the input and output of the DA generators is optimized for using Eq. 6.5. The NN required to compute these quantities from the available input variables is trained in advance using the same data on which the DA-WGAN is trained. The model is a CNN with an architecture identical to the WGAN generators, trained using MSE between input and output patches, dropout and early stopping. Sufficient performance is ensured using a validation split, a sample of which is shown in Fig. **B.1**. The SSIM for the target domain is above 0.87 and that of the source domain is above 0.93 for every channel. The source-domain score is more relevant here, as this part of the network is used for predicting CRE and cloud fraction from the synthetic observations. All results here indicate a reliable prediction of CRE and cloud fraction is possible with $\mathbf{C}_{toa}$.

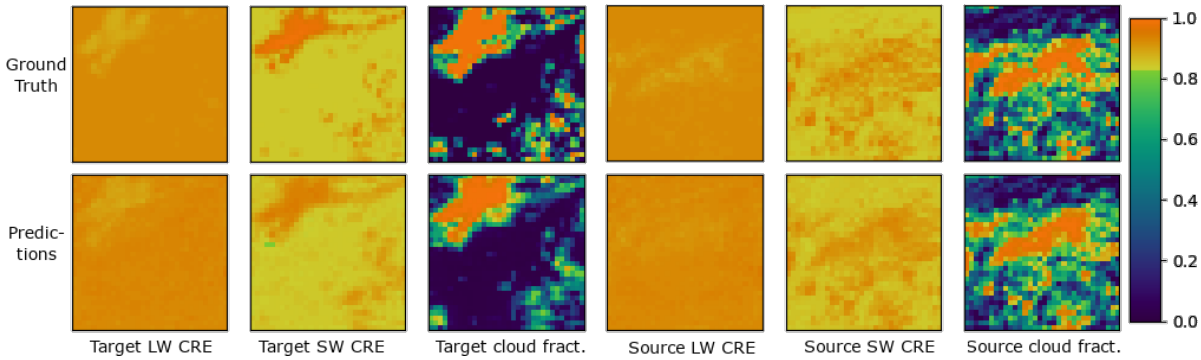


Figure B.1.: Sample from the validation set of $\mathbf{C}_{toa}$, showing the ground truth CRE and cloud fraction patches as well as those predicted with $\mathbf{C}_{toa}$. The structure of all six features is matched well, with only minimal smoothing. Again, source and target refer to ESA-CCI and ICON-A, respectively.

C. Automatic Hyperparameter Tuning of the WGAN

The WGAN has 17 hyperparameters that can be tuned to achieve better stability and performance, most importantly the number of filters in the generators and critics and the weighting of the many loss functions. Additionally, the use of dropout, the batch size and the patch size were fixed and therefore excluded from HPO. The tuning was performed with the `Ray Tune` library, using the asynchronous successive halving algorithm (Li et al., 2018). For a given set of hyperparameters, the algorithm trains the WGAN with the ability to stop early on converged or diverging losses/metrics. The performance of the generator $\mathbf{G}_1 : \mathcal{T} \to \mathcal{O}$ is then evaluated using the univariate JSD for each channel. Maximizing the mean increase of JSD with respect to the JSD between $\mathcal{S}$ and $\mathcal{T}$ then informs the choice of the next set of hyperparameters. This is repeated until no other possible combination of hyperparameters is available. The final set of hyperparameters with their search spaces is given in Table **C.1**.

In the following, the parameters in Table **C.1** which have yet not been defined in Section **6.3** will be discussed. The most important parameter is the gradient threshold ϵ above which the gradients are clipped. It acts to enforce Lipschitz-continuity of the critics, which is required to approximate the EMD. As clipping gradients is highly unelegant and the threshold value is entirely empirical, a specific WGAN gradient penalty has been suggested by Gulrajani et al. (2017). In theory, it can enforce Lipschitz-continuity dynamically and significantly improve training stability. However, the results of the present HPO show that it can as easily cause the training to diverge. The number of pretrain epochs denotes the number of epochs in which the WGAN is trained using mostly the standard WGAN updates Eq. 6.2 and Eq. 6.1, with $\mathcal{L}_{id}$ and $\mathcal{L}_{cyc}$ are reduced by a factor of 100 and the other losses are set to 0. The cyclic learning rate (Smith, 2017) varies the learning rate of the Adam optimizer (Kingma and Ba, 2015) periodically in the range $[\lambda, 50\lambda]$. The softplus activation function is a continuous version of

	Final	Search Space
Generator filters	96	$[24, 256]$
Critic filters	5	$[4, 13]$
λ_{GAN}	30	$[0, 50]$
λ_{id}	7.5	$[0, 10]$
λ_{wp}	0.002792	$[1 \cdot 10^{-5}, 2]$
λ_{cc}	9	$[0, 10]$
λ_{toa}	4	$[0, 5]$
Gradient penalty (GP)	no	yes/no
λ_{GP}	-	$[0.1, 100]$
Generator normalization	batchnorm	batchnorm/instancenorm/none
Generator L_2 Regularization	yes	yes/no
Num. pretrain epochs	0	$0, 1, 2$
Activation of critic blocks	softplus	RELU/softplus/linear
Gradient threshold ϵ	0.0498	$[0.0001, 0.1]$
Learning rate	$1 \cdot 10^{-5}$	$[1 \cdot 10^{-9}, 1 \cdot 10^{-4}]$
Cyclic learning rate	yes	yes/no
Update ratio $\frac{\#\mathbf{F}}{\#\mathbf{G}}$	1	$1, 5, 15, 50$
Batch size[1]	800	-
Patch size[1]	32	-
Dropout probability[1]	0.1	-

Table C.1.: Hyperparameters found by automatic tuning in the given search spaces. [1] Fixed and not subject to tuning.

RELU (Eq. C.1) which together with a linear activation are options for the activation of the hidden layers in the critics.

$$\text{Softplus}(x) = \frac{1}{\beta} \log\left(1 + e^{\beta x}\right). \tag{C.1}$$

In practice, β is usually set to 1, in general $\beta \geq 1$.

The generators may include normalization layers between the convolutional layers, which can be either the batchnorm (Ioffe and Szegedy, 2015) or the instancenorm (Ulyanov et al., 2016) variety. Both act as regularizers that stabilize training but can limit the expressiveness of the model.

A common problem in GAN training is getting stuck in a parameter range in which the discriminators/critics substantially outperform the generators, preventing any improvement in the latter. This problem is tackled here with the possibility of the generator having much more complexity than the critics via the number of filters. Multiple critic gradient updates per generator update (ratio $\frac{\#\mathbf{F}}{\#\mathbf{G}}$) are made possible, which is a common technique to stabilize training (Goodfellow et al., 2016b), but the optimal solution here is to always have simultaneous updates, which shows that the training being generator-limited and that the generator needs this higher complexity.

D. Structural Similarity Index Measure

The SSIM D_{SM} was introduced by Wang et al. (2004) to provide a measure of image quality that is in line with human-perceived quality. The SSIM simultaneously compares the luminance, contrast and structure of a potentially erroneous image and a ground truth image. D_{SM} satisfies (Wang et al., 2004):

- $D_{SM}(\mathbf{X}, \mathbf{Y}) = D_{SM}(\mathbf{Y}, \mathbf{X})$
- $D_{SM} \leq 1$
- $D_{SM}(\mathbf{X}, \mathbf{Y}) = 1$ if and only if $\mathbf{X} = \mathbf{Y}$

The implementation of the luminance, contrast and structure measures are given here by the `TensorFlow` implementation [1], which follows the original recommendations by Wang et al. (2004). After each training epoch D_{SM} is computed on the validation data for each variable individually and averaged. Computing $D_{SM}(\mathbf{X}, \mathbf{G}_i(\mathbf{X}))$ (real/synthetic) and $D_{SM}(\mathbf{X}, \mathbf{G}_j(\mathbf{G}_i(\mathbf{X})))$ (real/cycled) provides a good measure of how well individual synthetic scenes fit the source distribution perceptually. The measure of structure is especially helpful here, as it indicates the extent to which the WGAN "hallucinates" features, which generative models are known, and often intended to do. In this application hallucination should be avoided, as roughly the same atmospheric structure, most importantly cloud cover, should be reproduced. In a sense, an optimal SSIM is mutually exclusive with optimizing the other metrics, which become optimal for a pronounced change in the statistical distribution of the variables towards the source distribution.

[1] `https://www.tensorflow.org/api_docs/python/tf/image/ssim` (accessed 27$^{\text{th}}$ September 2023) for `TensorFlow` version 2.11 .

E. Full Joint Distributions from Domain Adaptation

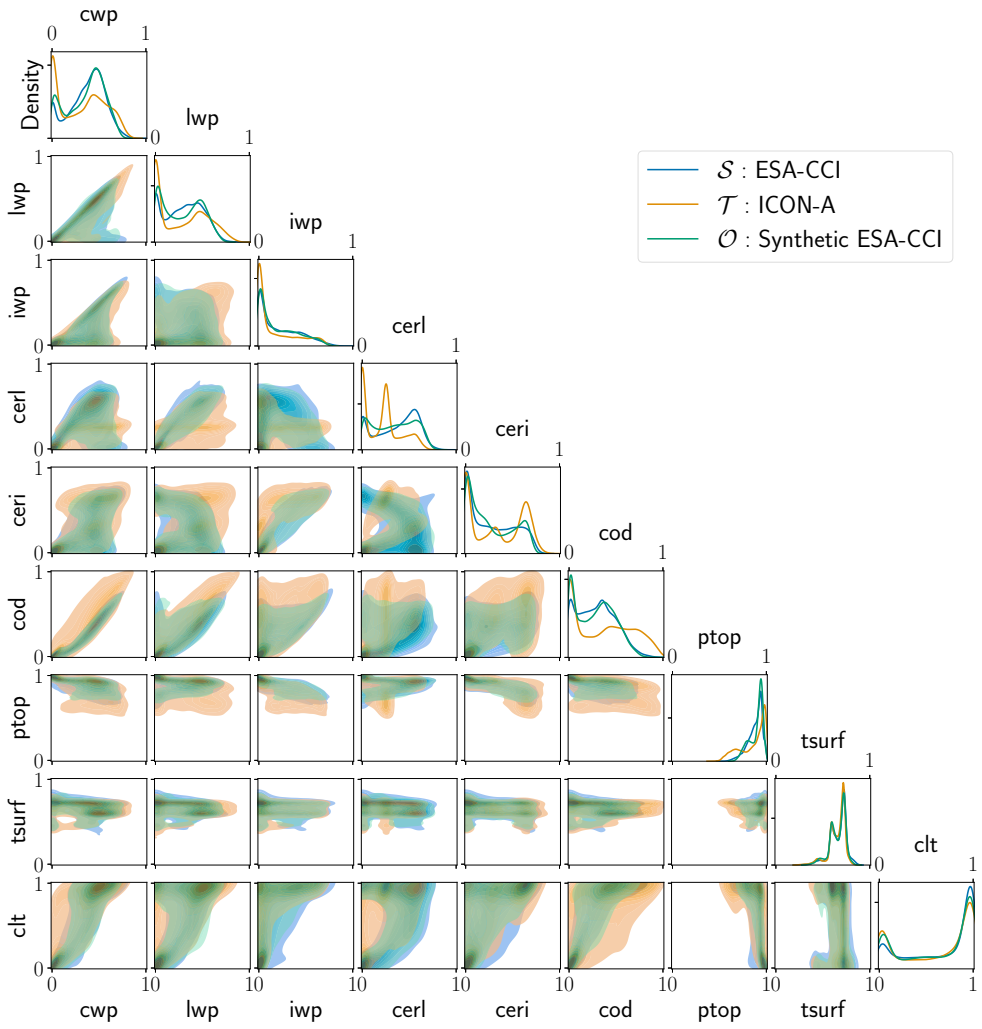


Figure E.1.: Complete version of Fig. **6.2** as reference for the distance metrics shown in Fig. **6.3**. The first eight variables are included in the DA, while the cloud fraction *clt* is diagnosed from them using $\mathbf{C}_{toa}$.

F. Models for COSP/WGAN-DA Comparison

Model	Simulated instruments	Ensemble-member	Time period	Horizontal resolution	Reference
CNRM-CM6-1	CALIPSO, ISCCP	r1i1p1f2	1979-2014	T127 grid; 250 km	Voldoire et al. (2019)
CNRM-ESM2-1	CALIPSO, ISCCP	r1i1p1f2	1979-2014	T127 grid; 250 km	Séférian et al. (2019)
GFDL-CM4	CALIPSO, ISCCP	r1i1p1f1	1979-2014	c96 grid; 250 km	Held et al. (2019)
HadGEM3-GC31-LL	ISCCP	r1i1p1f3	1979-2014	N96 grid; 250 km	Kuhlbrodt et al. (2018)
HadGEM3-GC31-MM	ISCCP	r2i1p1f3	1979-2014	N216 grid; 100 km	Kuhlbrodt et al. (2018)
IPSL-CM6A-LR	CALIPSO, ISCCP	r1i1p1f1	1979-2017	LMDZ grid; 250 km	Boucher et al. (2020)

Table F.1.: Models used to compute cloud cover changes induced by the COSP satellite simulator. All simulations are AMIP. Time periods are inclusive.

Acronyms

AI artificial intelligence 20
AMIP Atmospheric Model Intercomparison Project 40
AVHRR Advanced Very High Resolution Radiometer 36
CALIOP Cloud-Aerosol Lidar with Orthogonal Polarization 19
CALIPSO Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation 35
CAPE convective available potential energy 13
CC4CL Community Cloud retrieval for Climate 39
CCClim Cloud Class Climatology 3
CC-L 2B-CLDCLASS-LIDAR 35
cee cloud effective emissivity 36
cer effective radius of cloud particles 9
CERES Clouds and the Earth's Radiant Energy System 9
ceri effective radius of cloud ice particles 10
cerl effective radius of cloud water droplets 41
CIN convective inhibition 12
clt cloud fraction 31
CMIP Climate Model Intercomparison Project 33
CNN convolutional neural network 26
cod cloud optical depth 11
COSP Cloud Feedback Model Intercomparison Project Observation Simulator Package . 93
cph cloud top thermodynamic phase 36
CPR Cloud Profiling Radar 19
CRE cloud radiative effect 9
CS CloudSat 19
cwp cloud water path 18
DA domain adaptation 4
DL deep learning 3

List of Figures

List of Tables

References

Arias, P. et al. (2021). "Technical Summary". In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by V. Masson-Delmotte et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 33–144. DOI: 10.1017/9781009157896.002.

Arjovsky, M., S. Chintala, and L. Bottou (2017). "Wasserstein Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 214–223. URL: https://proceedings.mlr.press/v70/arjovsky17a.html.

Austin, R. T., A. J. Heymsfield, and G. L. Stephens (2009). "Retrieval of ice cloud microphysical parameters using the CloudSat millimeter-wave radar and temperature". In: *Journal of Geophysical Research: Atmospheres* 114.D8. DOI: 10.1029/2008jd010049.

Baum, B. A., W. P. Menzel, R. A. Frey, D. C. Tobin, R. E. Holz, S. A. Ackerman, A. K. Heidinger, and P. Yang (2012). "MODIS Cloud-Top Property Refinements for Collection 6". In: *Journal of Applied Meteorology and Climatology* 51.6, pp. 1145–1163. DOI: 10.1175/jamc-d-11-0203.1.

Bazi, Y., L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan (2021). "Vision Transformers for Remote Sensing Image Classification". In: *Remote Sensing* 13.3, p. 516. DOI: 10.3390/rs13030516.

Behrmann, J., W. Grathwohl, R. T. Q. Chen, D. Duvenaud, and J.-H. Jacobsen (2019). "Invertible residual networks". In: *International Conference on Machine Learning*. PMLR, pp. 573–582.

Beucler, T., M. Pritchard, S. Rasp, J. Ott, P. Baldi, and P. Gentine (2021). "Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems". In: *Physical Review Letters* 126.9, p. 098302. DOI: 10.1103/physrevlett.126.098302.

Blitzer, J., M. Dredze, and F. Pereira (2007). "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification". In: *Proceedings of the 45th annual meeting of the association of computational linguistics*, pp. 440–447.

Blitzer, J., R. McDonald, and F. Pereira (2006). "Domain adaptation with structural correspondence learning". In: *Proceedings of the 2006 conference on empirical methods in natural language processing*, pp. 120–128.

Bock, L. et al. (2020). "Quantifying Progress Across Different CMIP Phases With the ESMValTool". In: *Journal of Geophysical Research: Atmospheres* 125.21. DOI: 10.1029/2019jd032321.

Bodas-Salcedo, A., K. D. Williams, P. R. Field, and A. P. Lock (2012). "The Surface Downwelling Solar Radiation Surplus over the Southern Ocean in the Met Office Model: The Role of Midlatitude Cyclone Clouds". In: *Journal of Climate* 25.21, pp. 7467–7486. DOI: 10.1175/jcli-d-11-00702.1.

Bonneel, N., M. van de Panne, S. Paris, and W. Heidrich (2011). "Displacement interpolation using Lagrangian mass transport". In: *Proceedings of the 2011 SIGGRAPH Asia Conference*. ACM. DOI: 10.1145/2024156.2024192.

Bony, S., J.-L. Dufresne, H. L. Treut, J.-J. Morcrette, and C. Senior (2004). "On dynamic and thermodynamic components of cloud changes". In: *Climate Dynamics* 22.2-3, pp. 71–86. DOI: 10.1007/s00382-003-0369-6.

Bony, S. (2005). "Marine boundary layer clouds at the heart of tropical cloud feedback uncertainties in climate models". In: *Geophysical Research Letters* 32.20. DOI: 10.1029/2005gl023851.

— (2015). "Clouds, circulation and climate sensitivity". In: *Nature Geoscience* 8.26, pp. 261–268. DOI: DOI:10.1038/NGEO2398.

Bottou, L. (1998). "Online Learning and Stochastic Approximations". In: *Online learning in neural networks*.

Boucher, O. et al. (2020). "Presentation and Evaluation of the IPSL-CM6A-LR Climate Model". In: *Journal of Advances in Modeling Earth Systems* 12.7. DOI: 10.1029/2019ms002010.

Breiman, L. (1996). "Bagging predictors". In: *Machine Learning* 24.2, pp. 123–140. DOI: 10.1007/bf00058655.

— (2001). "Random Forests". In: *Machine Learning* 45.1, pp. 5–32. DOI: 10.1023/a:1010933404324.

Camps-Valls, G., D. Tuia, X. X. Zhu, and M. Reichstein (2021). *Deep Learning for the Earth Sciences. A Comprehensive Approach to Remote Sensing, Climate Science and Geosciences*. Wiley & Sons, Incorporated, John, p. 624. ISBN: 9781119646150.

Ceppi, P., D. T. McCoy, and D. L. Hartmann (2016). "Observational evidence for a negative shortwave cloud feedback in middle to high latitudes". In: *Geophysical Research Letters* 43.3, pp. 1331–1339. DOI: 10.1002/2015gl067499.

Chahine, M. T. (1974). "Remote Sounding of Cloudy Atmospheres. I. The Single Cloud Layer". In: *Journal of the Atmospheric Sciences* 31.1, pp. 233–243. DOI: 10.1175/1520-0469(1974)031<0233:rsocai>2.0.co;2.

Chen, Y. and A. D. D. Genio (2008). "Evaluation of tropical cloud regimes in observations and a general circulation model". In: *Climate Dynamics* 32.2-3, pp. 355–369. DOI: 10.1007/s00382-008-0386-6.

Cho, N., J. Tan, and L. Oreopoulos (2021). "Classifying planetary cloudiness with an updated set of MODIS Cloud Regimes". In: *Journal of Applied Meteorology and Climatology*. DOI: 10.1175/jamc-d-20-0247.1.

Choi, Y., M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo (2018). "StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation". In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE. DOI: 10.1109/cvpr.2018.00916.

Choi, Y., Y. Uh, J. Yoo, and J.-W. Ha (2020). "StarGAN v2: Diverse Image Synthesis for Multiple Domains". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Choromanska, A., M. Henaff, M. Mathieu, G. Ben Arous, and Y. LeCun (2015). "The Loss Surfaces of Multilayer Networks". In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, pp. 192–204. URL: https://proceedings.mlr.press/v38/choromanska15.html.

Copernicus Climate Change Service (2019). *ERA5-Land monthly averaged data from 2001 to present*. ECMWF. DOI: 10.24381/CDS.68D2BB30.

Coumou, D. and S. Rahmstorf (2012). "A decade of weather extremes". In: *Nature Climate Change* 2.7, pp. 491–496. DOI: 10.1038/nclimate1452.

Crueger, T. et al. (2018). "ICON-A, The Atmosphere Component of the ICON Earth System Model: II. Model Evaluation". In: *Journal of Advances in Modeling Earth Systems* 10.7, pp. 1638–1662. DOI: 10.1029/2017ms001233.

Cziczo, D. J. et al. (2013). "Clarifying the Dominant Sources and Mechanisms of Cirrus Cloud Formation". In: *Science* 340.6138, pp. 1320–1324. DOI: 10.1126/science.1234145.

Daume III, H. and D. Marcu (2006). "Domain Adaptation for Statistical Classifiers". In: *Journal of Artificial Intelligence Research* 26, pp. 101–126. DOI: 10.1613/jair.1872.

Delanoë, J. and R. J. Hogan (2010). "Combined CloudSat-CALIPSO-MODIS retrievals of the properties of ice clouds". In: *Journal of Geophysical Research* 115. DOI: 10.1029/2009jd012346.

Denby, L. (2020). "Discovering the Importance of Mesoscale Cloud Organization Through Unsupervised Classification". In: *Geophysical Research Letters* 47.1. DOI: 10.1029/2019gl085190.

Du, S., J. Lee, H. Li, L. Wang, and X. Zhai (2019). "Gradient Descent Finds Global Minima of Deep Neural Networks". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by K. Chaudhuri and R. Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1675–1685. URL: https://proceedings.mlr.press/v97/du19c.html.

Dubovik, O., G. L. Schuster, F. Xu, Y. Hu, H. Bösch, J. Landgraf, and Z. Li (2021). "Grand Challenges in Satellite Remote Sensing". In: *Frontiers in Remote Sensing* 2. DOI: 10.3389/frsen.2021.619818.

Dufresne, J.-L. and S. Bony (2008). "An Assessment of the Primary Sources of Spread of Global Warming Estimates from Coupled Atmosphere–Ocean Models". In: *Journal of Climate* 21.19, pp. 5135–5144. DOI: 10.1175/2008jcli2239.1.

Evan, A. T., A. K. Heidinger, and D. J. Vimont (2007). "Arguments against a physical long-term trend in global ISCCP cloud amounts". In: *Geophysical Research Letters* 34.4. DOI: 10.1029/2006gl028083.

Eyring, V. et al. (2021). "Human Influence on the Climate System". In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by V. Masson-Delmotte et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 423–552. DOI: 10.1017/9781009157896.005.

Eyring, V., S. Bony, G. A. Meehl, C. A. Senior, B. Stevens, R. J. Stouffer, and K. E. Taylor (2016). "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". In: *Geoscientific Model Development* 9.5, pp. 1937–1958. DOI: 10.5194/gmd-9-1937-2016.

Eyring, V. et al. (2019). "Taking climate model evaluation to the next level". In: *Nature Climate Change* 9.2, pp. 102–110. DOI: 10.1038/s41558-018-0355-y.

Fernando, B., A. Habrard, M. Sebban, and T. Tuytelaars (2013). "Unsupervised Visual Domain Adaptation Using Subspace Alignment". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Forster, P. et al. (2021). "The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity". In: *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by V. Masson-Delmotte et al. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press, pp. 923–1054. DOI: 10.1017/9781009157896.009.

François, B., S. Thao, and M. Vrac (2021). "Adjusting spatial dependence of climate model outputs with cycle-consistent adversarial networks". In: *Climate Dynamics* 57.11-12, pp. 3323–3353. DOI: 10.1007/s00382-021-05869-8.

Fulton, D. J., B. J. Clarke, and G. C. Hegerl (2023). "Bias correcting climate model simulations using unpaired image-to-image translation networks". In: *Artificial Intelligence for the Earth Systems*, pp. 1–56. DOI: 10.1175/AIES-D-22-0031.1.

Gates, W. L. (1992). "AMIP: The Atmospheric Model Intercomparison Project". In: *Bulletin of the American Meteorological Society* 73.12, pp. 1962–1970. DOI: 10.1175/1520-0477(1992)073<1962:atamip>2.0.co;2.

Gettelman, A. et al. (2020). "Simulating Observations of Southern Ocean Clouds and Implications for Climate". In: *Journal of Geophysical Research: Atmospheres* 125.21. DOI: 10.1029/2020jd032619.

Gettelman, A. and R. B. Rood (2016). *Demystifying Climate Models*. Springer Berlin Heidelberg. DOI: 10.1007/978-3-662-48959-8.

Gibson, P. B., S. E. Perkins-Kirkpatrick, P. Uotila, A. S. Pepler, and L. V. Alexander (2017). "On the use of self-organizing maps for studying climate extremes". In: *Journal of Geophysical Research: Atmospheres* 122.7, pp. 3891–3903. DOI: 10.1002/2016jd026256.

Giorgetta, M. A. et al. (2018). "ICON-A, the Atmosphere Component of the ICON Earth System Model: I. Model Description". In: *Journal of Advances in Modeling Earth Systems* 10.7, pp. 1613–1637. DOI: 10.1029/2017ms001242.

Goodfellow, I., Y. Bengio, and A. Courville (2016a). *Deep Learning*. MIT Press.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger. Vol. 27. Curran Associates, Inc. URL: http://www.github.com/goodfeli/adversarial.

— (2016b). "Generative Adversarial Nets". In: *Neural Information Processing Systems*. URL: http://www.github.com/goodfeli/adversarial.

Gordon, N. D., J. R. Norris, C. P. Weaver, and S. A. Klein (2005). "Cluster analysis of cloud regimes and characteristic dynamics of midlatitude synoptic systems in observations and a model". In: *Journal of Geophysical Research: Atmospheres* 110.D15. ISSN: 0148-0227. DOI: 10.1029/2004jd005027.

Gorooh, V. A., S. Kalia, P. Nguyen, K.-l. Hsu, S. Sorooshian, S. Ganguly, and R. Nemani (2020). "Deep Neural Network Cloud-Type Classification (DeepCTC) Model and Its Application in Evaluating PERSIANN-CCS". In: *Remote Sensing* 12.2, p. 316. DOI: 10.3390/rs12020316.

Grill, J.-B. et al. (2020). "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 21271–21284.

Groenke, B., L. Madaus, and C. Monteleoni (2020). "ClimAlign: Unsupervised statistical downscaling of climate variables via normalizing flows". In: *Proceedings of the 10th International Conference on Climate Informatics*. ACM. DOI: 10.1145/3429309.3429318.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville (2017). "Improved Training of Wasserstein GANs". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

Hahn, C. and S. Warren (2007). *A Gridded Climatology of Clouds over Land (1971-1996) and Ocean (1954-2008) from Surface Observations Worldwide (NDP-026E)*. en. Environmental System Science Data Infrastructure for a Virtual Ecosystem; Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States). DOI: 10.3334/CDIAC/CLI.NDP026E.

Haiden, T. (1997). "An analytical study of cumulus onset". In: *Quarterly Journal of the Royal Meteorological Society* 123.543, pp. 1945–1960. DOI: 10.1002/qj.49712354309.

Hansen, J. E. and L. D. Travis (1974). "Light scattering in planetary atmospheres". In: *Space Science Reviews* 16.4, pp. 527–610. DOI: 10.1007/bf00168069.

Harder, P., A. Hernandez-Garcia, V. Ramesh, Q. Yang, P. Sattigeri, D. Szwarcman, C. Watson, and D. Rolnick (2022). *Physics-Constrained Deep Learning for Climate Downscaling*. DOI: 10.48550/ARXIV.2208.05424.

Haynes, J. M., C. Jakob, W. B. Rossow, G. Tselioudis, and J. Brown (2011). "Major Characteristics of Southern Ocean Cloud Regimes and Their Effects on the Energy Budget". In: *Journal of Climate* 24.19, pp. 5061–5080. DOI: 10.1175/2011jcli4052.1.

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. DOI: 10.1109/cvpr.2016.90.

Held, I. M. et al. (2019). "Structure and Performance of GFDL's CM4.0 Climate Model". In: *Journal of Advances in Modeling Earth Systems* 11.11, pp. 3691–3727. DOI: 10.1029/2019ms001829.

Hersbach, H. et al. (2020). "The ERA5 global reanalysis". In: *Quarterly Journal of the Royal Meteorological Society* 146.730, pp. 1999–2049. DOI: 10.1002/qj.3803.

Heymsfield, A. J., S. Matrosov, and B. Baum (2003). "Ice Water Path–Optical Depth Relationships for Cirrus and Deep Stratiform Ice Cloud Layers". In: *Journal of Applied Meteorology* 42.10, pp. 1369–1390. DOI: 10.1175/1520-0450(2003)042<1369:iwpdrf>2.0.co;2.

Hines, C. O. (1997). "Doppler-spread parameterization of gravity-wave momentum deposition in the middle atmosphere. Part 1: Basic formulation". In: *Journal of Atmospheric and Solar-Terrestrial Physics* 59.4, pp. 371–386. DOI: 10.1016/s1364-6826(96)00079-x.

Hoffman, J., E. Tzeng, T. Park, J.-Y. Zhu, P. Isola, K. Saenko, A. A. Efros, and T. Darrell (2018). "CyCADA: Cycle-Consistent Adversarial Domain Adaptation". In: *Proceedings of the International Conference on Machine Learning 2018*.

Hornik, K., M. Stinchcombe, and H. White (1989). "Multilayer feedforward networks are universal approximators". In: *Neural Networks* 2.5, pp. 359–366. DOI: 10.1016/0893-6080(89)90020-8.

Hourdin, F. et al. (2017). "The Art and Science of Climate Model Tuning". In: *Bulletin of the American Meteorological Society* 98.3, pp. 589–602. DOI: 10.1175/bams-d-15-00135.1.

Hourdin, F. et al. (2021). "Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global". In: *Journal of Advances in Modeling Earth Systems* 13.6. DOI: 10.1029/2020ms002225.

Houze, R. A. (2014). "Nimbostratus and the Separation of Convective and Stratiform Precipitation". In: *International Geophysics*. Elsevier, pp. 141–163. DOI: 10.1016/b978-0-12-374266-7.00006-8.

Howard, L. (1803). "I. On the modifications of clouds, and on the principles of their production, suspension, and destruction; being the substance of an essay read before the Askesian Society in the session 1802–3". In: *The Philosophical Magazine* 17.65, pp. 5–11.

Huang, L., J. H. Jiang, Z. Wang, H. Su, M. Deng, and S. Massie (2015). "Climatology of cloud water content associated with different cloud types observed by A-Train satellites". In: *Journal of Geophysical Research: Atmospheres* 120.9, pp. 4196–4212. DOI: 10.1002/2014jd022779.

Huang, X., M.-Y. Liu, S. Belongie, and J. Kautz (2018). "Multimodal Unsupervised Image-to-Image Translation". In: *Computer Vision – ECCV 2018*. Springer International Publishing, pp. 179–196. DOI: 10.1007/978-3-030-01219-9_11.

Ioffe, S. and C. Szegedy (2015). "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by F. Bach and D. Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, pp. 448–456. URL: https://proceedings.mlr.press/v37/ioffe15.html.

IPCC (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Vol. In Press. Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press. DOI: 10.1017/9781009157896.

Isola, P., J.-Y. Zhu, T. Zhou, and A. A. Efros (2017). "Image-To-Image Translation With Conditional Adversarial Networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jakob, C. and G. Tselioudis (2003). "Objective identification of cloud regimes in the Tropical Western Pacific". In: *Geophysical Research Letters* 30.21. DOI: 10.1029/2003gl018367.

Jiang, J. H. et al. (2012). "Evaluation of cloud and water vapor simulations in CMIP5 climate models using NASA "A-Train" satellite observations". In: *Journal of Geophysical Research: Atmospheres* 117.D14. DOI: 10.1029/2011jd017237.

Jin, D., L. Oreopoulos, and D. Lee (2016). "Regime-based evaluation of cloudiness in CMIP5 models". In: *Climate Dynamics* 48.1-2, pp. 89–112. DOI: 10.1007/s00382-016-3064-0.

Kaps, **A.**, A. Lauer, G. Camps-Valls, P. Gentine, L. Gomez-Chova, and V. Eyring (2023a). "Machine-Learned Cloud Classes From Satellite Data for Process-Oriented Climate Model Evaluation". In: *IEEE Transactions on Geoscience and Remote Sensing* 61, pp. 1–15. DOI: 10.1109/TGRS.2023.3237008.

Kaps, **A.**, A. Lauer, and V. Eyring (2023b). *CCClim - A machine-learning powered cloud class climatology*. Zenodo. DOI: 10.5281/ZENODO.8369202.

Kaps, **A.**, A. Lauer, R. Kazeroni, M. Stengel, and V. Eyring (2023c). "Characterizing clouds with the CCClim dataset, a machine learning cloud class climatology". In: *Earth System Science Data Discussions*. in review. DOI: 10.5194/essd-2023-424.

Kawai, H., S. Yukimoto, T. Koshiro, N. Oshima, T. Tanaka, H. Yoshimura, and R. Nagasawa (2019). "Significant improvement of cloud representation in the global climate model MRI-ESM2". In: *Geoscientific Model Development* 12.7, pp. 2875–2897. DOI: 10.5194/gmd-12-2875-2019.

Kim, T., M. Cha, H. Kim, J. K. Lee, and J. Kim (2017). "Learning to Discover Cross-Domain Relations with Generative Adversarial Networks". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1857–1865. URL: https://proceedings.mlr.press/v70/kim17a.html.

Kingma, D. P. and J. L. Ba (2015). "Adam: A Method for Stochastic Optimization". In.

Kingma, D. P. and M. Welling (2013). *Auto-Encoding Variational Bayes*. DOI: 10.48550/ARXIV.1312.6114.

Kingma, D. P. and P. Dhariwal (2018). "Glow: Generative Flow with Invertible 1x1 Convolutions". In: *Advances in neural information processing systems, 31*.

Kodama, C., A. T. Noda, and M. Satoh (2012). "An assessment of the cloud signals simulated by NICAM using ISCCP, CALIPSO, and CloudSat satellite simulators". In: *Journal of Geophysical Research: Atmospheres* 117.D12. DOI: 10.1029/2011jd017317.

Kohonen, T. (1998). "The self-organizing map". In: *Neurocomputing* 21.1-3, pp. 1–6. DOI: 10.1016/s0925-2312(98)00030-7.

— (2013). "Essentials of the self-organizing map". In: *Neural Networks* 37, pp. 52–65. DOI: 10.1016/j.neunet.2012.09.018.

Kotsiantis, S. B. (2011). "Decision trees: a recent overview". In: *Artificial Intelligence Review* 39.4, pp. 261–283. DOI: 10.1007/s10462-011-9272-4.

Kramer, M. A. (1991). "Nonlinear principal component analysis using autoassociative neural networks". In: *AIChE Journal* 37.2, pp. 233–243. DOI: 10.1002/aic.690370209.

Kuhlbrodt, T. et al. (2018). "The Low-Resolution Version of HadGEM3 GC3.1: Development and Evaluation for Global Climate". In: *Journal of Advances in Modeling Earth Systems* 10.11, pp. 2865–2888. DOI: 10.1029/2018ms001370.

Kuma, P., F. A.-M. Bender, A. Schuddeboom, A. J. McDonald, and Ø. Seland (2023). "Machine learning of cloud types in satellite observations and climate models". In: *Atmospheric Chemistry and Physics* 23.1, pp. 523–549. DOI: 10.5194/acp-23-523-2023.

Kurihana, T., E. Moyer, R. Willett, D. Gilton, and I. Foster (2021). "Data-Driven Cloud Clustering via a Rotationally Invariant Autoencoder". In: *IEEE Transactions on Geoscience and Remote Sensing* 60, pp. 1–25. DOI: 10.1109/tgrs.2021.3098008.

Kurihana, T., E. J. Moyer, and I. T. Foster (2022). "AICCA: AI-Driven Cloud Classification Atlas". In: *Remote Sensing* 14.22, p. 5690. DOI: 10.3390/rs14225690.

L'Ecuyer, T. S. and J. H. Jiang (2010). "Touring the atmosphere aboard the A-Train". In: *Phys. Today* 63.7, pp. 36–41.

L'Ecuyer, T. S., Y. Hang, A. V. Matus, and Z. Wang (2019). "Reassessing the Effect of Cloud Type on Earth's Energy Balance in the Age of Active Spaceborne Observations. Part I: Top of Atmosphere and Surface". In: *Journal of Climate* 32.19, pp. 6197–6217. DOI: 10.1175/jcli-d-18-0753.1.

Lauer, A., L. Bock, B. Hassler, M. Schröder, and M. Stengel (2023). "Cloud Climatologies from Global Climate Models—A Comparison of CMIP5 and CMIP6 Models with Satellite Data". In: *Journal of Climate* 36.2, pp. 281–311. DOI: 10.1175/jcli-d-22-0181.1.

Lauer, A. et al. (2017). "Benchmarking CMIP5 models with a subset of ESA CCI Phase 2 data using the ESMValTool". In: *Remote Sensing of Environment* 203, pp. 9–39. DOI: 10.1016/j.rse.2017.01.007.

LeCun, Y. and Y. Bengio (1995). "Convolutional Networks for Images, Speech an Time-Series". In: *The handbook of brain theory and neural networks*.

LeCun, Y., Y. Bengio, and G. Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. DOI: 10.1038/nature14539.

Lee, J., R. C. Weger, S. K. Sengupta, and R. M. Welch (1990). "A neural network approach to cloud classification". In: *IEEE Transactions on Geoscience and Remote Sensing* 28.5, pp. 846–855. DOI: 10.1109/36.58972.

Lee, Y., C. D. Kummerow, and I. Ebert-Uphoff (2021). "Applying machine learning methods to detect convection using Geostationary Operational Environmental Satellite-16 (GOES-16) advanced baseline imager (ABI) data". In: *Atmospheric Measurement Techniques* 14.4, pp. 2699–2716. DOI: 10.5194/amt-14-2699-2021.

Li, J., J. Huang, K. Stamnes, T. Wang, Q. Lv, and H. Jin (2015). "A global survey of cloud overlap based on CALIPSO and CloudSat measurements". In: *Atmospheric Chemistry and Physics* 15.1, pp. 519–536. DOI: 10.5194/acp-15-519-2015.

Li, L., K. Jamieson, A. Rostamizadeh, E. Gonina, M. Hardt, B. Recht, and A. Talwalkar (2018). "A System for Massively Parallel Hyperparameter Tuning". In: DOI: 10.48550/ARXIV.1810.05934.

Liou, K. N. (2002). *Introduction to Atmospheric Radiation*. Elsevier Science & Technology Books, p. 583. ISBN: 9780080491677.

Liu, M.-Y., T. Breuel, and J. Kautz (2017). "Unsupervised Image-to-Image Translation Networks". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Long Beach, California, USA: Curran Associates Inc., pp. 700–708. ISBN: 9781510860964.

Liu, M.-Y. and O. Tuzel (2016). "Coupled Generative Adversarial Networks". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc.

Liu, S., M. Li, Z. Zhang, X. Cao, and T. S. Durrani (2020). "Ground-Based Cloud Classification Using Task-Based Graph Convolutional Network". In: *Geophysical Research Letters* 47.5. DOI: 10.1029/2020gl087338.

Lloyd, S. (1982). "Least squares quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2, pp. 129–137. DOI: 10.1109/tit.1982.1056489.

Loeb, N. G., B. A. Wielicki, D. R. Doelling, G. L. Smith, D. F. Keyes, S. Kato, N. Manalo-Smith, and T. Wong (2009). "Toward Optimal Closure of the Earth's Top-of-Atmosphere Radiation Budget". In: *Journal of Climate* 22.3, pp. 748–766. DOI: 10.1175/2008jcli2637.1.

Lohmann, U. and E. Roeckner (1996). "Design and performance of a new cloud microphysics scheme developed for the ECHAM general circulation model". In: *Climate Dynamics* 12.8, pp. 557–572. DOI: 10.1007/bf00207939.

Lott, F. (1999). "Alleviation of Stationary Biases in a GCM through a Mountain Drag Parameterization Scheme and a Simple Representation of Mountain Lift Forces". In: *Monthly Weather Review* 127.5, pp. 788–801. DOI: 10.1175/1520-0493(1999)127<0788:aosbia>2.0.co;2.

Malone, T. F. (1955). "Application of statistical models in weather prediction". In: *Proceedings of the National Academy of Sciences* 41.11, pp. 806–815. DOI: 10.1073/pnas.41.11.806.

Marais, W. J., R. E. Holz, J. S. Reid, and R. M. Willett (2020). "Leveraging spatial textures, through machine learning, to identify aerosols and distinct cloud types from multispectral observations". In: *Atmospheric Measurement Techniques* 13.10, pp. 5459–5480. DOI: 10.5194/amt-13-5459-2020.

Marchant, B., S. Platnick, K. Meyer, G. T. Arnold, and J. Riedi (2016). "MODIS Collection 6 shortwave-derived cloud phase classification algorithm and comparisons with CALIOP". In: *Atmospheric Measurement Techniques* 9.4, pp. 1587–1599. DOI: 10.5194/amt-9-1587-2016.

Marchant, B., S. Platnick, K. Meyer, and G. Wind (2020). "Evaluation of the MODIS Collection 6 multilayer cloud detection algorithm through comparisons with CloudSat Cloud Profiling Radar and CALIPSO CALIOP products". In: *Atmospheric Measurement Techniques* 13.6, pp. 3263–3275. DOI: 10.5194/amt-13-3263-2020.

Mateo-García, G., V. Laparra, D. López-Puigdollers, and L. Gómez-Chova (2021). "Cross-Sensor Adversarial Domain Adaptation of Landsat-8 and Proba-V Images for Cloud Detection". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. DOI: 10.1109/JSTARS.2020.3031741.

Mauritsen, T., G. Svensson, S. S. Zilitinkevich, I. Esau, L. Enger, and B. Grisogono (2007). "A Total Turbulent Energy Closure Model for Neutrally and Stably Stratified Atmospheric Boundary Layers". In: *Journal of the Atmospheric Sciences* 64.11, pp. 4113–4126. DOI: 10.1175/2007jas2294.1.

Mauritsen, T. et al. (2012). "Tuning the climate of a global model". In: *Journal of Advances in Modeling Earth Systems* 4.3, n/a–n/a. DOI: 10.1029/2012ms000154.

McDonald, A. J. and S. Parsons (2018). "A Comparison of Cloud Classification Methodologies: Differences Between Cloud and Dynamical Regimes". In: *Journal of Geophysical Research: Atmospheres* 123.19, pp. 11, 173–11, 193. DOI: 10.1029/2018jd028595.

McGarragh, G. R. et al. (2018). "The Community Cloud retrieval for CLimate (CC4CL) – Part 2: The optimal estimation approach". In: *Atmospheric Measurement Techniques* 11.6, pp. 3397–3431. DOI: 10.5194/amt-11-3397-2018.

McGovern, A., R. Lagerquist, D. J. Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith (2019). "Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning". In: *Bulletin of the American Meteorological Society* 100.11, pp. 2175–2199. DOI: 10.1175/bams-d-18-0195.1.

Meehl, G. A., C. A. Senior, V. Eyring, G. Flato, J.-F. Lamarque, R. J. Stouffer, K. E. Taylor, and M. Schlund (2020). "Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models". In: *Science Advances* 6.26. DOI: 10.1126/sciadv.aba1981.

Mirza, M. and S. Osindero (2014). *Conditional Generative Adversarial Nets*. DOI: 10.48550/ARXIV.1411.1784.

Nakajima, T. Y. and T. Nakajma (1995). "Wide-Area Determination of Cloud Microphysical Properties from NOAA AVHRR Measurements for FIRE and ASTEX Regions". In: *Journal of the Atmospheric Sciences* 52.23, pp. 4043–4059. DOI: 10.1175/1520-0469(1995)052<4043:wadocm>2.0.co;2.

Nakajima, T. and M. D. King (1990). "Determination of the Optical Thickness and Effective Particle Radius of CLouds from Reflected Solar Radiation Measurements Part I: Theory". In: *Journal of the Atmospheric Sciences*.

Nordeng, T. E. (1994). "Extended versions of the extended parametrization scheme at ECMWF and their impact on the mean and transient activity of the model in the tropics". In: *ECMWF Tech. Memo.*

Norris, J. R. (1998). "Low Cloud Type over the Ocean from Surface Observations. Part II: Geographical and Seasonal Variations". In: *Journal of Climate* 11.3, pp. 383–403. DOI: 10.1175/1520-0442(1998)011<0383:lctoto>2.0.co;2.

Oreopoulos, L., N. Cho, D. Lee, and S. Kato (2016). "Radiative effects of global MODIS cloud regimes". In: *Journal of Geophysical Research: Atmospheres* 121.5, pp. 2299–2317. DOI: 10.1002/2015jd024502.

Oreopoulos, L., N. Cho, D. Lee, S. Kato, and G. J. Huffman (2014). "An examination of the nature of global MODIS cloud regimes". In: *Journal of Geophysical Research: Atmospheres* 119.13, pp. 8362–8383. DOI: 10.1002/2013jd021409.

Oreopoulos, L. and W. B. Rossow (2011). "The cloud radiative effects of International Satellite Cloud Climatology Project weather states". In: *Journal of Geophysical Research* 116.D12. DOI: 10.1029/2010jd015472.

Otto, F. E. (2023). "Attribution of Extreme Events to Climate Change". In: *Annual Review of Environment and Resources* 48.1. DOI: 10.1146/annurev-environ-112621-083538.

Pan, B., G. J. Anderson, A. Goncalves, D. D. Lucas, C. J. W. Bonfils, J. Lee, Y. Tian, and H.-Y. Ma (2021). "Learning to Correct Climate Projection Biases". In: *Journal of Advances in Modeling Earth Systems* 13.10. DOI: 10.1029/2021MS002509.

Pang, Y., J. Lin, T. Qin, and Z. Chen (2022). "Image-to-Image Translation: Methods and Applications". In: *IEEE Transactions on Multimedia* 24, pp. 3859–3881. DOI: 10.1109/TMM.2021.3109419.

Pincus, R., S. Platnick, S. A. Ackerman, R. S. Hemler, and R. J. P. Hofmann (2012). "Reconciling Simulated and Observed Views of Clouds: MODIS, ISCCP, and the Limits of Instrument Simulators". In: *Journal of Climate* 25.13, pp. 4699–4720. DOI: 10.1175/jcli-d-11-00267.1.

Pincus, R. and B. Stevens (2013). "Paths to accuracy for radiation parameterizations in atmospheric models". In: *Journal of Advances in Modeling Earth Systems* 5.2, pp. 225–233. DOI: 10.1002/jame.20027.

Plant, R. S. and J.-I. Yano (2014). *Parameterization of Atmospheric Convection*. IMPERIAL COLLEGE PRESS. DOI: 10.1142/p1005.

Platnick, S., M. D. King, S. A. Ackerman, W. P. Menzel, B. A. Baum, J. C. Riedi, and R. A. Frey (2003). "The MODIS cloud products: algorithms and examples from terra". In: *IEEE*

Transactions on Geoscience and Remote Sensing 41.2, pp. 459–473. DOI: 10.1109/tgrs.2002.808301.

Platnick, S. et al. (2017). "The MODIS Cloud Optical and Microphysical Products: Collection 6 Updates and Examples From Terra and Aqua". In: *IEEE Transactions on Geoscience and Remote Sensing* 55.1, pp. 502–525. DOI: 10.1109/tgrs.2016.2610522.

Raddatz, T. J. et al. (2007). "Will the tropical land biosphere dominate the climate–carbon cycle feedback during the twenty-first century?" In: *Climate Dynamics* 29.6, pp. 565–574. DOI: 10.1007/s00382-007-0247-8.

Ramponi, A. and B. Plank (2020). *Neural Unsupervised Domain Adaptation in NLP—A Survey*. DOI: 10.48550/ARXIV.2006.00632. URL: https://arxiv.org/abs/2006.00632.

Rasp, S., M. S. Pritchard, and P. Gentine (2018). "Deep learning to represent subgrid processes in climate models". In: *Proceedings of the National Academy of Sciences* 115.39, pp. 9684–9689. DOI: 10.1073/pnas.1810286115.

Rasp, S., H. Schulz, S. Bony, and B. Stevens (2020). "Combining Crowdsourcing and Deep Learning to Explore the Mesoscale Organization of Shallow Convection". In: *Bulletin of the American Meteorological Society* 101.11, E1980–E1995. DOI: 10.1175/bams-d-19-0324.1.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat (2019). "Deep learning and process understanding for data-driven Earth system science". In: *Nature* 566.7743, pp. 195–204. DOI: 10.1038/s41586-019-0912-1.

Rosenblatt, F. (1961). *Perceptrons and the Theory of Brain Mechanisms*. Tech. rep. AGENCY. Cornell Aeronautical Laboratory.

Rossow, W. B. and R. A. Schiffer (1999). "Advances in Understanding Clouds from ISCCP". In: *Bulletin of the American Meteorological Society* 80.11, pp. 2261–2287. DOI: 10.1175/1520-0477(1999)080<2261:aiucfi>2.0.co;2.

Rossow, W. B., A. W. Walker, and L. C. Garder (1993). "Comparison of ISCCP and Other Cloud Amounts". In: *Journal of Climate* 6.12, pp. 2394–2418. DOI: 10.1175/1520-0442(1993)006<2394:coiaoc>2.0.co;2.

Saenko, K., B. Kulis, M. Fritz, and T. Darrell (2010). "Adapting visual category models to new domains". In: *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*. Springer, pp. 213–226.

Sasaki, H., C. G. Willcocks, and T. P. Breckon (2021). "UNIT-DDPM: UNpaired Image Translation with Denoising Diffusion Probabilistic Models". In: DOI: 10.48550/ARXIV.2104.05358.

Sassen, K. and Z. Wang (2011). "The Clouds of the Middle Troposphere: Composition, Radiative Impact, and Global Distribution". In: *Surveys in Geophysics* 33.3-4, pp. 677–691. DOI: 10.1007/s10712-011-9163-x.

Sassen, K., Z. Wang, and D. Liu (2008). "Global distribution of cirrus clouds from CloudSat/Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO) measurements". In: *Journal of Geophysical Research* 113. DOI: 10.1029/2008jd009972.

Schapire, R. E., Y. Freund, P. Bartlett, and W. S. Lee (1998). "Boosting the margin: a new explanation for the effectiveness of voting methods". In: *The Annals of Statistics* 26.5. DOI: 10.1214/aos/1024691352.

Schiffer, R. A. and W. B. Rossow (1983). "The International Satellite Cloud Climatology Project (ISCCP): The First Project of the World Climate Research Programme". In: *Bulletin of the American Meteorological Society* 64.7, pp. 779–784. DOI: 10.1175/1520-0477-64.7.779.

Schlund, M., V. Eyring, G. Camps-Valls, P. Friedlingstein, P. Gentine, and M. Reichstein (2020a). "Constraining Uncertainty in Projected Gross Primary Production With Machine Learning". In: *Journal of Geophysical Research: Biogeosciences* 125.11. DOI: 10.1029/2019jg005619.

Schlund, M., A. Lauer, P. Gentine, S. C. Sherwood, and V. Eyring (2020b). "Emergent constraints on equilibrium climate sensitivity in CMIP5: do they hold for CMIP6?" In: *Earth System Dynamics* 11.4, pp. 1233–1258. DOI: 10.5194/esd-11-1233-2020.

Schmidt, G. A. et al. (2017). "Practice and philosophy of climate model tuning across six US modeling centers". In: *Geoscientific Model Development* 10.9, pp. 3207–3223. DOI: 10.5194/gmd-10-3207-2017.

Schneider, T., C. M. Kaul, and K. G. Pressel (2019). "Possible climate transitions from breakup of stratocumulus decks under greenhouse warming". In: *Nature Geoscience* 12.3, pp. 163–167. DOI: 10.1038/s41561-019-0310-1.

Schneider, T., J. Teixeira, C. S. Bretherton, F. Brient, K. G. Pressel, C. Schär, and A. P. Siebesma (2017). "Climate goals and computing the future of clouds". In: *Nature Climate Change* 7.1, pp. 3–5. DOI: 10.1038/nclimate3190.

Schuddeboom, A. J. and A. J. McDonald (2021). "The Southern Ocean Radiative Bias, Cloud Compensating Errors, and Equilibrium Climate Sensitivity in CMIP6 Models". In: *Journal of Geophysical Research: Atmospheres* 126.22. DOI: 10.1029/2021jd035310.

Schuddeboom, A., A. J. McDonald, O. Morgenstern, M. Harvey, and S. Parsons (2018). "Regional Regime-Based Evaluation of Present-Day General Circulation Model Cloud Simulations Using Self-Organizing Maps". In: *Journal of Geophysical Research: Atmospheres* 123.8, pp. 4259–4272. DOI: 10.1002/2017jd028196.

Schulzweida, U. (2023). *CDO User Guide*. Version 2.3.0. DOI: 10.5281/zenodo.10020800. URL: https://doi.org/10.5281/zenodo.10020800.

Sedlar, J., L. D. Riihimaki, K. Lantz, and D. D. Turner (2021). "Development of a Random-Forest Cloud-Regime Classification Model Based on Surface Radiation and Cloud Products". In: *Journal of Applied Meteorology and Climatology* 60.4, pp. 477–491. DOI: 10.1175/jamc-d-20-0153.1.

Séférian, R. et al. (2019). "Evaluation of CNRM Earth System Model, CNRM-ESM2-1: Role of Earth System Processes in Present-Day and Future Climate". In: *Journal of Advances in Modeling Earth Systems* 11.12, pp. 4182–4227. DOI: 10.1029/2019ms001791.

Sherwood, S. C. et al. (2020). "An Assessment of Earth's Climate Sensitivity Using Multiple Lines of Evidence". In: *Reviews of Geophysics* 58.4. DOI: 10.1029/2019rg000678.

Sherwood, S. C., S. Bony, and J.-L. Dufresne (2014). "Spread in model climate sensitivity traced to atmospheric convective mixing". In: *Nature* 505.7481, pp. 37–42. DOI: 10.1038/nature12829.

Siebesma, A. P., S. Bony, C. Jakob, and B. Stevens, eds. (2020). *Clouds and Climate*. Cambridge University Press. DOI: 10.1017/9781107447738.

Smith, D. M. et al. (2015). "Earth's energy imbalance since 1960 in observations and CMIP5 models". In: *Geophysical Research Letters* 42.4, pp. 1205–1213. DOI: 10.1002/2014gl062669.

Smith, L. N. (2017). "Cyclical Learning Rates for Training Neural Networks". In: *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE. DOI: 10.1109/wacv.2017.58.

Smith, W. L. and C. M. R. Platt (2023). "Comparison of Satellite-Deduced Cloud Heights with Indications from Radiosonde and Ground-Based Laser Measurements". In: *Journal of Applied Meteorology (1962-1982)* 17.12, pp. 1796–1802.

Spänkuch, D., O. Hellmuth, and U. Görsdorf (2022). "What Is a Cloud? Toward a More Precise Definition". In: *Bulletin of the American Meteorological Society* 103.8, E1894–E1929. DOI: 10.1175/bams-d-21-0032.1.

Spreitzer, E. J., M. P. Marschalik, and P. Spichtinger (2017). "Subvisible cirrus clouds – a dynamical system approach". In: *Nonlinear Processes in Geophysics* 24.3, pp. 307–328. DOI: 10.5194/npg-24-307-2017.

Srivastava, N., G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1, pp. 1929–1958.

Stengel, M., O. Sus, S. Stapelberg, S. Finkensieper, B. Würzler, D. Philipp, R. Hollmann, and C. Poulsen (2019). *ESA Cloud_cci cloud property datasets retrieved from passive satellite sensors: AVHRR-PM L3C/L3U cloud products - Version 3.0*. Deutscher Wetterdienst (DWD). DOI: 10.5676/DWD/ESA_CLOUD_CCI/AVHRR-PM/V003.

Stengel, M. et al. (2017). "Cloud property datasets retrieved from AVHRR, MODIS, AATSR and MERIS in the framework of the Cloud_cci project". In: *Earth System Science Data* 9.2, pp. 881–904. DOI: 10.5194/essd-9-881-2017.

Stengel, M. et al. (2020). "Cloud_cci Advanced Very High Resolution Radiometer post meridiem (AVHRR-PM) dataset version 3: 35-year climatology of global cloud and radiation properties". In: *Earth System Science Data* 12.1, pp. 41–60. DOI: 10.5194/essd-12-41-2020.

Stephens, G. L. (1978). "Radiation Profiles in Extended Water Clouds. II: Parameterization Schemes". In: *Journal of the Atmospheric Sciences* 35.11, pp. 2123–2132. DOI: 10.1175/1520-0469(1978)035<2123:rpiewc>2.0.co;2.

Stephens, G., D. Winker, J. Pelon, C. Trepte, D. Vane, C. Yuhas, T. L'Ecuyer, and M. Lebsock (2018). "CloudSat and CALIPSO within the A-Train: Ten Years of Actively Observing the Earth System". In: *Bulletin of the American Meteorological Society* 99.3, pp. 569–581. DOI: 10.1175/bams-d-16-0324.1.

Stephens, G. L. (2005). "Cloud Feedbacks in the Climate System: A Critical Review". In: *Journal of Climate* 18.2, pp. 237–273. DOI: 10.1175/jcli-3243.1.

Stephens, G. L. and C. D. Kummerow (2007). "The Remote Sensing of Clouds and Precipitation from Space: A Review". In: *Journal of the Atmospheric Sciences* 64.11, pp. 3742–3765. DOI: 10.1175/2006jas2375.1.

Stephens, G. L. et al. (2002). "The CloudSat Mission and the A-Train". In: *Bulletin of the American Meteorological Society* 83.12, pp. 1771–1790. DOI: 10.1175/bams-83-12-1771.

Stevens, B. et al. (2021). "EUREC^{4}A". In: *Earth System Science Data* 13.8, pp. 4067–4119. DOI: 10.5194/essd-13-4067-2021. URL: https://essd.copernicus.org/articles/13/4067/2021/.

Stevens, B. and S. Bony (2013). "What Are Climate Models Missing?" In: *Science* 340.6136, pp. 1053–1054. DOI: 10.1126/science.1237554.

Stevens, B. et al. (2016). "The Barbados Cloud Observatory: Anchoring Investigations of Clouds and Circulation on the Edge of the ITCZ". In: *Bulletin of the American Meteorological Society* 97.5, pp. 787–801. DOI: 10.1175/bams-d-14-00247.1.

Stevens, B. et al. (2019). "Sugar, gravel, fish and flowers: Mesoscale cloud patterns in the trade winds". In: *Quarterly Journal of the Royal Meteorological Society* 146.726, pp. 141–152. DOI: 10.1002/qj.3662.

Stubenrauch, C. J., A. Chédin, G. Rädel, N. A. Scott, and S. Serrar (2006). "Cloud Properties and Their Seasonal and Diurnal Variability from TOVS Path-B". In: *Journal of Climate* 19.21, pp. 5531–5553. DOI: 10.1175/jcli3929.1.

Stubenrauch, C. J., A. G. Feofilov, S. E. Protopapadaki, and R. Armante (2017). "Cloud climatologies from the infrared sounders AIRS and IASI: strengths and applications". In: *Atmospheric Chemistry and Physics* 17.22, pp. 13625–13644. DOI: 10.5194/acp-17-13625-2017.

Sundqvist, H., E. Berge, and J. E. Kristjánsson (1989). "Condensation and Cloud Parameterization Studies with a Mesoscale Numerical Weather Prediction Model". In: *Monthly Weather Review* 117.8, pp. 1641–1657. DOI: 10.1175/1520-0493(1989)117<1641:cacpsw>2.0.co;2.

Sus, O. et al. (2018). "The Community Cloud retrieval for CLimate (CC4CL) – Part 1: A framework applied to multiple satellite imaging sensors". In: *Atmospheric Measurement Techniques* 11.6, pp. 3373–3396. DOI: 10.5194/amt-11-3373-2018.

Swales, D. J., R. Pincus, and A. Bodas-Salcedo (2018). "The Cloud Feedback Model Intercomparison Project Observational Simulator Package: Version 2". In: *Geoscientific Model Development* 11.1, pp. 77–81. DOI: 10.5194/gmd-11-77-2018.

Tselioudis, G., W. Rossow, Y. Zhang, and D. Konsta (2013). "Global Weather States and Their Properties from Passive and Active Satellite Cloud Retrievals". In: *Journal of Climate* 26.19, pp. 7734–7746. DOI: 10.1175/jcli-d-13-00024.1.

Tselioudis, G., W. B. Rossow, C. Jakob, J. Remillard, D. Tropf, and Y. Zhang (2021). "Evaluation of clouds, radiation, and precipitation in CMIP6 models using global weather states

derived from ISCCP-H cloud property data". In: *Journal of Climate*, pp. 1–42. DOI: 10.1175/jcli-d-21-0076.1.

Tsushima, Y., M. A. Ringer, M. J. Webb, and K. D. Williams (2012). "Quantitative evaluation of the seasonal variations in climate model cloud regimes". In: *Climate Dynamics* 41.9-10, pp. 2679–2696. DOI: 10.1007/s00382-012-1609-4.

Tuia, D., C. Persello, and L. Bruzzone (2016). "Domain Adaptation for the Classification of Remote Sensing Data: An Overview of Recent Advances". In: *IEEE Geoscience and Remote Sensing Magazine* 4.2, pp. 41–57. DOI: 10.1109/MGRS.2016.2548504.

Turbeville, S. M., J. M. Nugent, T. P. Ackerman, C. S. Bretherton, and P. N. Blossey (2022). "Tropical Cirrus in Global Storm-Resolving Models: 2. Cirrus Life Cycle and Top-of-Atmosphere Radiative Fluxes". In: *Earth and Space Science* 9.2. DOI: 10.1029/2021ea001978.

Tzallas, V., A. Hünerbein, M. Stengel, J. F. Meirink, N. Benas, J. Trentmann, and A. Macke (2022). "CRAAS: A European Cloud Regime dAtAset Based on the CLAAS-2.1 Climate Data Record". In: *Remote Sensing* 14.21, p. 5548. DOI: 10.3390/rs14215548.

Tzeng, E., J. Hoffman, N. Zhang, K. Saenko, and T. Darrell (2014). "Deep Domain Confusion: Maximizing for Domain Invariance". In: DOI: 10.48550/arXiv.1412.3474.

Ulyanov, D., A. Vedaldi, and V. Lempitsky (2016). *Instance Normalization: The Missing Ingredient for Fast Stylization*. DOI: 10.48550/ARXIV.1607.08022.

Unidata/University Corporation For Atmospheric Research (2003). *Historical Unidata Internet Data Distribution (IDD) Global Observational Data*. en. DOI: 10.5065/9235-WJ24.

Vignesh, P. P., J. H. Jiang, P. Kishore, H. Su, T. Smay, N. Brighton, and I. Velicogna (2020). "Assessment of CMIP6 Cloud Fraction and Comparison with Satellite Observations". In: *Earth and Space Science* 7.2. DOI: 10.1029/2019ea000975.

Voldoire, A. et al. (2019). "Evaluation of CMIP6 DECK Experiments With CNRM-CM6-1". In: *Journal of Advances in Modeling Earth Systems* 11.7, pp. 2177–2213. DOI: 10.1029/2019ms001683.

Wall, C. J., T. Storelvmo, J. R. Norris, and I. Tan (2022). "Observational Constraints on Southern Ocean Cloud-Phase Feedback". In: *Journal of Climate* 35.15, pp. 5087–5102. DOI: 10.1175/jcli-d-21-0812.1.

Walther, A., W. Straka, and A. K. Heidinger (2013). "NOAA NESDIS Center for Satellite Applications and Research ABI Algorithm Theoretical Basis Document For Daytime Cloud Optical and Microphysical Properties (DCOMP)". In: June, pp. 1–66. URL: https://www.goes-r.gov/products/baseline-cloud-opt-depth.html.

Wang, M. and W. Deng (2018). "Deep visual domain adaptation: A survey". In: *Neurocomputing* 312, pp. 135–153. DOI: 10.1016/j.neucom.2018.05.083.

Wang, T., E. J. Fetzer, S. Wong, B. H. Kahn, and Q. Yue (2016). "Validation of MODIS cloud mask and multilayer flag using CloudSat-CALIPSO cloud profiles and a cross-reference of their cloud classifications". In: *Journal of Geophysical Research: Atmospheres* 121.19. DOI: 10.1002/2016jd025239.

Wang, Z., A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli (2004). "Image Quality Assessment: From Error Visibility to Structural Similarity". In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612. DOI: 10.1109/tip.2003.819861.

Wang, Z. (2019a). *CloudSat 2B-CLDCLASS-LIDAR Product ProcessDescription and Interface Control Document*. Version p1_R05. URL: https://www.cloudsat.cira.colostate.edu/data-products/2b-cldclass-lidar.

— (2019b). *Level 2 Cloud Scenario ClassificationProduct Process Description andInterface Control Document. 2B-CLDCLASS*. Version P1_R05. CloudSat ProjectA NASA Earth System Science Pathfinder Mission.

Wehr, T. et al. (2023). "The EarthCARE mission – science and system overview". In: *Atmospheric Measurement Techniques* 16.15, pp. 3581–3608. DOI: 10.5194/amt-16-3581-2023.

Williams, K. D. and G. Tselioudis (2007). "GCM intercomparison of global cloud regimes: present-day evaluation and climate change response". In: *Climate Dynamics* 29.2-3, pp. 231–250. DOI: 10.1007/s00382-007-0232-2.

Williams, K. D. and M. J. Webb (2008). "A quantitative performance assessment of cloud regimes in climate models". In: *Climate Dynamics* 33.1, pp. 141–157. DOI: 10.1007/s00382-008-0443-1.

Williams, K. D., C. A. Senior, A. Slingo, and J. F. B. Mitchell (2005). "Towards evaluating cloud response to climate change using clustering technique identification of cloud regimes". In: *Climate Dynamics* 24.7-8, pp. 701–719. DOI: 10.1007/s00382-004-0512-z.

Winker, D. M., J. R. Pelon, and M. P. McCormick (2003). "The CALIPSO mission: spaceborne lidar for observation of aerosols and clouds". In: *SPIE Proceedings*. Ed. by U. N. Singh, T. Itabe, and Z. Liu. SPIE. DOI: 10.1117/12.466539.

WMO (2023). *International Cloud Atlas*. URL: https://cloudatlas.wmo.int/ (visited on 2023-04-20).

Wood, R. (2012). "Stratocumulus Clouds". In: *Monthly Weather Review* 140.8, pp. 2373–2423. DOI: 10.1175/mwr-d-11-00121.1.

Xu, M., M. Wu, K. Chen, C. Zhang, and J. Guo (2022). "The Eyes of the Gods: A Survey of Unsupervised Domain Adaptation Methods Based on Remote Sensing Data". In: *Remote Sensing* 14.17, p. 4380. DOI: 10.3390/rs14174380.

Young, A. H., K. R. Knapp, A. Inamdar, W. Hankins, and W. B. Rossow (2018). "The International Satellite Cloud Climatology Project H-Series climate data record product". In: *Earth System Science Data* 10.1, pp. 583–593. DOI: 10.5194/essd-10-583-2018.

Zantedeschi, V., F. Falasca, A. Douglas, R. Strange, M. J. Kusner, and D. Watson-Parris (2019). "Cumulo: A Dataset for Learning Cloud Classes". In: *Tackling Climate Change with Machine Learning Workshop, NeurIPS*.

Zelinka, M. D., S. A. Klein, Y. Qin, and T. A. Myers (2022a). "Evaluating Climate Models' Cloud Feedbacks Against Expert Judgment". In: *Journal of Geophysical Research: Atmospheres* 127.2. DOI: 10.1029/2021jd035198.

Zelinka, M. D., T. A. Myers, D. T. McCoy, S. Po-Chedley, P. M. Caldwell, P. Ceppi, S. A. Klein, and K. E. Taylor (2020). "Causes of Higher Climate Sensitivity in CMIP6 Models". In: *Geophysical Research Letters* 47.1. DOI: 10.1029/2019gl085782.

Zelinka, M. D., D. A. Randall, M. J. Webb, and S. A. Klein (2017). "Clearing clouds of uncertainty". In: *Nature Climate Change* 7.10, pp. 674–678. DOI: 10.1038/nclimate3402.

Zelinka, M. D., I. Tan, L. Oreopoulos, and G. Tselioudis (2022b). "Detailing cloud property feedbacks with a regime-based decomposition". In: *Climate Dynamics*. DOI: 10.1007/s00382-022-06488-7.

Zhang, C., X. Zhuge, and F. Yu (2019). "Development of a high spatiotemporal resolution cloud-type classification approach using Himawari-8 and CloudSat". In: *International Journal of Remote Sensing* 40.16, pp. 6464–6481. DOI: 10.1080/01431161.2019.1594438.

Zhang, J., P. Liu, F. Zhang, and Q. Song (2018). "CloudNet: Ground-Based Cloud Classification With Deep Convolutional Neural Network". In: *Geophysical Research Letters* 45.16, pp. 8665–8672. DOI: 10.1029/2018gl077787.

Zhang, M. H. (2005). "Comparing clouds and their seasonal variations in 10 atmospheric general circulation models with satellite measurements". In: *Journal of Geophysical Research* 110.D15. DOI: 10.1029/2004jd005021.

Zhou, K., Z. Liu, Y. Qiao, T. Xiang, and C. C. Loy (2022). "Domain Generalization: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20. DOI: 10.1109/tpami.2022.3195549.

Zhu, J.-Y., T. Park, P. Isola, and A. A. Efros (2017). "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *Proceedings of the IEEE international conference on computer vision*.

Code and Data Availability

CCClim can be downloaded from zenodo without restrictions under DOI:10.5281/zenodo.8369201 (Kaps et al., 2023b). The ESA-CCI dataset and related information are available under DOI:10.5676/DWD/ESA_CLOUD_CCI/AVHRR-PM/V003 (Stengel et al., 2019). The code related to the Random Forest produced in Chapter **4** can be found at DOI:10.5281/zenodo.7248773. The code for the production of CCClim and the analysis shown in Chapter **5** is available under DOI:10.5281/zenodo.10279991. The code for training and analyzing the domain adaptation models is released under DOI:10.5281/zenodo.10284130. The CUMULO data was downloaded from https://github.com/FrontierDevelopmentLab/CUMULO [1]. The ERA5 data was downloaded from the Copernicus Climate Change Service (C3S) (2023) (Copernicus Climate Change Service, 2019). The results contain modified Copernicus Climate Change Service information 2020. Neither the European Commission nor ECMWF is responsible for any use that may be made of the Copernicus information or data it contains.

[1] Last accessed 27th November 2023

Acknowledgments

This research was funded by the European Research Council (ERC) Synergy Grant "Understanding and modeling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant Agreement No. 855187). This work used resources of both, the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1179 and the supercomputer JUWELS at the Juelich Supercomputing Centre (JSC) under the Earth System Modelling Project (ESM). The author gratefully acknowledges the Leibniz Supercomputing Centre for funding this project by providing computing time on its Linux-Cluster.

I would like to extend my heartfelt appreciation to my supervisor and first examiner, Prof. Dr. Veronika Eyring, for her insightful guidance, support, and expert advice throughout the entire dissertation process. Her mentorship has been instrumental in shaping the quality and depth of this research.
I am also immensely grateful to my scientific advisor, Dr. Axel Lauer, who truly kept this research on track by helping me to see the more positive sides of sometimes unsatisfying results. His invaluable contributions, constructive feedback, and commitment to our research have no doubt significantly improved its quality. The development and execution of this dissertation have significantly been enriched by his expertise.
For their constructive feedback and guidance over the years as well as their contribution as co-authors, I would like to thank Prof. Dr. Luis Gómez-Chova, Prof. Dr. Gustau Camps-Valls and Prof. Dr. Pierre Gentine.
Also, my appreciation goes to Prof. Dr. Hartmut Bösch for agreeing to be the second examiner of this dissertation.
Dr. Rémi Kazeroni deserves my thanks for contributing to our article by dealing with the ICON-A model for me and for generally providing competent advice and support.
My sincere thanks go to Dr. Martin Stengel for agreeing to be a co-author and in this capacity providing invaluable feedback to our research.
Furthermore, I want to express my thanks to the entire PA-EVA department for providing an intellectually stimulating and most importantly cheerful environment. The collective efforts of all department members have provided me with a platform for growth and learning in which it was a pleasure to work and which I hope to find again in the future.
I would also like to thank Manuel Schlund for providing the LaTeX source code as a template.

The code for his dissertation, which constitutes parts of the code used to produce this document, is published on GitHub `https://github.com/schlunma/dissertation/`[1].
To everyone mentioned above, thank you for your support, encouragement, and the positive impact you have had on my academic journey.
Also, thanks to the office coffee machine for its dutiful service over the years.

[1] Last accessed 5$^{\text{th}}$ of December 2023