Understanding and Modelling Convection with Machine Learning

Doctoral Dissertation of

Gunnar Behrens

2024

Understanding and Modelling Convection with Machine Learning

Am Institut für Umweltphysik
vom Fachbereich für Physik und Elektrotechnik
der Universität Bremen

zur Erlangung des akademisches Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation

von

Gunnar Behrens

wohnhaft in Bremen

Erster Gutachter:	Prof. Dr. Veronika Eyring
Zweiter Gutachter:	Prof. Dr. Pierre Gentine
Eingereicht am:	February 8th 2024
Tag des Promotionskolloquiums:	May 10th 2024

Note: This published version of the doctoral dissertation includes few modifications compared to the submitted version. These are the result of linguistic optimizations and detailed proofreading. Moreover some references and how they were introduced in the thesis was adjusted due to updated versions of the references got published since the submission of this thesis. No results or figures have been modified in this published version.

Abstract

Machine Learning (ML) has demonstrated its potential as a powerful technique to improve the performance of an Earth System Model (ESM), yet many challenges remain. ESMs are essential tools to project and understand climate change. While they have been improved over the last decades, longstanding systematic errors compared to observations and uncertainties in climate projections remain. Convective processes are in general unresolved as their typical length scale is smaller than the grid size of ESMs. The effect of such subgrid processes is traditionally taken into account with parameterizations. These parameterizations, such as mass-flux schemes that represent the effect of convective processes on e.g., large-scale dynamic and thermodynamic fields, are often attributed to be the main sources of model errors and translate into uncertainties in climate projections. A way to reduce these limitations of ESMs is to take advantage of novel ML or deep learning (DL) algorithms that learn actively on output from convection permitting high-resolution simulations. These high-resolution simulations can resolve a large fraction of the relevant processes explicitly. The resulting data-driven parameterizations are then coupled with an ESM and replace existing traditional subgrid parameterizations in hybrid (physics + ML) ESMs. This thesis presents novel approaches to transform DL algorithms from data science concepts towards an operational use in ESM simulations. First, a DL algorithm is developed that enables to better understand subgrid convective processes and interactions with the large-scale environment. Second, a novel DL algorithm ensemble approach is developed, that provides an improved representation of convective processes. Third, it is demonstrated that the more realistic uncertainty quantification of the ensembles capturing the chaotic nature of subgrid processes reduces longstanding biases in a hybrid model run of an ESM.

Specifically, a Variational Encoder Decoder (VED) is used to deep-learn and to interpret subgrid convective processes in a superparameterized climate model with an aquaplanet setup. The developed VED combines a realistic reproduction of convective processes and disentangles large-scale drivers of convective processes and convective regimes in its latent space, a lower dimensional manifold between the encoding and decoding part of the DL model. In the latent space the information is compressed into only five latent nodes, which is a substantial dimensionality reduction compared to artificial neural nets developed in previous work. Interpreting the latent space allows a detailed investigation of drivers of convective processes and convective regimes along distinct latent dimensions. The latent space of the VED reveals that the temperature differences between the poles and the tropics in combination with the characteristics of subtropical and subpolar air masses along the mid-latitude storm tracks are key drivers of convective processes. Moreover, the VED separates key characteristics

of shallow convective, cumulus, cirrus-like and deep convection regimes. This demonstrates the improved understanding of convective processes gained with the latent space of a VED.

Apart from the improved understanding of convective processes with interpretable DL algorithms, these algorithms can be combined to form ensembles. I constructed and coupled a number of these DL ensembles to an ESM with realistic topography. This makes it possible to account for uncertainties due to the chaotic nature of subgrid processes. Some of these novel ensembles improve the representation of subgrid processes compared to individual DL models. Moreover, the DL ensembles realistically capture the variability of subgrid processes with their uncertainty quantification. This realistic uncertainty quantification is a crucial step to enhance the trustworthiness of DL models for their application in hybrid ESMs. The two best performing DL ensembles are then coupled to an ESM replacing the prediction of a superparameterization except for cloud water tendencies. The resulting hybrid ESMs with the two DL ensemble parameterizations run stably over more than five months. Compared to a traditional convection scheme the novel DL schemes, despite some induced biases in large-scale fields, improve the reproduction of precipitation extremes and the diurnal cycle of continental precipitation similar to a superparameterization. This demonstrates the applicability and advantages of DL ensemble parameterizations when coupled to an ESM, especially over individual deterministic DL models that fail within the first days.

My thesis thus advances the modelling of convective processes with DL in Earth system sciences via enhanced representation and understanding of convective processes in ESMs. It provides ways to reduce limitations of state-of-the-art ML models and paves a way forward to the operational use of DL and ML in the next generation of ESMs.

Integrated Author's References

Parts of this thesis (text, figures, and tables) are already published in several peer-reviewed or studies that are in review currently. Especially Chapter 3, 4 and 5 are reproducing the methods, results and conclusion sections of my published work with minor modifications. Therefore it has to be noted that these chapters have to be seen as direct citations from my published and work that is in review. In these chapter I will use the pronoun "we" to acknowledge the contributions from my co-authors. More details on the published and work that is currently in review and author contributions can be found in section 1.3 and at the beginning of the main chapters 3, 4 and 5.

Publications as first author and co-author

Behrens, **G.**, Beucler, T., Gentine, P., Iglesias-Suarez, F., Pritchard, M., & Eyring, V. (2022). Non-Linear Dimensionality Reduction With a Variational Encoder Decoder to Understand Convective Processes in Climate Models [e2022MS003130 2022MS003130]. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS003130. https://doi.org/10.1029/2022MS003130

Behrens, **G.**, Beucler, T., Iglesias-Suarez, F., Yu, S., Gentine, P., Pritchard, M., Schwabe, M., & Eyring, V. (2024). Improving Atmospheric Processes in Earth System Models with Deep Learning Ensembles and Stochastic Parameterizations. *In Review to Journal of Advances in Modeling Earth Systems*. https://doi.org/10.48550/arXiv.2402.03079

Yu, S., Hannah, W. M., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., **Behrens**, **G.**, Busecke, J. J. M., Loose, N., Stern, C., Beucler, T., Harrop, B. E., Hilman, B. R., Jenney, A. M., Ferretti, S. L., Liu, N., Anandkumar, A., ... Pritchard, M. S. (2023). ClimSim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators. https://doi.org/10.48550/arXiv.2306.08754

Contents

1. Introduction

1.1. Motivation

In the last decades, our understanding of anthropogenic climate change has broadened due to numerous studies that have been synthesized in the Intergovernmental Panel on Climate Change (IPCC) assessment reports. These assessment reports are based on different phases of the Coupled Model Intercomparison Project (CMIP) (Eyring et al. 2016), which consists of orchestrated Earth System Model (ESM) simulations following a consistent experimental protocol. An overarching aim of these ESMs is to skilfully represent the current and past climate state of the Earth system (Eyring et al. 2016), which enables long-term climate projections throughout the 21^{st} century. The climate itself is defined as the statistics of accumulated weather phenomena in all components of the Earth system over a reference period (30 years based on protocols of the World Meteorological Organisation (WMO), Gettelman and Rood 2016). The purpose of the ESMs is to represent these crucial statistics, internal variability, effects of the anthropogenic forcing and other factors influencing the Earth system via the primitive equations of resolved large-scale fields e.g., temperature or specific humidity in the atmosphere in the past, present and future. However, this task of an ESM involves a crucial scale separation between slower, large-scale processes that are directly simulated with the numerical core of the ESM and faster, small-scale phenomena. Due to the typical coarse horizontal resolution of ESMs e.g., 50 - 100 km, there are important processes for the climate system acting on typical length scales below the ESM horizontal grid size, such as convection i.e., the processes in a moist atmospheric environment that drive cloud formation, impact radiation, the general circulation and many other large-scale processes that are resolved in ESMs. Thus their effects have to be approximated via parameterizations in ESMs. Throughout the thesis, I will refer to those small-scale processes that need to be parameterized as "subgrid processes". These subgrid processes are not explicitly represented in ESMs.

In the last decades, ESMs have substantially improved for example reducing the biases in specific humidity fields compared to observations (Eyring et al. 2021). This progress in Earth system modelling has led to the attribution of the recent climate change with virtual certainty to anthropogenic causes (Eyring et al. 2021). Similarly key effects of climate change like surface air temperatures warming, sea level rise, sea ice reduction in polar latitudes and many other effects got more and more certain (IPCC 2021). Despite this significant progress in Earth system modelling, long-standing systematic errors remain in ESMs compared to observations and result in pronounced uncertainties in their future projections. One example is the "double Inter-Tropical Convergence Zone bias" over the tropical southwestern Pacific

Ocean southeast of Papua New Guinea, where ESMs simulate too much precipitation in comparison to observations (i.e., Bock et al. 2020). Another important example of uncertainties in future projections is the persistent range of Equilibrium Climate Sensitivity (ECS). The ECS is defined as the temperature difference before a doubling of the atmospheric carbon dioxide (CO_2) concentration and after reaching a steady state after the CO_2 doubling (i.e., Bock et al. 2020). In fact, current ESMs participating in the last CMIP phase (CMIP6) yield an increased ECS range compared to previous phases (CMIP3 and CMIP5), with values between 1.8 $\frac{K}{CO_2\ doubling}$ to 5.6 $\frac{K}{CO_2\ doubling}$ (Schlund et al. 2020). In other words, the uncertainty in ECS has grown despite substantial improvements in ESMs, partly associated with the representation of clouds (Bock et al. 2020; Lauer et al. 2023). To a large extent, systematic biases in ESMs and uncertainties in their future projections are associated with deficiencies of traditional convection schemes (i.e., Bock et al. 2020; Bony et al. 2015; Gentine et al. 2021; Rasp et al. 2018), see further details in chapter 2 of the thesis.

Storm Resolving Models (SRMs) with a horizontal resolution below 5 km overcome this "convective deadlock" and the reliance on convection parameterizations to represent deep convection (Gentine et al. 2018; Randall et al. 2003; Randall 2013). Deep convective processes can be explicitly resolved with these high resolutions. Thus SRMs simulate the Inter-tropical Convergence Zone, precipitation extremes or the diurnal cycle of precipitation with clearly improved skill compared to ESMs (Satoh et al. 2019; Stevens et al. 2020). Despite these encouraging results, SRM simulations remain computationally demanding. Therefore state-of-the-art SRM runs are often restricted to periods from seasonal to annual time scales.

Deep Learning, where machine learning, an optimizable algorithm like a neural network, performs a multi-dimensional nonlinear regression task with up to a millions degrees of freedom, has already demonstrated great potential in learning subgrid processes in ESMs (Gentine et al. 2021). Moreover, deep learning makes it possible to translate a large portion of the advantages of SRM simulations into ESM simulations while decreasing the computational costs compared to SRM simulations (Rasp et al. 2018). In such trailblazing "hybrid model experiments", deep neural networks and random forests based parameterizations, coupled with the large-scale dynamics of the host general circulation model, have shown substantial improvements in the performance of the ESM due to a better representation of the influence of subgrid processes (Rasp et al. 2018; Yuval and O'Gorman 2020). Nevertheless, the complexity of deep learning models with up to a million degrees of freedom severely limits their interpretability and the overall understanding of the reproduced convective processes. In this thesis, therefore, I develop novel deep learning network architectures that are by construction highly interpretable. While the application of these interpretable deep learning techniques has been restricted to idealized models and focused on single variables in climate science in the past, I utilize here these techniques in a realistic setting based on multivariate climate model data to broaden our understanding of convective processes and the large-scale environment in which they occur. The application in a multivariate setting that is presented in this thesis is a fundamental step from theoretical experiments towards more routine use of deep learning based postprocessing to help climate scientists to better understand complex simulated

processes in ESMs. Furthermore, initial deep learning hybrid ESMs showed a weaker reproduction skill at pressure levels on which convective processes exhibit a more stochastic and turbulent behaviour (e.g., Rasp et al. 2018), such as in the planetary boundary layer. Thus, it is intuitive to ask whether we may obtain an improved reproduction skill using deep learning ensemble approaches that account for variations in subgrid processes related to stochasticity (Han et al. 2023). Moreover, the interpretability, and as a result the trustworthiness, of single deterministic neural networks is in general hampered by the lack of uncertainty quantification, i.e., placing the prediction within a realistic variance. Likewise the stability of the hybrid models, using the ESM in combination with a novel data-driven parameterization, may well be dependent on the deficiencies of the DL algorithm (Han et al. 2023; Lin et al. 2023; Rasp et al. 2018). It is hypothesized that DL ensembles can be helpful in this context as the ensemble may provide a compensation of individual deficiencies of DL models and boost the stability of the hybrid model (Brenowitz et al. 2020; Han et al. 2023).

1.2. Key Science Questions

Based on the introductory remarks in the previous section and the goal of my thesis to better understand and model convection with Machine Learning, I pose three overarching scientific questions that set the scope of this thesis.

1. Can deep learning enhance the understanding of convection and large-scale drivers of convection?

2. Can stochastic and deterministic ensemble deep learning parameterizations that take into account the stochasticity improve the representation of subgrid convective processes "offline" based on ESM data?

3. Do stochastic and deterministic ensemble parameterizations with calibrated uncertainty quantification of subgrid processes have an effect on the stability and improve the quality of hybrid ESM simulations?

1.3. Structure of the Thesis

This thesis consists of parts that are published or in review to peer-reviewed journals (two lead author papers and one co-author study). A list of my studies (published or in review) that are used in this thesis is displayed on page vi. If parts from these studies are presented in this thesis, "we" is utilized to facilitate readability and clarity by omitting the passive voice. Moreover I use "we" to acknowledge all involved co-authors. However I declare, unless stated otherwise, that all content from these publications (text, figures, and tables) displayed in this thesis originates from me as the author of this thesis. Contributions to these studies are listed at the beginning of the corresponding chapters.

This thesis is structured as follows. Chapter 1 provides an introduction to the topic, the key scientific questions addressed in my thesis, and the overall structure of the thesis. Chapter 2 reflects the scientific background of this thesis. Section 2.1 gives an introduction into clouds, convective processes and the general atmospheric circulation. The subject of section 2.2 is the theory behind convective processes. It is followed by section 2.3, which explains how convective processes are represented in Earth system models with convection schemes, with a superparameterization and Storm Resolving Models. Finally, section 2.4 illustrates how machine learning or deep learning approaches enable the development of novel data-driven parameterizations of convective processes.

Chapters 3 to 5 describe the results of this thesis. Chapter 3 presents the results of the first study of this thesis which has been published by the *Journal of Advances in Modeling Earth Systems* in **Behrens** et al. 2022. This chapter focuses on the first key scientific question and shows how a latent space of a VED in combination with generative modelling can be utilized to broaden our understanding of convective processes and large-scale drivers of convection in a climate model. Chapter 4 is based on my work in **Behrens** et al. 2024. The related paper (**Behrens** et al. 2024) is in review to *Journal of Advances in Modeling Earth Systems*. This chapter targets the second research question and focuses on an evaluation of the general reproduction of convective processes with stochastic and deterministic ensemble parameterizations compared to individual deep learning models based on test data from a realistic global Earth system model simulation. Chapter 5 builds on the previous chapter and forms the second part of my work in **Behrens** et al. 2024. The related paper (**Behrens** et al. 2024) is in review to *Journal of Advances in Modeling Earth Systems*. This chapter covers the third key research question in detail. Herein, I shed light on the quality of the resulting uncertainty estimates of the different stochastic and deterministic ensemble parameterizations. I then evaluate the performance of the most skillful ensemble parameterizations when coupled back into an Earth System Model compared to individual neural networks and existing convection schemes. Chapter 6 summarizes the key findings of this thesis and puts them into the broader context for climate science and their impact for the Earth system modelling community. Appendices A and B present supporting information of **Behrens** et al. 2022 and **Behrens** et al. 2024 together with a glossary for all used abbreviations, figures and tables of this thesis. The last part of the thesis consists of the references used in this thesis and acknowledgements for particular persons and funds that were essential for the progress of this thesis.

2. Scientific Background

In this chapter, the scientific background of convective processes and how these are represented in Earth system models is provided. Section 2.1 focuses on the background of clouds, convective processes and the atmospheric general circulation. The theory of convective processes is discussed in section 2.2. Section 2.3 illustrates how convective processes are represented in Earth system models and sheds light on known deficiencies of these models with respect to convective processes. Section 2.4 highlights the potential of machine learning algorithms to provide an improved representation of convective processes in Earth system models.

2.1. Clouds, Convective Processes and the Atmospheric General Circulation

2.1.1. Clouds: In the retrospective

Clouds are a visual result of convective processes in the troposphere. The research related to clouds and convective processes is one of the oldest meteorological research topics. Dating back to the early 1800s when Luke Howard (Howard 1894, the third volume of his original manuscript from 1803) started to classify clouds in the sky based on their appearance and texture. Such cloud atlases were one of the first ways to understand clouds, convective processes and different cloud types (convection regimes) in the troposphere. As an example, today every weather observation station has to report the cloud type of low, middle and high clouds or the cloud cover in their regular observations to fulfil World Meteorological Organisation (WMO) standards.

From an observational perspective a further major advance in meteorology related to clouds, was the increase of observation stations around the globe during the late 19^{th} and early 20^{th} century. This enabled more robust weather forecasts and a basic understanding of the weather situation and in particular cloud fields associated with extra-tropical cyclones over Europe and North America. Likewise the invention of radiosondes in the 1920s / 1930s and their integration into the weather observation over the next decades was a step forward in better understanding convective processes. In detail radiosondes enabled for the first time the measurements of vertical temperature and humidity profiles throughout the troposphere that are crucial characteristics to understand convective processes. Based on this, estimates about the cloud base or the cloud top height could be drawn only by comparing the ambient temperature and theoretical adiabatic and pseudo-adiabatic lapse rates.

Since the 1970s Earth observing satellites are playing a crucial role in meteorology. With them we could observe important quantities like cloud cover not only based on the weather station network, but get a quasi global map based on geostationary satellites like Meteosat (Holmlund et al. 2021). Furthermore satellites provide measurements of the cloud top heights via the outgoing longwave radiation and retrieved brightness temperatures (Lohmann et al. 2016). Satellite products, e.g., European Space Agency's Climate Change Initiative Cloud (ESACCI-CLOUD), and related reanalysis products like, e.g., European Centre for Medium-Range Weather Forecasts fifth-generation reanalysis (ERA5), give us key information about key properties of clouds, allowing us to obtain the cloud ice water and cloud liquid water path or the precipitation of a cloudy column respective grid cell (e.g., Lauer et al. 2023).

Today cloud observations in combination with gained physical understanding about their driving processes play an essential role for climate science.

In the following subsection I will briefly explain how convective processes and the atmospheric general circulation are connected in the Earth system.

2.1.2. Convective Processes and Atmospheric General Circulation

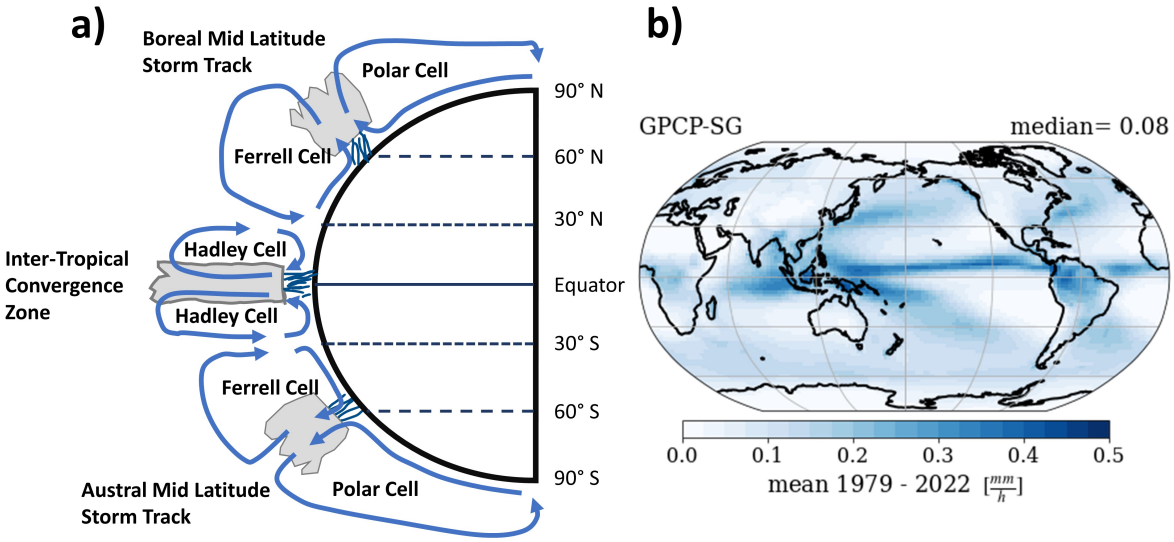

Figure 2.1.: Schematic of the atmospheric general circulation: The left subplot shows a simplified version of the general circulation from the North Pole to the South Pole. The blue arrows indicate the atmospheric general circulation in vertical or meridional direction. The clouds inside the schematic depict the regions with dominant convective processes that play an important role for the atmospheric general circulation. Additionally the left subplot shows the meridional cell structure of the atmospheric general circulation. The right subplot shows the global temporal average precipitation field of Global Precipitation Climatology Project - Satellite-Gauge (GPCP-SG) for the period 1979 to 2022 as a proxy for the spatial structure of convective processes. The spatial median of the temporal average precipitation field of GPCP-SG is shown top right above the right subplot.

The advances in observing convective processes with satellites and associated reanalysis products over the last decades enabled us to explore how convective processes are connected with the atmospheric general circulation (Bony et al. 2015). Figure 2.1a shows a simplified version of the observed atmospheric general circulation on a meridional-vertical plane from the North to the South Pole. Figure 2.1b displays the global map of annual mean precipitation averaged over the period from 1979 to 2022 based on the monthly data set of the satellite-gauged product Global Precipitation Climatology Project - Satellite-Gauge (GPCP-SG)(see Huffman et al. 2023 for its latest version). The observed atmospheric general circulation consists of three cells per hemisphere in meridional direction. The Coriolis force plays a crucial role in shaping the general circulation of the Earth's atmosphere and the formation of circulation cells, such as the Hadley, Ferrel, and Polar cells. The Coriolis force influences the large-scale wind patterns, including the formation of the trade winds, westerlies, and polar easterlies, by deflecting atmospheric flows towards the right of the flow direction on the Northern hemisphere and towards the left of the flow direction on the Southern hemisphere. In the following I will explain the connection of these cells with precipitation patterns seen in Figure 2.1b. For a detailed description of the distinct cells I point the reader to Liljequist and Cehak 2013 and to Bony et al. 2015.

In Figure 2.1b the pattern of the Inter-Tropical Convergence Zone (ITCZ), a zone of increased precipitation in the tropics near the Equator, is visible over the equatorial Pacific Ocean, the equatorial Atlantic Ocean and the eastern equatorial Indian Ocean with elevated precipitation. In these regions deep convection is predominating with cloud tops near the tropopause (~ 16 to 20 km). This deep convection and the resulting strong diabatic heating due to the very moist and warm cloud air in the ITCZ drives the ascending branch of the Hadley Cell on both hemispheres (Figure 2.1a).

In contrast to the elevated precipitation in the ITCZ region near the Equator, we see in general no or negligible precipitation in the subtropics around 30° N or S. An explanation for this is the general subsidence of air masses in the subtropics, that forms the descending branch of the Hadley and Walker cells (Figure 2.1a). Due to that subsidence the air masses are generally dry in the subtropics related to a pronounced adiabatic heating. The subsidence and the low humidity of air masses typically forms an inversion layer above the planetary boundary layer. This inversion limits the vertical extent and the strength of convective processes. Therefore no or only shallow convection, that is limited to the planetary boundary layer and the lower free troposphere, is predominating in the subtropics and over the upwelling regions in the tropical eastern Pacific and Atlantic Ocean. Thus, barely precipitation is formed in these regions.

A second pattern with elevated precipitation is present along the mid latitude storm tracks between 45° and 60 ° N or S (Figure 2.1b), that forms the ascending branch of the Ferrell and Polar Cell (Figure 2.1a). Low pressure systems associated with the meandering jet stream characterize the mid latitude storm tracks on both hemisphere. Often weaker precipitation occurs near the warm front of the extratropical cyclones that is dominated by stratiform clouds. Deep convection characterizes the cold front of extra-tropical cyclones and forms a second major source of diabatic heating for the atmospheric general circulation.

Over the Arctic Ocean and Antarctica the precipitation rates are similarly negligible like in the subtropics (Figure 2.1b). High latitudes are also characterized by subsidence of air masses in general, which forms the descending branch of the Polar Cell (Figure 2.1a). Moreover the temperatures on polar latitudes are on average well below 273 K due to the high surface albedos, low average solar insolation and pronounced radiative cooling of the atmosphere, which results in extremely low specific humidities. These two factors result in typically only weak convective processes that we see in polar latitudes.

These results indicate how deep convection along the ITCZ and the mid latitude storm tracks drives the atmospheric general circulation. In contrast, subsidence related to the general circulation limits the strength of convective processes in the subtropics and only little or shallow convection is predominating on these latitudes. This demonstrates the necessity to investigate convective processes together with the large-scale environment in which they are forming to better understand the complex interaction between them (Bony et al. 2015). I will investigate this interplay in detail in chapter 3 with novel interpretable DL methods in a climate model.

The next section explains the theory of convective processes and how we can obtain some properties of clouds based on observations.

2.2. Theory of Convective Processes

2.2.1. The vertical extent of clouds and convective processes in the troposphere

Satellite observations and radiosonde measurements provide concepts to help to better understand convective processes in the atmosphere, especially in the context of investigating the vertical extent of convective processes (e.g., estimate the cloud top height).

One way to investigate the vertical extent of convective processes is to use the measured temperature profiles of a radiosonde in combination with adiabatic and pseudo-adiabatic lapse rates. They can be plotted in a "tephigram" that allows to estimate the cloud base, cloud top and the vertical extent of the acting convective processes. Figure 2.2 illustrate a tephigram. The y-axis of the tephigram is given by pressure coordinates **P**. As the isobars depend also on variations in temperature and specific humidity they are slightly curved and not straight lines. The x-axis is defined by the isolines of saturation specific humidities q_s, which are not orthogonal to the y-axis in this case.

In Figure 2.2 the measured vertical temperature profile T_{env} is denoted by the solid black line. One approach to determine the cloud base is to assume that the lifted air parcel has a prescribed higher temperature at the surface than the environment. This enables the uplift of the air parcel due to its positive buoyancy. From the surface the air parcel follows a dry adiabatic lapse rate along an isoline of potential temperature. In the following the theoretical uplift profile is denoted as T_{lapse}. The first intersection between T_{env} and T_{lapse} defines the Lifting Condensation Level (LCL). The LCL defines the level where condensation starts due to lifting. The respective level when the condensation starts (cloud base) is known as Convective

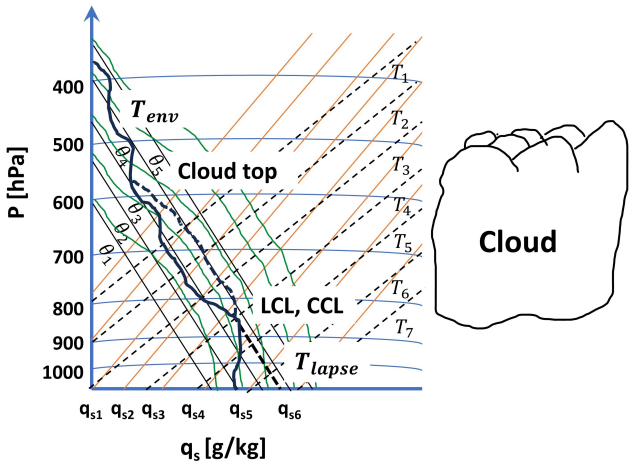

Figure 2.2.: Schematic of a tephigram: The tephigram shows the vertical profile of the temperature of environmental air (T_{env}, bold black line) and the theoretical uplift curve (T_{lapse}, dotted bold black line). The schematic further illustrates the Lifting Condensation Level (LCL), Convective Condensation Level (CCL) and the approximate cloud top height. In the background the solid black lines indicate isolines of constant potential temperature θ (which are identical to the lines based on the dry adiabatic lapse rate Γ_{dry}) and the dashed black lines are distinct isotherms T. The green curves indicate theoretical pseudo-adiabatic uplift curves. Additionally the isolines of saturation specific humidity (q_s, orange lines) and isobars (blue curves) are shown.

Condensation Level (CCL) and agrees for cumulus convection well with LCL, as it is shown in Figure 2.2 (Liljequist and Cehak 2013; Lohmann et al. 2016). From the CCL the air parcel that is saturated with respect to liquid water follows a pseudo-adiabatic lapse rate. Due to the release of latent heat of condensation Convective Available Potential Energy (CAPE) is built up. CAPE can be explained as an estimate of the added kinetic energy for the uplift of a parcel due to the release of latent heat (Lohmann et al. 2016). CAPE is produced until T_{lapse} intersects with T_{env} for a second time. This second intersection approximates the cloud top height in theory, because the air parcel is not anymore buoyant with respect to its surroundings. However the cloud top height in reality is usually situated on higher levels as it is indicated in Figure 2.2, when all the CAPE is transformed into Convective Inhibition (CIN) (Lohmann et al. 2016). CIN is an estimate of the mechanical work that is applied on the air parcel that is needed for the uplift or to sustain the uplift, respectively (Lohmann et al. 2016). So the vertical integral over CIN and CAPE is zero for the entire column (Lohmann et al. 2016).

Apart from radiosonde soundings, the cloud top height can be also estimated with passive satellite measurements of outgoing longwave radiation. The cloud top heights can be inferred with brightness temperatures retrieved from the measured outgoing longwave radiation under the assumptions that clouds are almost perfect black bodies for these wavelengths. Another and more accurate way to determine the cloud top height is active remote sensing (Hagihara et al. 2014). Active remote sensing helped to understand convective processes on finer resolutions than with passive sensors (Hagihara et al. 2014). Examples for active remote sensing are the cloud profiling radar of CloudSat or the cloud aerosol lidar Cloud-Aerosol Lidar and Infrared

Pathfinder Satellite Observation (CALIPSO). The cloud aerosol lidar also allows to retrieve the cloud base height and the vertical extent of convective processes (Hagihara et al. 2014).

The combination of passive and active satellite observations over the last decades and meteorological measurements over the last centuries shaped and widened our understanding of convective processes in the troposphere. Furthermore these active observational products increased their quality dramatically in comparison to passive satellite observations, which allowed the scientific community to investigate longstanding questions of the connection between convective processes and the atmospheric general circulation. The next subsection briefly illustrates the theoretical background of the occurrence of convective processes in the atmosphere, which will be essential to understand and simulate convective processes in ESMs.

2.2.2. Theoretical background of convective processes and condensation

Convective processes play a crucial role for a moist air parcel to reach saturation with respect to liquid water or ice water. Saturation means that a phase transition of water vapour via condensation or deposition takes place based on principle thermodynamic laws in moist air. The critical saturation water vapour pressure $e_{s,w}(T)$ for condensation in thermodynamic equilibrium for a given temperature T (equation 2.2) can be approximated using the Clausius-Clapeyron relationship (equation 2.1, Lohmann et al. 2016)

$$\frac{de_{s,w}}{dT} = \frac{s_v - s_w}{\alpha_v - \alpha_w} = \frac{L_v}{T \times (\alpha_v - \alpha_w)} \simeq \frac{L_v}{T \times \alpha_v} = \frac{L_v \times e_{s,w}}{R_v \times T^2} \tag{2.1}$$

$$\rightarrow e_{s,w}(T)|_{L_v=const} = e_{s,w}(T_{trip}) \times \exp\left(\frac{L_v}{R_v}(\frac{1}{T_{trip}} - \frac{1}{T})\right) \tag{2.2}$$

In equation 2.1 $s_{v/w}$ are the specific entropies of water vapour and liquid water, respectively, $a_{v/w}$ are the specific volumes of water vapour and liquid water and L_v is the latent heat release due to condensation. The specific volume of water vapour is magnitudes larger than the respective specific volume of liquid water. This leads to the approximated form of the Clausius-Clapeyron relationship. This allows the computation of the saturation vapour pressure of liquid water $e_{s,w}(T)|_{L_v=const}$ as a function of temperature of the moist air parcel and assuming a constant latent heat release. R_v is the specific gas constant of water vapour. $e_{s,w}(T_{trip}) = 611.2\ hPa$ and $T_{trip} = 273.15\ K$ are the reference saturation water vapour pressure and temperature at the triple point.

Condensation of pure water vapour occurs only in a thermodynamically supersaturated environment as a result of the strong surface tension forces σ_{surf} (Lohmann et al. 2016). This supersaturation that is necessary to onset a phase transition with homogeneous nucleation can be explained with the curvature effect (Lohmann et al. 2016). This means that the saturation vapour pressure on the curved surface of an initial cloud droplet $e_{s,w}(r)$ of radius r is exponentially higher than over a plane water surface $e_{s,w}(\infty)$ (equation 2.3). Moreover the needed supersaturation decreases with the radius of the initial cloud droplet, which means

that condensation of pure water vapour occurs only in extreme supersaturated conditions (Lohmann et al. 2016).

$$e_{s,w}(r) = e_{s,w}(\infty) \times \exp\left(\frac{2 \times \sigma_{surf}}{r \times R_v \times \rho_w \times T}\right) \tag{2.3}$$

Aerosols in the atmosphere act as cloud condensation nuclei and reduce the necessary saturation water vapour pressure for condensation to values of about the theoretical one given by the Clausius-Clapeyron relationship. The so-called "solute effect" further reduces the supersaturation required for the condensation if the cloud condensation nuclei contains water soluble components. Heterogeneous nucleation acts as a predominant source of cloud droplets in the atmosphere. Common cloud condensation nuclei over the ocean are sea salt and mineral dust, sulfate particles or soot over the continents.

In the following, I will explain processes in the atmosphere that can lead to condensation. The uplift of air parcels and the related adiabatic cooling is a way to reach the LCL or CCL, as I already showed in the tephigram (Figure 2.2). This convective uplift due to positive buoyancy is often accompanied by fine-scale turbulence. This fine-scale turbulence i.e., by entraining dry ambient air masses into moist air, influences when the moist air is saturated with respect to water. These two processes often result in cumuliform convective regimes (Lohmann et al. 2016). Apart from convective processes, the horizontal advection of specific humidity, temperature or moist static energy, may lead to conditions that enable condensation. It is hypothesized that advection and related convergence may play a role for convective self aggregation (Wing et al. 2018). Moreover radiative cooling and the related temperature decrease is another way that can lead to condensation in near-surface layers resulting in the formation of fog or shallow convection especially during night time (Lohmann et al. 2016). Radiative cooling results often in stratiform clouds (Lohmann et al. 2016). Likewise surface diabatic fluxes especially over the ocean play a role for shallow convection and the formation of clouds near the surface within the planetary boundary layer (Wing et al. 2018). Apart from these thermodynamic factors, changes in aerosols and thus the cloud condensation nuclei can influence the formation of clouds. In particular anthropogenic sources of cloud condensation nuclei play a role to form "artificial" clouds as can be seen for e.g., ship tracks or contrails. Moreover natural sources of aerosols may influence the formation of clouds i.e., mineral dust that is lifted up from the deserts into the troposphere (Lohmann et al. 2016). The processes discussed above are only a fraction of all processes that influence convection. Drivers of convective processes are often superposed on each other which limits the general understanding what drives the respective processes. I will demonstrate a novel deep learning technique in this thesis in chapter 3, that provides a new data-driven point of view of the interplay between convective processes and the large-scale thermodynamic state related to the atmospheric circulation.

Convective processes and cloud formation occur predominantly on spatial scales that are far smaller than the horizontal grid size of an Earth system model. Therefore their effect on

the Earth system together with other subgrid factors have to be parameterized in Earth system model, while some of them could be directly resolved with "storm resolving models" with a km-scale horizontal resolution. The next section explains how convective processes are treated in Earth system models, explicitly simulated with storm resolving models and approximated with a superparameterization.

2.3. Convective Processes in Earth System Models and Storm Resolving Models

This section covers the topic how convective processes are represented in Earth system models and storm resolving models. The first part of this section highlights the way and the caveats how subgrid convection schemes represent convective processes in Earth system models with multiple examples of convection schemes. The second part will explain the concept of the superparameterizataion and its benefits compared to traditional convection schemes that I will use for the investigation in this thesis. The third and last part of this section shows solutions to overcome known limitations of traditional convection schemes with storm resolving models and novel numerical techniques based on atmospheric simulations with increased resolution.

Beforehand I will briefly explain ESMs and a few limitations with respect to convective processes. An ESM is a key tool to investigate the historical evolution and future changes of the Earth system (Gettelman and Rood 2016). ESMs represent typically a large portion of the climate-relevant processes in the atmosphere, ocean, land or cryosphere of the Earth system from a bio-geo-chemical stand point. This may include the chemical cycles of carbon and nitrogen and to some extent interactive chemistry in all components of the model (Gettelman and Rood 2016). The setup of an ESM or hereafter also simply climate model depends on the particular research questions. This means that not all components of the Earth system have to be necessarily included and instead prescribed boundary conditions can be used to account for the excluded component. Also computationally heavy calculations like interactive chemistry, the simulation of closed bio-geo-chemical cycles or coupling between components can be adjusted to the needs of the particular research (Gettelman and Rood 2016). Large-scale atmospheric dynamics are solved in ESMs via the primitive equations of state based on general fluid dynamics (Gettelman and Rood 2016). Non-resolved processes however need to be parameterized. This includes for example convection.

ESMs taking part in the model intercomparison projects CMIP3 to CMIP6 show persistent biases in convection related fields when compared to observations (Bock et al. 2020). One example for such long-standing biases is the "double ITCZ bias" with many ESMs overestimating the mean precipitation over the tropical southwestern Pacific Ocean in a region south of the Equator from Papua New Guinea towards the Date Line compared to observations (Bock et al. 2020; Lauer et al. 2023). Another example is the underestimation of shallow convection over the upwelling regions in the tropical and subtropical Atlantic, Indian and Pacific Ocean

(Bock et al. 2020). The reduced cloud cover related to too weak shallow convection e.g., to the west of the Peruvian coast results in a large underestimation of shortwave cloud radiative effect (swcre) of more than 30 $\frac{W}{m^2}$ (Bock et al. 2020; Lauer et al. 2023).

These two biases are often attributed to the uncertainties arising from the use of conventional convection parameterizations (i.e., **Behrens** et al. 2022; Bock et al. 2020; Bony et al. 2015; Gentine et al. 2021) The ESMs participating in CMIP6 utilize a "small zoo" of in complexity varying subgrid parameterizations for processes like cloud microphysics (essential for the simulation of realistic precipitation rates), cumulus convection itself (representing the uplift, entrainment, detrainment of air masses due to clouds), cloud cover (key factor for a realistic radiation budget at the top of atmosphere), radiation (e.g., to simulate the effect of cloud water and ice on shortwave and longwave radiation) and also small-scale turbulence, gravity waves and boundary layer physics (which affect the simulated evaporation and other boundary conditions for convective processes). All these factors have an influence on the representation of convective processes in an ESMs and are potential sources for the large regional biases with respect to observations of various convection related variables (Lauer et al. 2023). Furthermore this variety of different parameterizations hampers our overall understanding of convective processes in ESMs, because it is challenging to distinguish between a e.g., realistic precipitation pattern or an artifact from a subgrid parameterization or even the interplay of various of them and the dynamical core of the ESM.

All these different parameterizations consist of one or multiple equations that calculate an estimate for a given variable of interest (e.g., cloud cover, radiative or mass fluxes, ...) based on selected large-scale state variables. Often these equations contain tuning parameters that are not constrained by physics and that allow to adjust the simulated processes towards an improved agreement with observations. Additional uncertainty exists inside the parameterization related to processes like i.e., fine-scale turbulence that are of stochastic nature. This stochasticity cannot be entirely captured with an deterministic parameterization in an ESM.

The following part of the thesis describes selected subgrid convection schemes and shows why it is challenging to adjust tuning parameters in an existing scheme.

2.3.1. Subgrid Convection Schemes in Earth System Models

The general aim of subgrid convection schemes is to mimic the average effects of convective processes that are far smaller than the typical horizontal length scale of the climate model's grid cells. In the following, three examples of subgrid convection schemes of ESMs are given that are useful for the general understanding of this thesis. The presented schemes depend on the same theoretical background with respect to thermodynamics. In detail, the Zhang-McFarlane scheme, that is used in this thesis, can be seen as a simplification of the Arakawa-Schubert scheme. Moreover all presented convective schemes show general concepts that are applied in the superparameterization (see section 2.3.2), which is the main benchmark scheme for the developed deep learning algorithms in this thesis.

Arakawa-Schubert Scheme

One way to parameterize the effect of convective processes in a climate model is via mass-flux schemes. These mass-fluxes schemes can be used to update the large-scale thermodynamic and dynamic variables in the primitive equations, that are also called general equations, of the numerical core of the ESM. Convection mass-flux schemes typically distinguish between four crucial processes. Additionally these schemes distinguish between a cloud free area fraction $\mathbf{A}_{no\ cloud}$ and a cloudy area fraction $\mathbf{A}_{cloud}$ within a grid cell (Arakawa and Schubert 1974). The first process is the entrainment of ambient air from the cloud-free area into the cloudy area. This process is especially important at the lateral boundary of the cloudy area and at the cloud base (Arakawa and Schubert 1974). Entrainment **E** into the cloudy area from a mass-flux point of view is a convergence of mass at a given level. This convergence causes an upward mass-flux $\mathbf{M}_{cloud}$ as a result of the mass conservation described by the continuity equation and the buoyancy inside the cloudy area, which is the second key process. In the cloudy area the cloud air rises up to a certain level, where all buoyancy of the cloud air with respect to its environment reaches zero. This is similar to the production and dissipation of CAPE as shown in the tephigram in Figure 2.2. At this level the cloud air is detrained into the cloud free area. Detrainment **D** is the third crucial process in a mass-flux convection parameterization. Because of the conservation of mass, the cloud free area is characterized by a net subsidence, which is the fourth key process in a convection scheme. The total net mass flux of the combined cloud and cloud free areas can be computed as the product of the air density ρ and the average vertical velocity of the entire grid cell $\overline{\omega}$. Using the net mass flux term and $\mathbf{M}_{cloud}$ yields equation 2.4 for the downward mass flux in the cloud free area $\mathbf{M}_{no\ cloud}$.

$$\mathrm{M}_{no\ cloud} = \rho \times \overline{\omega} - \mathrm{M}_{cloud} \tag{2.4}$$

The downward mass flux in the cloud free area closes the equations in this mass flux scheme.

The Arakawa-Schubert scheme for one grid column is illustrated in Figure 2.3. One modification in this scheme is that instead of one single confined cloud area, we have an ensemble of clouds varying in time and space (Arakawa and Schubert 1974). The individual clouds of the ensemble can have different cloud bases and top heights. This means that the total entrainment **E** and detrainment **D** of the cloud ensemble depends on the detrainment $\mathbf{D_i}$ and entrainment terms $\mathbf{E_i}$ of individual members at various vertical levels of the climate model (Figure 2.3). The same applies for the total mass flux in the cloudy area $\mathbf{M}_{cloud}$ and the total cloudy area $\mathbf{A}_{cloud}$.

For coupling the Arakawa-Schubert scheme to the dynamical core of the climate model the total mass flux of the cloudy area $\mathbf{M}_{cloud}$, the total detrainment **D** and the mixing ratio of cloud liquid water on the level of quasi-neutral buoyancy of the individual detraining cloud members is needed (Arakawa and Schubert 1974). Crucial parameters of the Arakawa-Schubert scheme are the treatment of the mass fluxes from sub-cloud layers and especially the planetary

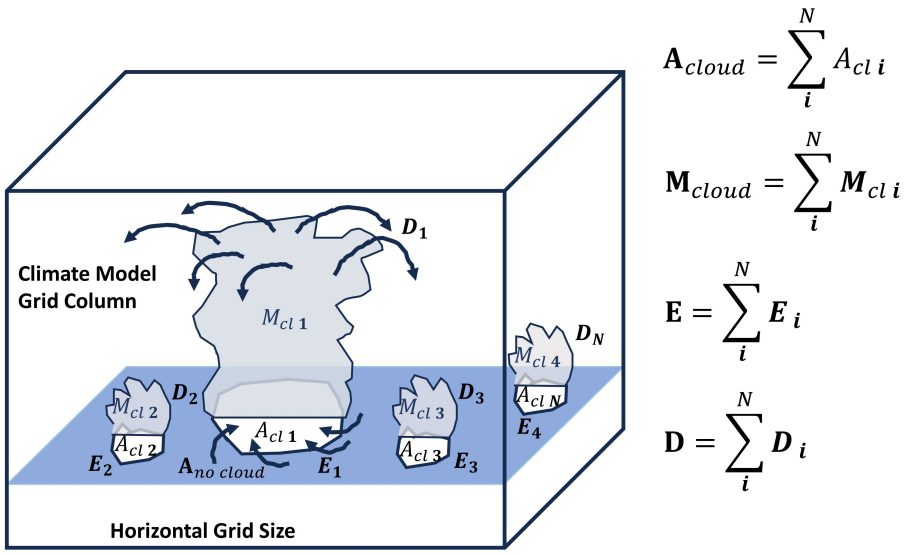

Figure 2.3.: Schematic of the Arakawa-Schubert Scheme (Arakawa and Schubert 1974): The schematic shows an ensemble of **N** clouds inside a grid column of a climate model. The cloudy area A_{cloud} can be computed by summing over all N partial area fractions of the individual clouds $A_{cl\,i}$ at each vertical level. The same applies for the total mass flux of the cloud area M_{cloud}, the total entrainment **E** and the detrainment **D** based on the contributions from individual clouds.

boundary layer (Arakawa and Schubert 1974). Also, initial values of entrainment and detrainment have to be prescribed for the cloud area of the individual clouds Arakawa and Schubert 1974. Moreover the Arakawa-Schubert scheme depends on thermodynamic assumptions, i.e., a necessary instability at a the respective level of cloud bases. (Arakawa and Schubert 1974). The Arakawa-Schubert scheme has certain advantages as it allows to approximate mass fluxes due an ensemble of clouds with varying entrainment and detrainment levels. However, known limitations include the treatment of the subcloud layer and of the cloud microphysics and the need for a critical mixing ratio as a tuning parameter to distinguish between cloud free from cloud air.

Yanai Scheme

Another way to represent the thermodynamic effect of clouds in a climate model is to calculate the heating and moistening due to clouds directly and use these terms in the general equations for temperature and humidity in the climate model. Such an approach is based on the convection scheme of Yanai et al. 1973. First they defined a combined term for the effect of subgrid convection, turbulence and radiation on the general equation for dry static energy $s = c_p \times T + g \times z$, called Q1 (equation 2.5).

$$Q1 = L_v \times (c - e) - \frac{d}{dp}\overline{\omega' s'} + Q_{rad} \tag{2.5}$$

The first part of the sum in Q1 gives the latent heating of the ambient air with respect to the difference in the condensation rate **c** (increase in the mass mixing ratio of cloud water per

time step, Yanai et al. 1973) and the evaporation rate **e** (reduction of the mass mixing ratio of cloud water and time step, Yanai et al. 1973). The second term in equation 2.5 represents the vertical eddy static energy flux due to turbulence that may be influenced by convective processes (Yanai et al. 1973). The third term $\mathbf{Q}_{rad}$ denotes the influence of shortwave and longwave radiative fluxes on the change in dry static energy or heating (Yanai et al. 1973). The general equation of dry static energy can be rewritten so that Q1 can be used as a source term to update the large-scale temperature T. In the superparameterization that I use in this thesis, usually dT/dt with respect to subgrid convection and other processes is computed instead of Q1, but both terms are closely related to each other.

The second term Q2 (Yanai et al. 1973) parameterizes the combined effects of convection and turbulence on the general equation of the specific humidity.

$$Q2 = L_v \times (c - e) + L_v \times \frac{d}{dp}\overline{\omega' q'} \tag{2.6}$$

In Q2 (equation 2.6) the first term represents again the latent heating due to condensation and evaporation, while the second term is the vertical eddy specific humidity flux due to turbulence (Yanai et al. 1973). In the superparameterization Q2 divided by L_v is identical to the change in specific humidity dq/dt to update the large-scale state of q.

Bulk parameterizations based on the general concepts of Yanai et al. 1973 have the advantage that they include the combined effect of subgrid effects of convective processes related turbulence and radiation on the large-scale thermodynamic state variables in one scheme. One clear disadvantage is that Q1 and Q2 reflect a combined effect, which requires expert knowledge to understand whether a heating or moistening is associated with subgrid convective processes or related processes. Similar to the Arakawa-Schubert scheme, the Yanai scheme relies on microphysical assumptions such as when condensation or evaporation starts.

Zhang-McFarlane Scheme

Due to the complexity of the Arakawa Schubert scheme and computational limitations in the last two decades of the 20^{th} century, a class of simplified Arakawa-Schubert schemes were developed for an application in climate models. This class includes the Zhang-McFarlane scheme (Zhang and McFarlane 1995), that is the standard cumulus convection scheme in the Community Atmosphere Model (CAM) (Collins et al. 2006) and the Community Earth System Model (CESM) (Danabasoglu et al. 2020) with a few small adjustments. The analysis in this thesis is based on CAM version 3 (Collins et al. 2006) and CESM version 2 (Danabasoglu et al. 2020). Thus it is intuitive to focus on the Zhang-McFarlane scheme in detail.

One modification of the Zhang-McFarlane scheme with respect to the Arakawa-Schubert scheme is penetrating cumulus convection. Therefore the updraft ensemble (mass fluxes of cloud air) consists of updrafts that could penetrate a conditionally unstable layer in the lower troposphere (Zhang and McFarlane 1995). This prevents some of the "shallow convection" that is simulated with the Arakawa-Schubert scheme. The upward mass flux at the base of all

updraft members is set to a constant value (Zhang and McFarlane 1995). Moreover, the CAPE related to the updrafts decays exponentially with time based on the adjustment time scale. The adjustment time scale is a typical non-physical parameter ("tuning parameter") that has to be defined for subgrid convective schemes. These modifications simplify the computation of the approximated subgrid convective processes.

With these assumptions the heating due to convective fluxes in the convective area can be computed as:

$$c_p \times \left(\frac{dT}{dt}\right)_{conv} = -\frac{1}{\rho} \times \frac{d}{dz}\left(M_u \times s_u + M_d \times s_d - M_{conv} \times s\right) + L_v \times (c - e) \tag{2.7}$$

, where M_d, M_u, M_{conv} are the downward, upward and averaged mass fluxes in the convective area. s, s_u, s_d are the respective dry static energies (Zhang and McFarlane 1995). The second term in equation 2.7 includes the adjustment due to latent heating or cooling from condensation and evaporation, respectively, similar to the Yanai scheme.

The corresponding change in specific humidity $\frac{dq}{dt}$ in the convective area can be written as:

$$\left(\frac{dq}{dt}\right)_{conv} = -\frac{1}{\rho} \times \frac{d}{dz}\left(M_u \times q_u + M_d \times q_d - M_{conv} \times q\right) + c - e \tag{2.8}$$

based on the respective components in the upward, downward branch and averaged over the convective area.

I will use the Zhang-McFarlane scheme in this thesis to benchmark my data-driven parameterization in chapter 5.

Similar to the Arakawa-Schubert scheme, the Zhang-McFarlane scheme uses a critical threshold when convection starts in a conditionally unstable layer and enough CAPE to penetrate this layer. The prescribed decay of CAPE and the treatment of the sub-cloud layer are limitations of the Zhang-McFarlane scheme.

The following part of the thesis will describe the superparameterization that I use to optimize my novel data-driven parameterizations.

2.3.2. Superparameterization

The Superparameterization (SP) in comparison to the conventional convection schemes discussed above is similar to a set of nested high-resolution atmospheric circulation models that each consist of only one vertical grid column. Such an approach has the advantage that these nested grid columns permit to resolve convective processes and their effects directly without an additional convection parameterization. The type of SP that is used in this thesis was first developed and described in Grabowski 2001 and Khairoutdinov and Randall 2001. The ESM and the atmosphere model of the SP simulations of this thesis are CESM version 2 (Danabasoglu et al. 2020) and its atmospheric component CAM in different versions. In chapter 3 CAM version 3 (Collins et al. 2006) is used in an aquaplanet setup (a simulation where the globe is

covered with an ocean and topography is omitted), whereas in chapter 4 and 5 CAM version 6 is utilized as atmospheric component of CESM version 2 (Danabasoglu et al. 2020). For brevity the superparemeterized CAM is called Super Parameterized Community Atmosphere Model (SPCAM) henceforth and the superparameterized CESM version 2 is denoted Super Parameterized Earth System Model (SPCESM).

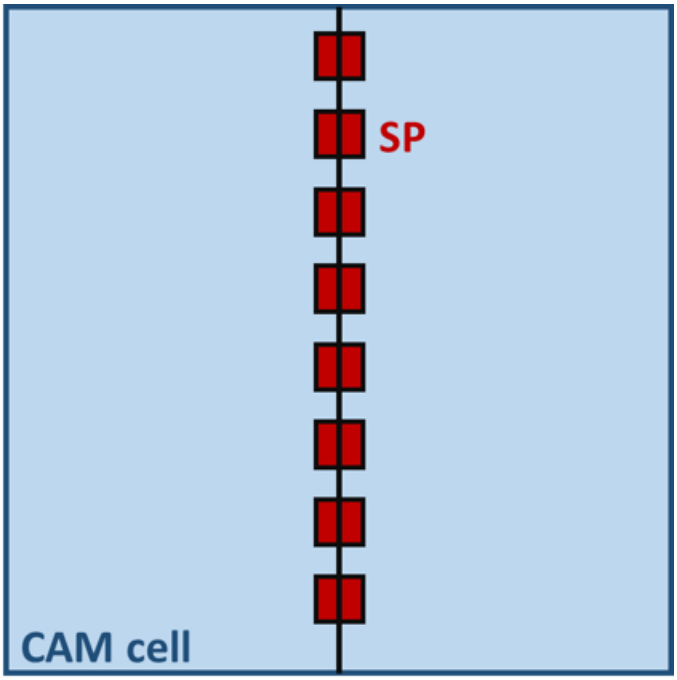

Figure 2.4.: Schematic of the Super Parameterized Community Atmosphere Model (SPCAM) configuration based on Pritchard and Bretherton 2014. Within each coarse Community Atmosphere Model (CAM) cell (blue frame) are in this example 8 nested high resolution grid Superparameterization (SP) (red small squares) cells situated. SP directly simulates subgrid effects of radiation, turbulence and convective processes (Gentine et al. 2018), which is then given back to the numerical core of CAM.

Figure 2.4 is a schematic of the SP setup that was used in SPCAM or SPCESM. The large blue cell represents a grid cell of CAM or CESM 2. In this setup, CAM and CESM 2 have a horizontal grid size in the order of 160 km at the Equator. The set of red cells in Figure 2.4 represents the SP. The number of nested grid columns is a critical tuning parameter of the SP (Pritchard et al. 2014) and strongly affects the necessary computational power. A larger number of nested high-resolution grid columns in combination with the necessary subcyling of the numerical core of each grid column requires considerably larger computational resources than, for example, traditional convection parameterizations. Decreasing the number of nested SP columns from 32 grid columns to 8 grid columns, Pritchard et al. 2014 showed that the CPU time scales by the same factor of 4. In addition they found that the reduction in high-resolution grid columns has no obvious impact on the simulated Madden Julian Oscillation (MJO), an important eastward propagating oscillation of a deep convective system over the tropical Indo-Pacific (Zhang 2005). However, the choice of the number of nested high-resolution grid columns has an impact on cloud liquid water and precipitation due to changes in the vertical mixing (Pritchard et al. 2014).

In detail, a smaller number of nested high-resolution grid columns increases the strength of convective processes in regions that are predominated by deep convection. In section 3 a SP with 8 nested high-resolution grid columns is used in an aquaplanet setup of CAM version 3. In sections 4 and 5 a SP setup with 32 nested high-resolution grid columns is utilized in CESM 2 with a realistic topography and coupled to ocean and land surface model. In both cases the high-resolution grid columns are oriented in meridional direction with an equal spacing between the columns (Figure 2.4). Each of these high resolution grid columns has a horizontal extent of 4 km, which enables the explicit simulation of a large fraction of cumulus convection. However, the SP has its own parameterization for subgrid turbulence and cloud microphysics as these processes cannot be directly resolved in the high-resolution grid cell (Rasp et al. 2018). The nested grid columns have their own numerical core and thermodynamic general equations based on the numerical core of System for Atmospheric Modeling (SAM) (Khairoutdinov and Randall 2003). The vertical axis of SP is coarse with only 30 or 26 levels on a hybrid-sigma grid as the levels are identical to the vertical levels of the host climate model. Such a small number of vertical levels may influence the realism of the representation of subgrid turbulence in the planetary boundary layer or the vertical mixing related to convective processes compared to a storm resolving model with a finer vertical resolution. One advantage of using the same vertical axis in SP and the host climate model is that the information from the large-scale climate model to the SP, and vice versa, can be directly exchanged and used as source terms in the respective general equations without further interpolation. Here, the input data from the climate model to the SP consists of thermodynamic fields only, but also large-scale horizontal velocity fields could be added (Pritchard and Bretherton 2014; Rasp et al. 2018).

In the SP setup that I am using in thesis the vertical profiles of specific humidity **q(p)**, temperature **T(p)** and the meridional wind component **v(p)** are used as input for the SP. Moreover the SP uses the surface pressure $\mathbf{P}_{surf}$, solar insolation Q_{sol}, the surface latent heat flux Q_{lat} and the surface sensible heat flux Q_{sens} as scalar input variables. Additionally, in the chapter 4 and 5 the input variable list of SP includes the vertical profiles of cloud liquid $\mathbf{q}_{cl}$ and cloud ice water $\mathbf{q}_{ci}$. Based on these inputs of SP, the numerical core of each nested high-resolution grid column computes the response of its vertical profile of temperature $\mathbf{T(p)}_i$ and specific humidity $\mathbf{q(p)}_i$, where i symbolizes the number of the respective high-resolution grid column. In chapter 4 and 5 the vertical profiles of cloud liquid water $\mathbf{q(p)}_{cl,i}$ and cloud ice water $\mathbf{q(p)}_{ci,i}$ are added to the output variable list of SP.

The nested high-resolution grid columns have a finer time stepping than the coupled climate model (Khairoutdinov and Randall 2001). This "subcycling" enables the evolution of the resolved processes in the SP columns. At the coarser time step of the climate model an average over all nested high resolution grid columns of the simulated variables is computed (Grabowski 2001; Khairoutdinov and Randall 2001). During the subcycling of SP a radiative transfer model and a microphysics parameterization is called to compute the radiative and surface fluxes (e.g., the precipitation rates) conditioned on the resolved subgrid processes in SP (Khairoutdinov and Randall 2001; Rasp et al. 2018). The precipitation rates and distinct radiative fluxes are also added to the output variables list of averages. These averages are

used to calculate the respective subgrid tendency terms of the thermodynamic state variables in the host climate model and can be used as source terms representing the subgrid processes instead of e.g., a conventional convection parameterization (Khairoutdinov and Randall 2001; Rasp et al. 2018). The same applies for the additionally computed surface and radiation terms that can be used to update the large-scale state in each grid cell of the climate model (Rasp et al. 2018).

One clear advantage of SP compared to conventional convection schemes is that a large fraction of the effects of subgrid convective processes can be explicitly simulated with the nested high-resolution grid columns. Also, the nested high-resolution grid columns do not have to be newly initialized at each climate model time step, but use the information from the previous time steps as initial conditions (Khairoutdinov and Randall 2001). This allows to store information about the convective states inside the high-resolution grid columns from one time step of the climate model to the next. It can be seen as a way to incorporate memory effects of convective processes in a climate model, which is not the case for most other conventional convection parameterizations (Khairoutdinov and Randall 2001).

As a result the SP reduces the uncertainty of the represented convective processes in a climate model compared to conventional convection schemes (Khairoutdinov et al. 2005; Rasp et al. 2018). This leads to an enhanced realism with SP of the represented convective processes (Jones et al. 2019b; Khairoutdinov et al. 2005; Rasp et al. 2018) and of the related spatio-temporal variability (Khairoutdinov et al. 2005; Rasp et al. 2018).

However, the SP setup that I utilize in this thesis, has a few known limitations. One is that the horizontal advection of variables in the SP columns is only possible through the numerical core of the host climate model, e.g., from the nested high-resolution grid columns in one climate model grid cell to the ones in a neighbouring climate model grid cell. This may lead to a general damping of subgrid convection related variables (e.g., subgrid anomalies in specific humidity that vanish due to the averaging and the advection via the coarse grid of the host climate model) and an imperfect transport of cloud liquid and ice water concentrations from one nested to the adjacent nested SP (Jansson et al. 2022). More advanced SP setups enable the transport of tracers from one high-resolution SP to the next in the neighbouring climate model grid cell without relying on the numerical core of the host climate model (Jansson et al. 2022). A second limitation of the SP setup used in this thesis is that it simulates the effect of subgrid turbulence on thermodynamic variables, but there is no direct subgrid momentum transport from the high-resolution grid columns into the host climate model (Rasp et al. 2018). A third limitation is a double ITCZ bias when coupled to CESM with realistic topography and surface coupling (Woelfle et al. 2018).

Despite these known limitations, the SP that I use to optimize my machine learning algorithms in chapters 3 to 5 shows clear improvements to conventional convection schemes in terms of biases of convective processes in climate models when compared to storm resolving models. Examples here are an improved agreement of precipitation extremes (Rasp et al. 2018), an improved diurnal cycle of precipitation (chapter 5) and an improved reproduction of a MJO-like oscillation (chapter 3, Khairoutdinov et al. 2005). The increased realism of the

represented convective processes with a SP is a significant step to enhance our understanding about convection in a climate model (chapter 3) compared to using a conventional convection scheme.

The next part of the section introduces storm resolving models used for atmospheric modelling with horizontal resolutions between 1 km to 5 km that permit the explicit simulation of cumulus convection.

2.3.3. Storm Resolving Models

A Storm Resolving Model (SRM) is an atmospheric general circulation model with a horizontal resolution that allows to explicitly simulate convective storms (deep convective cells). This type of general circulation model typically has a horizontal resolution between 1 km to 5 km, which allows to model a large fraction of the effects of cumulus convection without the use of the parameterizations (Stevens et al. 2019). SRMs with their high-resolution bear the potential to break the "convective deadlock" (Gentine et al. 2018; Randall et al. 2003; Randall 2013), the reliance on imperfect subgrid convection parameterizations and related closure assumptions. However, there may still have to be shallow convection schemes applied in SRMs. Due to the high-resolution nature of SRMs, they were initially used to model convective processes (Grabowski et al. 1996) or the interaction of convection and radiation (Tompkins and Craig 1998) on regional scales in the order of hundred kilometers and less. Especially the modelling and the comparison against existing observational products from e.g., regional measurement campaigns was one of their first main applications (Grabowski et al. 1996; Xu and Randall 1996). One of the first global simulations with a SRM configuration was presented in Tomita and Satoh 2004, which evolved into the Nonhydrostatic Icosahedral Atmospheric Model (NICAM) (Satoh et al. 2008). NICAM does not include a hydrostatic assumption to simulate vertical velocities in its numerical core. Moreover the grid tiles have a triangular surface in horizontal direction, which alleviates biases due to the discretization in polar latitudes that a regular rectangular grid has. Similar to NICAM the quasi-global System for Atmospheric Modeling (SAM) presented in Khairoutdinov and Randall 2003 provided an improved simulation of convective processes and widened the understanding about convection and its interplay with the general atmospheric circulation. Especially for a realistic simulation of tropical convection and convection related variability SRMs proved to be valuable tools (Bony et al. 2015; Satoh et al. 2019). However comparing different SRMs participating in the Dynamics of the Atmospheric General Circulation Modeled On Non-Hydrostatic Domains (DYAMOND) project (Stevens et al. 2019) showed that there is a large degree of variability in model results related to convective processes, especially on short time and small horizontal scales. This can be seen for example in the cloud condensate fields associated with extratropical frontal systems. The same applies for mesoscale convective clusters in the tropics and maritime shallow convection over upwelling regions, where a large spread among the SRMS exists in DYAMOND (Stevens et al. 2019). An explanation for this may be that different SRMs have different implementations of microphysics, fine-scale turbulence schemes and how convective processes in the planetary

boundary layer are treated (Stevens et al. 2019). There are also indications that the vertical velocities, characteristics of updrafts and downdrafts related to convective processes may vary from SRM to SRM (Mooers et al. 2023).

Despite these uncertainties, SRMs provide a considerable improvement in the representation of precipitation patterns on the basis of daily averages and a more realistic energy cascade between large-scale and small-scale processes of atmospheric processes compared to climate models (Stevens et al. 2020). Furthermore, subseasonal variability that is related to convection like the MJO or structural features of tropical cyclones are realistically reproduced with SRMs (Satoh et al. 2019), which remains largely challenging with convectional climate models or ESMs.

Despite these advantages over coarse-resolution climate models, SRMs remain computationally expensive even on the latest high performance computers. DYAMOND (Stevens et al. 2019) was limited to 40 days, its successor Next Generation Earth System Models (nextGEMS) aimed at runs over 2 years. In both cases the high-resolution output of the SRMs were limited to a set of essential variables related to processes of interest e.g., convection. Even these limited high-resolution fields had to be postprocessed on the fly to reduce the original resolution to an amount of data that can be stored on disks. The postprocessed data with decreased resolution enabled a further evaluation due to lower memory requirements (Hohenegger et al. 2023).

Despite significant advances in the last two years in constructing an ESM-like multi-component SRM that allows coupled high-resolution simulations of ocean and atmosphere (Hohenegger et al. 2023), long-term simulations on current state-of-the-art high performance computers remains almost impossible. It is argued that SRM-like configurations without the use of a conventional convection scheme on coarser resolutions of 40 to 80 km enable longer simulations and might alleviate known biases of ESMs (Hohenegger et al. 2020). Thus such coarse SRM-like configurations showed a pronounced increase of biases in global averaged radiative fluxes compared to a reference SRM run (Hohenegger et al. 2020). As a result of that coarse SRM-like configurations that were run over 40 days may well introduce biases in longterm simulations similar to ESMs. A different option to enlarge the duration of SRM simulations may be a chain-like approach (Hoefler et al. 2023). In this case the SRM is run over a certain affordable period. Then a novel machine learning algorithm is used as a "gap-filler" towards another SRM simulation with different climate conditions. In the end, the resulting simulation may consist of several SRM runs that are connected together by predictions of a machine learning algorithm. Theoretically, such an approach may allow a hybrid-SRM simulation over decades and may enable projections of the future climate of the Earth system. Despite the attractiveness from a data science point of view, there remain certain challenges of such a chain-like approach. The large memory requirements of SRMs are only to some extent alleviated with the machine learning predictions. Moreover, the integration of machine learning may interfere with the SRM simulations leading to biases in reproduced processes, unintentional model drifts and in the worst case to complete model crashes. Therefore, a chain-like SRM approach requires both large computational efforts and expert knowledge in climate modelling to handle these challenges.

The computational costs can be strongly reduced with machine learning algorithms that train actively on SRM data and emulate certain processes in an SRM. In the recent years we saw a new phase of evolution of machine learning and associated hardware that resulted in neural networks that are well suited for both large data sets and complex non-linear relationships (Gentine et al. 2021; Reichstein et al. 2019). The following section of the thesis briefly illustrates the concept of machine learning and a few neural network structures that I am using in this thesis and in **Behrens** et al. 2022; **Behrens** et al. 2024. Additionally developments in machine learning for the parameterization of convective processes are discussed and put in the context of understanding convective processes and modelling with machine learning.

2.4. Machine Learning for the Parameterization of Convective Processes

This section begins by introducing three neural network architectures, that will form the foundations for the investigations in this thesis. Subsection 2.4.1 explains the class of artificial neural networks (ANNs). It is followed in subsection 2.4.2 by an explanation of an AutoEncoder Decoder (AED) and a Variational Auto Encoder (VAE) with a lower dimensional space, called "latent space" between the encoding and decoding parts of the network. After these subsections focused on model architectures, I will discuss in subsection 2.4.3 recent advances in machine learning for the parameterizations of convective processes. Afterwards I will introduce the field of stochastic machine learning in climate science in section 2.4.4.

2.4.1. Artificial Neural Networks (ANNs)

An Artificial Neural Network (ANN) is a class of neural networks, that has been originally used to learn and represent subgrid physical processes in ESMs (e.g., Gentine et al. 2018; Rasp et al. 2018). An ANN consists of an input layer, a set of fully connected layers behind the input layer and finally an output layer (Figure 2.5). First, a set of input variables, after some normalization, is fed into the input layer (i.e., its dimension correspond to the number of input variables). Normalizing the input variables in a multi-variate setup ensures that all have approximately similar importance during the ANN optimization (Rasp et al. 2018). The input layer is followed by a number of hidden layers where the computation takes place (Figure 2.5). Each hidden layer consists of nodes which are connected to all nodes (or neurons) of the previous and succeeding layer as it is shown in Figure 2.5. Therefore such a layer is called "fully-connected" in data science. The name "hidden layer" comes from the fact that these layers are situated within the ANN, between the input and output layers, and the computation taking place during training or prediction are not part of the final output of the ANN (Goodfellow et al. 2016). These hidden layers and their nodes allow the network to learn complex non-linear dynamical systems or processes, and are optimized during training (Goodfellow et al. 2016). Shallower ANNs with fewer hidden layers are in general favourable with respect to interpreting their predictions based on some inputs. However previous studies showed that

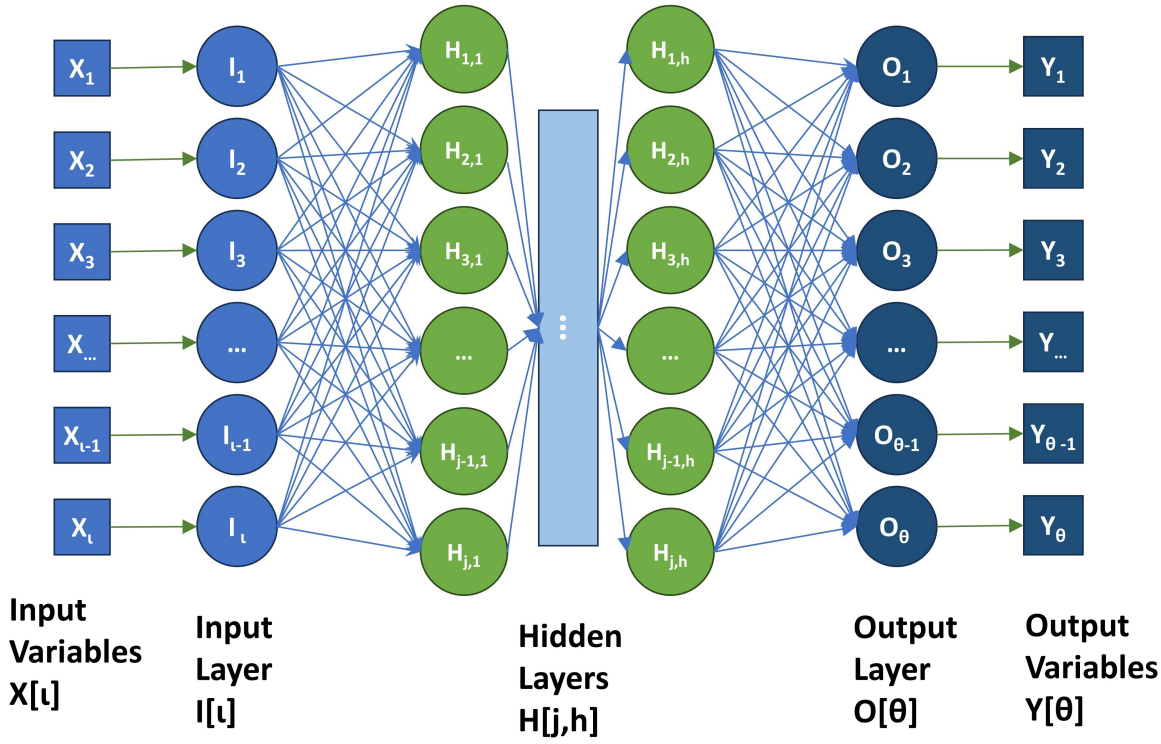

Figure 2.5.: Schematic of an Artificial Neural Network: The schematic shows a typical Artificial Neural Network (ANN). The network uses a set of input variables X[ι] that is fed into the input layer I[ι] with ι nodes. The input layer is fully connected to the first hidden layer H[j, 1] with j nodes. The information then propagates through in total h hidden layers. The final output layer O contains θ nodes corresponding to an array of output variables Y[θ]. This schematic is inspired by a similar one presented in Beucler et al. 2019.

deeper ANNs with a larger number of hidden layers have advantages in reproducing complex non-linear processes such as convection compared to shallower architectures (e.g., Gentine et al. 2018; Mooers et al. 2021; Rasp et al. 2018). Finally, the output layer transforms the processed information from the hidden layers into the ANN's output (Figure 2.5). Its purpose is to map the incoming information into the shape of the desired output variables (Figure 2.5) that the network is optimizing (Goodfellow et al. 2016).

So far, we have described the architecture (layers) of an ANN. Next, I describe how the individual layers are propagating information throughout the ANN, which is determined by the "activation" functions (Goodfellow et al. 2016).

The input layer of an ANN has typically a linear activation function, while the other layers usually include non-linear functions (Figure 2.5, Goodfellow et al. 2016). Equation 2.9 illustrates the purpose of these activation functions for one distinct node of a hidden or output layer. X denotes the value of one particular input variable. A linear regression is performed on the incoming signal X by adding a weight W_z and a bias b_z. The result of this linear function is then used as input for the activation function G() of the node. A linear activation function G() of the node Z would be simply the identity of $W_z \times X + b_Z$. For a non-linear activation, the result of the linear regression is transformed with a non-linear function G(). This allows

the ANN to adapt to non-linear relationships and breaks the linearity between X and Z. Due to the non-linear activation functions an ANN acts as a non-linear regression with a large number of internal degrees of freedom.

$$Z = G(W_Z \times X + b_Z) \tag{2.9}$$

With the definition of the response of each node Z to the incoming signals X in equation 2.9, the basis to optimize the weights and biases (parameters) of the ANN is set. During training, the aim is, therefore, to optimize the ANN's parameters applying a so called "loss function" or "cost function" $\mathcal{L}$ (Goodfellow et al. 2016). At the beginning of the training, all weights and biases are assigned usually to randomly drawn initial values. Then, the data set is passed to the ANN in batches (i.e., a number of training samples utilized in one iteration, Goodfellow et al. 2016). The "batch size" is a hyperparameter that can have a substantial impact on the ANN performace. Specifically, there exist general differences between a training with smaller and larger batch sizes. Small batch sizes may enable a faster convergence towards a maximum of optimization and a better generalization of the ANN, but introduce noise during the optimization (Goodfellow et al. 2016). Larger batch sizes may reduce the fluctuations in skill between different batches during training. However the training with large batch sizes needs larger computational resources due to the larger data amount that is fed into the ANN (Goodfellow et al. 2016). The ANN's predictions for each sample Y^{pred} are used to compute the loss function $\mathcal{L}$. This is also called "forward propagation" in data science (Goodfellow et al. 2016). A common metric used in $\mathcal{L}$ is the mean square error between Y^{pred} and the true data Y (i.e., equation 4.4, Goodfellow et al. 2016). In detail, in my thesis all constructed ANNs will utilize a Mean Squared Error (MSE) as loss function for multi-variate data sets. The gradient of the chosen loss function is then computed with respect to all weights and biases of the ANN, a process called "back propagation" in data science (Goodfellow et al. 2016). Back propagation ensures that the weights and biases of the ANN are optimized during training, accounting for each parameter's individual impact on the loss function via matrix multiplication. A second essential hyperparameter for the optimization of an ANN is the "learning rate". It defines the downward gradient step on the surface of the chosen loss function. A smaller learning rate may slow down the convergence towards a minimum of the loss function. However at the end of the optimization the distance towards a minimum may be smaller and the skill of the ANN may be improved compared to a training with larger learning rates (Goodfellow et al. 2016). In contrast, larger learning rates may accelerate the optimization and reduce the risk that the ANN get stuck into a local minimum. Despite this, larger learning rates may influence negatively on the optimization due to e.g., an oscillation of the algorithm between multiple minima of the loss function (Goodfellow et al. 2016). To summarize an optimization step, the ANN computes for each batch of data a new value of the loss function conditioned on the weights and biases. Afterwards a down-gradient step of the loss function (the learning rate) is applied, which is then backpropagated through the network and weights and biases are adjusted to it (Goodfellow et al. 2016).

An epoch consists of multiple optimization steps and is a complete pass through the entire training data set. To evaluate the robustness and especially an overfitting of the predictions of the ANN, the network is tested after each epoch to the validation data set. For the validation data set the loss function and performance metrics are computed, but the weights and biases are not adjusted. Overfitting occurs, if the training skill of the ANN is larger than the validations skill measured with the respective losses (Goodfellow et al. 2016). Reducing the overfitting may enhance the generalization of an ANN, which means that the ANN adaption to out-of-sample data (i.e., coming from unseen and different environmental conditions) is improved (Goodfellow et al. 2016). Over the course of the training, which may include a number of epochs, ideally the loss function is slowly decreasing. The training is ending after a predefined epoch or if an "early-stopping" is applied. The latter means that the training ends when i.e., the difference between training and validation skill of an ANN exceeds a predefined threshold (Goodfellow et al. 2016).

To increase the efficiency of training an ANN, there are a number of additional strategies. First, an extensive hyperparameter search, finding suitable initial learning rates and batch sizes, would help further optimize the ANN. Second, state-of-the-art optimizers, algorithms that perform a stochastic gradient descent via incorporating noise, can help overcome local minima of the loss function during training (Goodfellow et al. 2016). Third, shuffling the training data, so a permutation in every epoch that results in batches with varying loss statistics, introduces noise and helps to reduce overfitting. Finally, a learning rate schedule, which reduces the initial learning rate after a certain epoch, aids to achieve a more efficient training. In the first epochs a large learning rate ensures that the optimization is not getting stuck in the nearest local minimum. In latter epochs a smaller learning rate reduces the distance towards a specific minimum of the loss function and secures skillful predictions of the ANN. As an example a learning rate schedule was helpful to achieve a realistic reproduction of a multi-variate data set related to convective processes (Rasp et al. 2018).

To achieve a realistic reproduction of non-linear processes like convection, the respective ANNs tend to consists of multiple hidden layers and large total node sizes (i.e., the networks of Gentine et al. 2018; Rasp et al. 2018 or Mooers et al. 2021). As an example the ANN used in Rasp et al. 2018 had in the order of 500k trainable parameters or degrees of freedom with in total 9 hidden layers. Thus a quantification of the influence of an input variable on a specific output variable is cumbersome due to the complexity of the ANNs. Especially for the application of ANNs in Earth science, this lack of interpretability is unsatisfying (Mamalakis et al. 2022), where the focus may lie on improving the understanding of non-linear processes conditioned on a large-scale thermodynamic and dynamic background states. One suitable step forward to interpret the behaviour of an ANN is to use explainable artificial intelligence, determining the importance of the input variables on which the ANN makes its predictions (Saranya and Subhashini 2023). These explainable artificial intelligence algorithms have one general caveat, namely their computational cost. High qualitative explainable artificial intelligence, such as SHAP values (Lundberg and Lee 2017), allow a detailed interpretability of complex processes captured by a neural network. However, their applicability on climate data sets of a few hun-

dred gigabytes with multiple input and output variables remained challenging (Mamalakis et al. 2022). Nevertheless, SHAP values can be helpful to validate the interpretability of the underlying processes driving the dynamical system at hand (e.g., convective processes). An example to test the applicability of SHAP for multi-variate climate data, was the correct identification of spurious correlations in an ANN, i.e., convective processes in the lower troposphere that were driven by specific humidity in the stratosphere (Iglesias-Suarez et al. 2024).

2.4.2. Autoencoder Decoders (AEDs) and Variational Autoencoder Decoders (VAEs)

Therefore it is intuitive to ask the question whether models with a latent space, a lower-order manifold between the encoding and decoding part of the network, have the potential to obtain interpretability without relying on computational expensive explainable artificial intelligence. Such lower-order models will build the base for the general aim of the thesis to better understand subgrid convective processes in an ESM. In detail this thesis will evaluate the applicability of VAE and Encoder Decoder (ED) structures to obtain an improved understanding about convective processes in this context.

More generally, Variational Auto Encoder (VAE), but also AutoEncoder Decoder (AED) structures, can help investigate a lower dimensional representation of an input image (or input array) in the "latent space" (e.g., Kingma and Welling 2019). VAEs have even generative modelling capabilities, meaning that one could construct a new image by drawing a sample from the latent space and feeds it into the decoding part of the network (Kingma and Welling 2019). In mathematical notation, the key task of VAEs and AEDs are the mapping from input variables $\mathbf{X}$ to the reconstructed variables $\mathbf{X}^{pred}$. The respective function can be defined in pseudocode as X^{pred} = Decoder (Encoder(X)), where the lower-dimensional representation, also know as latent variables $\mathbf{z}$, within the latent space can be obtained via z = Encoder(X). While "latent dimension" indicates the space defined by one specific latent variable, "latent node" refers to the structural element inside the network of one particular latent variable, and "latent space width" is the overall dimensionality of the latent variables.

Figure 2.6 shows the overall structure of a VAE (upper) and an AED (lower panel). Both networks have a set of predefined input variables $\mathbf{X}$ that are fed into the Encoder. The Encoder decreases progressively the dimensionality of the incoming signal from layer to layer in both cases towards the latent space. The latent space is formed by a number of latent nodes N_{latent}. For an AED the latent variables $\mathbf{z}$ are directly determined by the Encoder. Whereas for a VAE the mean μ and the logarithmic variance $\ln \sigma^2$ are computed for every latent node in the setup that I am using (Kingma and Welling 2014). μ and $\ln \sigma^2$ are utilized in a so called "reparameterization" (Kingma and Welling 2014). This reparameterization maps the encoded distribution μ and $\ln \sigma^2$ on an isotropic Gaussian. The latent variable $\mathbf{z}$ is then drawn from the resulting reparameterized distribution (Kingma and Welling 2014). As a result of this difference in the latent space, an AED can be classified as a deterministic deep learning model, while a VAE has also a non-deterministic stochastic component. The latent variables $\mathbf{z}$ are

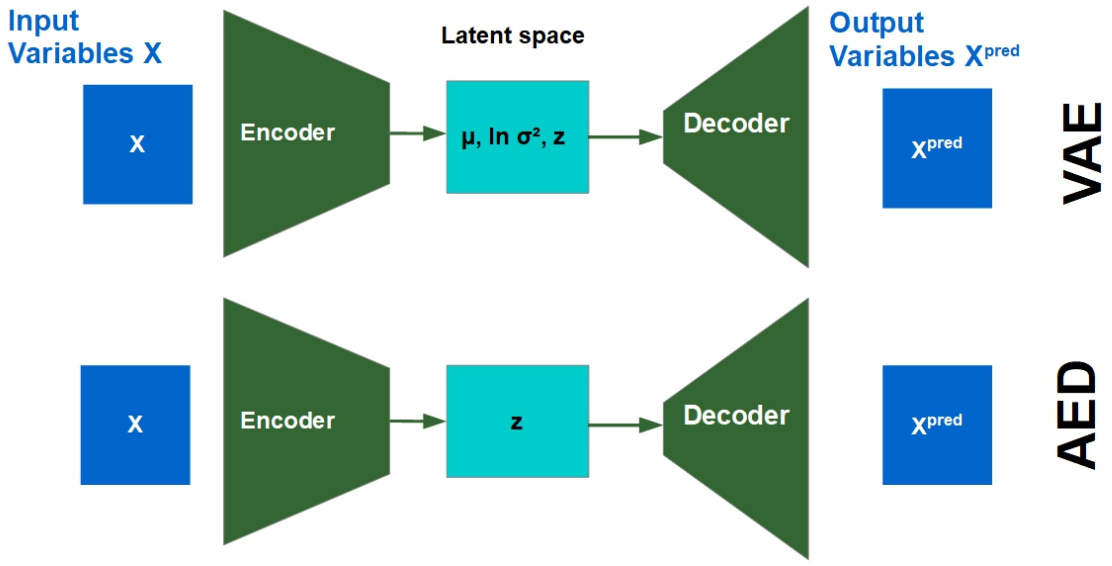

Figure 2.6.: Schematic of a Variational Auto Encoder (VAE) and AutoEncoder Decoder (AED). The VAE and AED uses a set of input variables **X** that is fed into the respective Encoder. The Encoder maps the information into a latent space that includes a mean μ, logarithmic variance ln σ^2 and the latent variable z in the case of the VAE. For the AED the latent space only consists of z. For both networks z is fed into the Decoder that increases the dimensionality to the output layer, returning the reconstructed variables $\mathbf{X}^{pred}$.

the only input to the Decoder. The Decoder progressively increases the dimensionality of the propagated signals from layer to layer towards the output layer, that is formed by the set of reconstructed variables $\mathbf{X}^{pred}$.

Equation 2.10 shows a standard VAE loss (based on the equation shown in Mooers et al. 2023). The first term indicates the reconstruction error between the input variables $\boldsymbol{x}$ and the generated samples $p_{\eta(x|z)}$ of the Decoder p_η, based on $\boldsymbol{x}$ and the latent variables z (Kingma and Welling 2014). The generative model's parameters is indicated by $\boldsymbol{\eta}$, and the subscript q_γ symbolises the sampling from the latent space or variational distribution determined by the Encoder (Kingma and Welling 2014). Sampling from the latent space is based on the latent variables z conditioned on $\boldsymbol{x}$, where $\boldsymbol{\gamma}$ indicates the variational parameters. The second term of Equation 2.10 is the Kullback-Leibler (KL) loss term. It depends on the KL divergence between the distribution represented by μ and ln σ^2 and an isotropic Gaussian distribution. N_{batch} is the respective batch size and N_{latent} is the number of latent dimensions. λ is an positively defined annealing factor (Alemi et al. 2018). λ increases through the course of the training and gives the KL term increasing relative importance from epoch to epoch compared to the reconstruction error. Alternatively, λ can be set to a constant that regularizes the KL term to increase the reproduction capabilities of the VAE to the expanse of disentanglement inside its latent space. In a deterministic setup (deterministic loss function), the expected reconstruction error can be calculated by a squared error between $\boldsymbol{X}$ and the reconstructed input variables $\boldsymbol{X}^{pred}$. For the AEDs, the respective loss function is determined by the reconstruction errors without the additional KL term.

$$\mathcal{L}_{\text{VAE}}\left(\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{x}\right) = \mathbb{E}_{q_{\gamma}(z|\boldsymbol{x})}[\log p_{\eta}(\boldsymbol{x}|z)] + \lambda \times \underbrace{\frac{1}{2N_{\text{batch}}} \sum_{k=1}^{N_{\text{batch}}} \sum_{z=1}^{N_{\text{latent}}} \left(\mu_{z,k}^2 + \sigma_{z,k}^2 - \ln \sigma_{z,k}^2 - 1\right)}_{\text{KL Loss}}, \ \lambda \in \mathbb{R}^+ \tag{2.10}$$

Building machine learning convective parameterizations, however, requires predicting the effects of subgrid convective processes on the large-scale state of the system. Variational Encoder Decoder (VED) networks enable a $\mathbf{X} \rightarrow \mathbf{Y}^{pred}$ mapping, with varying input $\mathbf{X}$ and output variables $\mathbf{Y}$ (see chapter 3). For example, thermodynamic state variables, such as temperature profiles $\mathbf{T}$(p) and specific humidity profiles $\mathbf{q}$(p), may be part of the input fields $\mathbf{X}$. While $\mathbf{Y}$ includes the temporal derivatives of the respective state variables that reflect the effects of subgrid processes. This key difference translates also into a modified loss function (equation 2.11) compared to the traditional ones used for VAE networks (equation 2.10).

$$\mathcal{L}_{\text{VED}}\left(\boldsymbol{Y}, \boldsymbol{Y}^{\mathbf{pred}}\right) = \text{MSE}\left(\boldsymbol{Y}, \boldsymbol{Y}^{\mathbf{pred}}\right) + \lambda \times \underbrace{\frac{1}{2N_{\text{batch}}} \sum_{k=1}^{N_{\text{batch}}} \sum_{z=1}^{N_{\text{latent}}} \left(\mu_{z,k}^2 + \sigma_{z,k}^2 - \ln \sigma_{z,k}^2 - 1\right)}_{\text{KL Loss}}, \ \lambda \in \mathbb{R}^+ \tag{2.11}$$

Equation 2.11 shows a VED loss case (see chapter 3), that uses the MSE between the true output variables $\boldsymbol{Y}$ as reconstruction term in contrast to traditional VAEs.

For the general aim of the thesis to understand convective processes simulated and reproduced in $\boldsymbol{Y}$ or $\boldsymbol{Y}^{pred}$ it is intuitive to construct a direct mapping $\boldsymbol{X} \rightarrow \boldsymbol{Y}^{pred}$ with our VED and EDs, and not the traditional mapping $\boldsymbol{X} \rightarrow \boldsymbol{X}^{pred}$. In chapter 3, I will even show that a combined mapping $\boldsymbol{X} \rightarrow \boldsymbol{Y}^{pred} + \boldsymbol{X}^{pred}$ has clear advantages over the $\boldsymbol{X} \rightarrow \boldsymbol{Y}^{pred}$ for the general interpretability of convective processes in the latent space of a VED. To evaluate the uncertainties quantification related to stochasticity of convective processes, I will restrict the mapping to $\boldsymbol{X} \rightarrow \boldsymbol{Y}^{pred}$ of VEDs in the two succeeding chapters 4 and 5.

2.4.3. Recent Advances in Parameterizing Convective Processses with Machine Learning

As the previous section 2.3 suggests, convective processes in ESMs bear large uncertainties due to their reliance on subgrid convection schemes. A superparameterization or SRMs are high-resolution alternatives but remain computationally expensive. Therefore, machine learning algorithms that skillfully learn from such high-resolution convective processes are a valuable alternative (Gentine et al. 2018; Gentine et al. 2021). These machine learning algorithm could shape our understanding about convective processes in ESMs (Gentine et al. 2021). Likewise machine learning algorithms may provide realistic uncertainty quantification of convective processes that are a further key information to improve our understanding about convective processes quantitatively (see chapter 5). When these machine learning algorithms are coupled

into the numerical core of a climate model they have a fraction of the computational costs of the original superparameterization (Rasp et al. 2018) or the SRM (Krasnopolsky et al. 2013).

In recent years machine learning subgrid processes based on high resolution training data has flourished (Gentine et al. 2021). One of the first studies that replaced an existing radiation scheme with a ML based version was presented in Krasnopolsky et al. 2010. Furthermore it was shown that neural networks are able to learn subgrid convective processes based on SRM simulations (Krasnopolsky et al. 2013). Thus they were applied as a data-driven parameterization in a climate model in pioneering experiments (Krasnopolsky et al. 2013). This data-driven scheme had a compatible skill as traditional convection schemes in a decade-long simulation over the tropical Pacific Ocean (Krasnopolsky et al. 2013).

A few years later in a second phase of machine learning for climate modelling it was proved that these subgrid convective processes from a superparameterization can be deep learned with an ANN on global scales (Gentine et al. 2018). The ANN reproduced global temperature tendency fields dT/dt (similar to Q1 in equation 2.5 of the Yanai scheme in section 2.3), specific humidity tendencies dq/dt (similar to Q2 in equation 2.6), the longwave heating and shortwave heating rates skilfully (Gentine et al. 2018). In the analysis the ANN predictions were compared against data of the superparameterized Community Atmosphere Model (CAM) in an aquaplanet setup. This is a simulation where topography is excluded and the entire model domain is covered by an ocean as lower boundary of the atmospheric model. In a follow-up study it was proved that an ANN can be used instead of a traditional subgrid convection parameterization in a global aquaplanet simulation with CAM (Rasp et al. 2018). The used ANN reproduced a large portion of the spatio-temporal variability related to convective processes that a superparameterization would have but a convection scheme could not represent (Rasp et al. 2018). The initial hybrid model with machine learned subgrid parameterizations was followed by examples based on a random forest (Yuval and O'Gorman 2020) or an ANN (Brenowitz and Bretherton 2019) with the System for Atmospheric Modeling (SAM) in an aquaplanet setup. Apart from aquaplanets, ANN or residual neural networks were able to reproduce subgrid processes in a real geography setup in prognostic tests (Han et al. 2020; Mooers et al. 2021). This was a further step towards the use of deep learning parameterization for convective processes in an ESM. Wang et al. 2022b and Han et al. 2023 also showed that their architectures allowed to run stable hybrid model simulates in CAM and CESM2 over the course of a few years. Despite these advances Wang et al. 2022b used an atmosphere only setup, while Han et al. 2023 side-stepped deep learning surface radiative fluxes that are essential for coupling to the other model components of an ESM. This thesis will build on these limitation and shows ways forward in this respect. Recently it was showed that causal discovery can be used to improve the representation of subgrid processes with an ANN in hybrid simulations via the identifications of spurious correlations between inputs and outputs of the ANN (Iglesias-Suarez et al. 2024). Moreover causal discovery enabled an improved understanding of convective processes in this respect (Iglesias-Suarez et al. 2024).

Apart from causal discovery and explainable artificial intelligence, VAEs and VEDs may have the potential to improve the understanding about convective processes, that will be

investigated in this thesis. Initially VAEs were used to gain an understanding about the stochasticity and chaotic processes of the Lorenz 96 model (Lorenz 1996; Tibau Alberdi et al. 2018) or to cluster different phases of the boreal polar vortex based on reanalysis data (Krinitskiy et al. 2019). A VAE was also used to investigate characteristics of simulated tropical convection with its latent space (Mooers et al. 2020). During the preparation of this thesis and the related paper (**Behrens** et al. 2022), that I will show in chapter 3, two more studies complemented the advances in improving the general understanding of convective processes with interpretable deep learning models. The latent representation of an AED helped an ANN to improve the prediction of extreme precipitation (Shamekh et al. 2023). This study showed that the latent space stored key information about the convective aggregation and enhanced both the interpretability and reproduction of convective processes (Shamekh et al. 2023). The interpretability of the latent space of a VAE allowed also a model intercomparison of SRMs (Mooers et al. 2023). This study indicated via a latent space analysis that considerable differences of vertical velocities exists related to convective processes between the different SRMs (Mooers et al. 2023).

2.4.4. Stochastic Machine Learning in climate science

Despite these recent advances in modelling and understanding convective processes with machine learning discussed above, known limitations of these architectures remained. One limitation is the weak reproduction skill of ANNs for convective processes in the planetary boundary layer. It has been hypothesized (i.e., Gentine et al. 2018; Mooers et al. 2021), that the weak reproduction with single deterministic ANNs might be related to stochasticity. Individual models may have deficiencies in representing to some extent stochastic processes (Han et al. 2023). This limitation of deterministic algorithms appeared also in Earth system modelling in the past. It led to the development of ensembles, to obtain an improved estimate of the average subgrid convective processes in a climate model (Jones et al. 2019a, 2019b). Such deterministic ensembles may have limitations in reproducing the variability of convective processes (Jones et al. 2019b). To improve the representations of subgrid variability in climate models stochastic schemes were developed (Berner et al. 2017). Some of these traditional approaches to generate stochasticity were translated from numerical weather prediction to climate modelling. One example in this context is the Stochastic Perturbed Parameter Tendencies scheme (SPPT)(Buizza et al. 1999; Christensen et al. 2015), a scheme where subgrid tendency terms are perturbed with multiplicative random noise. Building upon SPPT, further work improved the representation of subgrid turbulence and air-sea interactions (Bessac et al. 2021) and showed the possibility to upgrade the scheme with information from SRM simulations (Christensen 2020). Despite these advances in traditional stochastic parameterizations, stochastic deep learning is still in its infancy in climate science as of today. Various studies (i.e., Bhouri and Gentine 2022; Gagne II et al. 2020; Parthipan et al. 2022) showed that stochastic machine learning improve estimating uncertainties of the conceptual Lorenz 96 model (Lorenz 1996). A stochastic entrainment and detrainment scheme for shallow convection based on a

Monte Carlo dropout, where a certain percentage of the links inside a network are randomly clipped during the repetitive predictions, outperformed an existing traditional scheme (Shin and Baik 2022). A Conditional Generative Adversarial Network (CGAN) enabled a realistic estimation of uncertainties of moistening and heating profiles related to tropical maritime convective processes (Nadiga et al. 2022). The representation of shallow convection was improved with a stochastic multi-plume scheme related to dry and shallow convective processes (Chinita et al. 2023). Moreover characteristics of cloudiness in SRM simulations over Germany were corrected with an stochastic sampling of mass fluxes at the cloud bases (Sakradzija and Klocke 2018). Apart from atmospheric convective and turbulent processes, in oceanography stochastic neural network proved to be valuable tools to realistically reproduce the effects of subgrid eddies on the oceanic general circulation (Guillaumin and Zanna 2021; Perezhogin et al. 2023).

Despite these great opportunities of deterministic deep learning ensembles and stochastic deep learning discussed above, there exist two limitations of the latter. State-of-the-art stochastic deep learning schemes focused on quantifying uncertainties for individual variables or processes. However calibrated uncertainty quantification for a multi-variate data set of subgrid processes i.e., like with a superparameterization are not investigated yet. Therefore, chapter 5 will show a detailed analysis in this context. Furthermore for such multi-variate input and output data sets individual stochastic deep learning networks have in general a weaker reproduction skill compared to individual deterministic counterparts (Yu et al. 2023). In this work, I trained an ED based on the network architecture, that will be introduced in chapter 3, with a strongly reduced dimensionality in its latent space of only five nodes. It was shown that this ED had an improved reproduction of subgrid variables compared to the participating hyperparameter-tuned stochastic models (Yu et al. 2023). So it is intuitive to wonder whether there is a potential to combine both, accurate reproduction skills of deterministic models with realistic uncertainty estimates for stochastic convective processes. This topic will be covered in chapter 4.

3. Understanding Convective Processes in a Climate Model using Non-Linear Dimensionality Reduction of a Variational Encoder Decoder

The following chapter is reproduced from **Behrens** et al. 2022. In this chapter I will present ways to investigate convective regimes and large-scale drivers of convective processes based on the latent space of one Variational Encoder Decoder using generative modelling. This chapter is structures as follows. Section 3.1.1 explains the climate model that is used in this chapter. Section 3.1.2 illustrates the Variational Encoder Decoder that is designed to understand convective processes and large-scale drivers together with other benchmarking machine learning algorithms. It is followed by section 3.2 that evaluates the reproduction and encoding capabilities with respect to the interpretability of the latent space of the Variational Encoder Decoder. Section 3.3 contains the latent space investigation of the Variational Encoder Decoder to understand the encoded large-scale drivers of convection and distinct convective regimes. Section 3.4 summarizes the key results of **Behrens** et al. 2022. For **Behrens** et al. 2022 I, as the author of the thesis, contributed all figures, tables and large parts of the code to produce them. In the Figure 3.11 of **Behrens** et al. 2022 summarizing this chapter one published schematic from Schneider et al. 2017 was added together with the reference pointing to the original publication. Furthermore I let the writing and the analysis of the published paper.

3.1. Data and Methods

This section reproduces the section Data and Methods of **Behrens** et al. 2022 with negligible modifications.

3.1.1. Data: Superparameterized Aquaplanet Simulation

We use a 2-year aquaplanet simulation of the superparameterized Community Atmosphere Model v3.0 (SPCAM) (Collins et al. 2006; Khairoutdinov et al. 2005) under the configuration of Pritchard and Bretherton 2014 in which Sea Surface Temperatures (SST) were imposed following a realistic zonally symmetric distribution (Andersen and Kuang 2012). The SST maximum in the tropics is slightly displaced to 5° N and decreases meridionally towards

the poles to reduce exact equatorial symmetry. The solar forcing is fixed to Austral Summer conditions (no seasonal variability), but includes diurnal variability. The model has a coarse horizontal resolution corresponding to a typical grid size of 300 km near the equator. The vertical axis extends from the surface to ~ 40 km (3.5 hPa) following a hybrid coordinate with 30 levels (22 levels below 100 hPa). The GCM uses a 30-minute time step. Following Pritchard et al. 2014, the superparameterized (SP) component consists of 8 nested 2D columns oriented meridionally on the same vertical axis and with a subgrid size of 4 km (Grabowski 2001; Khairoutdinov and Randall 2001). Deep convection is explicitly resolved every 20 seconds and a Smagorinsky 1.5-order turbulence closure, and a one-moment microphysics parameterization (Khairoutdinov and Randall 2003) are used. SPCAM in this configuration yields a realistic reproduction of the ITCZ and tropical wave-spectra with a pronounced Madden-Julian-Oscillation (MJO)-like signal, as well as improved precipitation distributions compared to the host GCM (CAM, Pritchard et al. 2014). However, this SPCAM setup neglects momentum transport, and for our approach, we sidestep the SP of cloud ice and water sources and sinks and instead emulate their radiative consequences through the total diabatic heating, as in Rasp et al. 2018.

3.1.2. Model: Variational Encoder Decoder

We develop a variational encoder decoder (see schematic in Figure 3.1) to holistically learn subgrid-scale processes in SPCAM. VAEs traditionally reproduce their inputs, e.g., learning a mapping from large-scale variables to themselves. Here, our goal is to map large-scale to subgrid-scale variables. Therefore, we adopt a variational encoder decoder (VED) architecture to include the emulation of subgrid-scale variables. We include convection, turbulence, and radiation by simultaneously predicting the total diabatic heating and moistening tendencies alongside a decoded reconstruction of the relevant input data that summarize local large-scale state information prior to radiative-convective adjustment. Compared to deep feed-forward neural nets, the variational encoder decoder enhances the interpretability of convective processes and how they are connected to the driving large-scale climate via its latent space of reduced dimensionality. Regarding the input fields (**X**), we closely mirror the established precedent of Rasp et al. 2018 by using profiles of specific humidity **q(p)** in $\frac{kg}{kg}$ and temperature **T(p)** in K on 30 vertical levels each, as extracted from the end of the host model dynamics or the beginning of the physics package. **X** additionally includes the scalar values of solar insolation $\mathbf{Q}_{sol}$ in $\frac{W}{m^2}$, surface latent heat flux $\mathbf{Q}_{lat}$ in $\frac{W}{m^2}$ and surface sensible heat flux $\mathbf{Q}_{sens}$ in $\frac{W}{m^2}$, and surface pressure $\mathbf{P}_{surf}$ in Pa. That is, **X** is a concatenation of these two vectors and four scalars, [**q(p)**, **T(p)**, $\mathbf{Q}_{sol}$, $\mathbf{Q}_{lat}$, $\mathbf{Q}_{sens}$, $\mathbf{P}_{surf}$], into a 64-element input vector. The variational encoder decoder is trained to predict **O**, which combines the reconstruction of the same large-scale input data (as described above) with the subgrid-scale process rate output fields targeted by Rasp et al. 2018 **Y** (i.e., a parameterization): vertical profiles of total diabatic specific humidity tendency **dq(p)/dt** in $\frac{kg}{kg \times s}$ and total diabatic temperature tendency **dT(p)/dt** in $\frac{K}{s}$ defined on 30 pressure levels, as well as scalar values for shortwave and longwave radiative heat fluxes

at the model top ($\mathbf{Q}_{sw\ top}$ and $\mathbf{Q}_{lw\ top}$) and at the surface ($\mathbf{Q}_{sw\ surf}$ and $\mathbf{Q}_{lw\ surf}$) in $\frac{W}{m^2}$, and precipitation rate **precip** in $\frac{m}{s}$.

The full predicted vector $\mathbf{O}$ = [**dq(p)/dt**, **dT(p)/dt**, $\mathbf{Q}_{sw\ top}$, $\mathbf{Q}_{sw\ surf}$, $\mathbf{Q}_{lw\ top}$, $\mathbf{Q}_{lw\ surf}$, **precip**, **q(p)**, **T(p)**, $\mathbf{Q}_{sol}$, $\mathbf{Q}_{lat}$, $\mathbf{Q}_{sens}$, $\mathbf{P}_{surf}$] has a dimension of 129.

As it will be the main ML model used in this study, we henceforth abbreviate the variational encoder decoder structure simultaneously predicting subgrid-scale convective processes and large-scale climate conditions to "VED" for simplicity. A prior experiment with a $\text{VED}_{X\rightarrow Y}$ that was trained on $\mathbf{X}$ to predict $\mathbf{Y}$, similar to the established precedent of Rasp et al. 2018, does not encode the large-scale climate variables $\mathbf{X}$ as much in its latent space compared to VED. This limited our ability to gain insight into convective predictability with $\text{VED}_{X\rightarrow Y}$ (see Appendix A with supporting material section A.4.1 and Figure A.16 for details). In contrast the combined reproduction of subgrid-scale processes and large-scale climate variables with VED together with our generative modeling method allows us to explore convective regimes and corresponding large-scale climate conditions.

The encoding part of the VED (Encoder) consists of 6 hidden layers, which progressively reduce the dimensionality from 463 nodes in the first hidden layer down to 5 nodes (the latent variables) in the latent space. These values were chosen following a formal hyperparameter search (see the Appendix A and section A.2). We will test the sensitivity of emulations of the VED with respect to the number of latent nodes in section 3 in detail. In the following we will refer to one distinct latent variable in the context of the network architecture as "latent node". While we will use the notation "latent space" for the manifold spanned by all latent variables. Within this latent space, the mean μ and logarithmic variance $\ln \sigma^2$ are computed for each node, where σ is the standard deviation of the posterior (Kingma and Welling 2014). Then a so-called 'reparameterization trick' (Kingma and Welling 2014) is utilized to map the original distribution based on μ and $\ln \sigma^2$ onto an isotropic gaussian distribution. We used the $\ln \sigma^2$ instead of σ^2 for the construction of the network to simplify the reparameterization and the computation of the VED loss. The resulting latent variables $\mathbf{z}$ (5 dimensions) are used to investigate convective processes and drivers of convective predictability. Henceforth we will use the notation "latent dimension" to describe the subspace spanned by one particular latent variable. We will show in section 3.3 that characteristic convective regimes and large-scale climate states are encoded in $\mathbf{z}$. The latent variables $\mathbf{z}$ are the only input fed to the decoding part of the VED (Decoder), which reconstructs both large-scale and subgrid-scale fields. In the decoder, the dimensionality is progressively increased to 463 in the last hidden layer before the 129-node output layer. We use the rectified linear unit (relu) as activation function of all hidden layers of the Encoder and Decoder except for the Decoder output layer, where we use an exponential linear unit (elu) based on prior hyperparameter testing (see S.1). In the latent space, μ and $\ln \sigma^2$ are linearly activated, whereas for the latent variables $\mathbf{z}$ we call the reparameterization function. In summary, the Encoder and Decoder of the VED consist of 388,440 and 418,469 total trainable parameters, respectively.

We train the VED over 40 epochs (number of iterations through training data), during which the weights and biases are updated to minimize the VED loss function (see Equation 3.1).

$$\text{VED loss} = \text{reconstruction loss} + \lambda \text{ KL loss} \tag{3.1}$$

The loss function is the sum of a reconstruction and a Kullback-Leibler (KL, Equation 3) loss term. The first term measures the mean-square error (MSE, Equation 3.2) between the predicted ($\mathbf{O}^{emul}$) and the ground truth data ($\mathbf{O}$).

$$\text{reconstruction loss} = \frac{1}{M} \times \frac{1}{N} \sum_{i=1}^{(M=129)} \sum_{j=1}^{(N=\text{batch size})} (O_{ij} - O_{ij}^{emul})^2 \tag{3.2}$$

The KL loss term can be interpreted as a regularizer of the resulting latent distributions (Kingma and Welling 2014), which penalizes the complexity in the latent space based on the KL divergence.

$$\text{KL loss} = \frac{1}{2} \times \frac{1}{N} \sum_{j=1}^{(N=\text{batch size})} \sum_{k=1}^{(K=\text{latent space width})} \left[-1 - \ln \sigma_{jk}^2 + \mu_{jk}^2 + \sigma_{jk}^2 \right] \tag{3.3}$$

$$\lambda \, \epsilon \, \mathbb{R}_+ \tag{3.4}$$

We apply a KL annealing approach that multiplies the KL loss term by an annealing factor λ (equation 3.4) with initial value 0. The annealing factor then grows after a certain epoch during the training process (Alemi et al. 2018). This generally improves the reproduction capabilities of VAEs due to lowering the impact of the regularizing KL term (Mooers et al. 2020), avoiding a posterior collapse (Alemi et al. 2018), which negatively affects training. During a training step a 2D batch (dimensions 714×64) of 714 samples, the batch size, is fed into the VED to optimize the weights and biases. We use Adam as the VED's optimizer (Kingma and Ba 2014). The purpose of an optimizer is to improve the networks performance (minimization of the networks loss function in our case) during the training process based on stochastic gradient descent. We choose this particular optimizer to follow the same strategy like in the preceding study of Rasp et al. 2018. The learning rate (the applied down-gradient step to optimize the loss) has an initial value of 0.00074594 based on a formal hyperparameter tuning and is divided by factor 5 after every 7^{th} epoch over the course of the training. The batch size and the initial learning rate were chosen based on a formal hyperparameter search. Further optimized hyperparameters and a description of the hyperparameter search can be found in Appendix A Table A.1 and section A.2. The chosen hyperparameters represent a suitable local minimum for the optimization of the VED architecture but should not be considered as the optimal hyperparameter setting.

3.1.3. Benchmarking

To benchmark the performance of our VED, we construct three reference networks with different architectures. The first reference network is an Encoder Decoder (ED). The ED closely mirrors the architecture of the VED except that there is no KL regularization, meaning

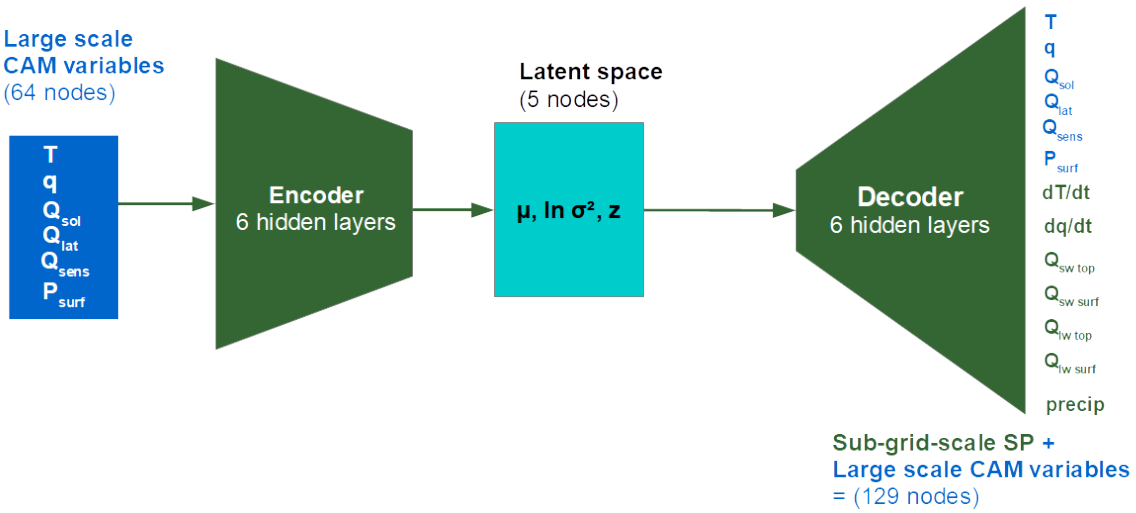

Figure 3.1.: Schematic of the constructed VED which uses large scale CAM variables to investigate simulated subgrid-scale convective processes of SP. The latent space consists of mean μ, a logarithmic variance $\ln \sigma^2$ layer and the latent variables **z**. The output data **O** of the decoder includes a reconstruction of the input data **X** to the encoder to encourage a latent space that can additionally compress the large- scale climate variables, in addition to their mapping to the target subgrid-scale fields **Y**. This Figure was directly reproduced from **Behrens** et al. 2022.

that the calculation of $\ln \sigma^2$ and μ is omitted. Furthermore, the ED's loss function only relies on the reconstruction loss. The second reference network, LR, is a further simplification of the ED, for which linear activations are used, which can be viewed as an equivalent to a principal component regression except that the latent space is not orthogonal. That is, the LR network can be interpreted as the combination of linear dimensionality reduction and regression modules. We use a reference deep artificial neural net (reference ANN) with its original output normalization based on Rasp et al. 2018, which was proven to be a skilful emulator of SPCAM. Note that to reproduce Rasp et al. 2018, meridional wind profiles were used as input fields to construct and train the reference ANN network. As an additional baseline model, we implement a linear version of our reference ANN. Similar to the reference ANN, this "Reference Linear Model" uses 256 nodes and 9 hidden layers but replaces all of the ANN's activation functions with the identity function (i.e. passing the values unchanged). Finally, we constructed one further VED structure and a conditional VAE in the run-up of this study, which are presented in the supporting material (see Appendix A section A.4) together with their strengths and limitations. Our goal is to strike a balance between the successful emulation of the target subgrid-scale output data **Y** with compression, and the usefulness of scientific interpretation for convective processes and large-scale climate states. The VED we have chosen (see Figure 3.1) is optimal on these fronts.

We split the SPCAM simulation into space-time shuffled training, unshuffled validation and unshuffled test data sets spanning 3 months (~ 4400 time steps) each. The input data **X** is normalised by subtracting the mean of each variable at each vertical level and dividing

by the range between minimum and maximum of the resulting anomalies. Furthermore, we normalize the output of the VED, ED and LR as described in the Appendix A (see section A.2). The output normalization, i.e., scaling to the same order of magnitude, allows us to achieve comparable reproduction skills across all fields. We show the impact of the existing differences of the VED output normalization and the reference ANN output normalization (Rasp et al. 2018) on the evaluation of mean reproduction skills of the networks in section A.3 in the Appendix A.

In the next section we will evaluate the performance of the VED with respect to common reproduction metrics, and discuss the interpretability of the information encapsulated in the latent space.

3.2. Evaluation of the VED

This section is directly reproduced from **Behrens** et al. 2022 with negligible modifications.

In this section, we assess the predictive skill of the VED, and compare its mean regimes / statistics and tropical variability against reference networks. Furthermore, we evaluate the interpretability of the VED's latent space with respect to climate and convective variables. With this analysis, we are investigating the overall decoding (reproduction) and encoding (dimensionality reduction, interpretability) abilities of the VED to learn convective processes.

3.2.1. Mean Regimes and Statistics

We start by evaluating the accuracy of the VED predictions to assess the impact of its dimensionality reduction on the overall performance. We use the mean squared error (MSE) to assess the performance of the VED predictions across subgrid scale fields $\mathbf{Y}$ for the training, validation, and test sets based on our VED output normalization. Overall, the VED shows good reproduction skills (see Appendix A Table A.4). The VED (test MSE = 0.165) clearly outperforms the linear model LR (test MSE = 0.243) in all data sets. The difference in predictive skills between VED and ED (test MSE = 0.165) is negligible. However, both networks express increased but comparable MSE with respect to reference ANN (test MSE= 0.135), in spite of the reference ANN having a substantially larger dimensionality (no latent manifold with a dramatic dimensionality reduction down to 5 nodes). These results are robust to the choice of output normalization (VED's versus reference ANN's, Rasp et al. 2018), as demonstrated in the Appendix A.3 section A.3.

In the following, we explore whether a latent space of 5 nodes is a good compromise between accuracy to reproduce convective processes and physical interpretability in the latent space. Figure 3.2 shows the VED performance (MSE) on test, validation, and training data as a function of the latent space width. We find a substantial sensitivity of the VED's performance to the latent space width - smaller width results in reduced accuracy associated with increased dimensionality reduction. Even for a latent space of two nodes, the VED has a higher predictive skill than the reference linear model, confirming the necessity of using nonlinear models to

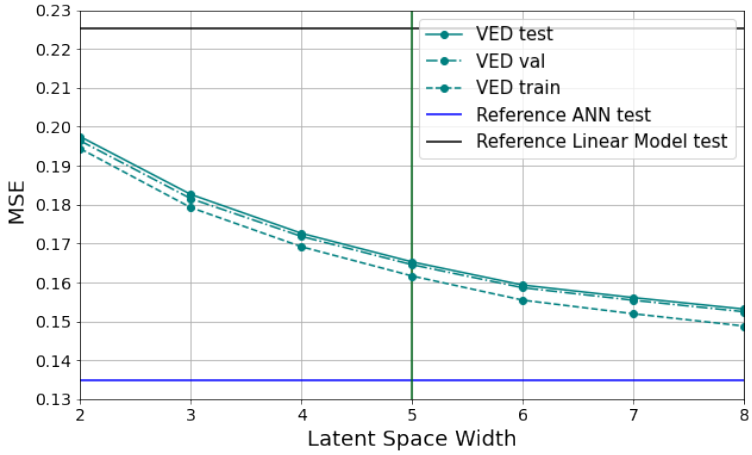

Figure 3.2.: Mean Squared Error (MSE) as a function of Latent Space Width of the VED for test (solid cyan), validation (dashed-dotted cyan) and training data set (dashed cyan curve) using our VED output normalization. The horizontal solid blue / black line represents the MSE scores of the reference ANN of Rasp et al. 2018 / a linear version of this network (Reference Linear Model) on test data with fixed layer width of 256 nodes in the 9 hidden layers. This Figure was directly reproduced from **Behrens et al. 2022**

faithfully represent subgrid-scale processes. Moreover, the VED's performance is converging towards the reference ANN for larger latent space widths (8 nodes). A latent space of 5 nodes results in a small reduction of predictive skills compared to the 'wider' latent space (Figure 3.2), indicated by a MSE decrease of only ≈ 0.012 between a latent space of 5 nodes and 8 nodes. Additionally, we will show later (in section 3.3) that such a latent space width enables the characterisation of realistic convective regimes and drivers of convective processes on specific nodes. This suggests that the overlap between different nodes is small. Despite this small overlap, we will show in section 3.3 that the resulting five latent nodes govern both SP convective processes and CAM climate states in most cases. For larger latent space widths of 6 nodes and more, the interpretability of resulting convective regimes gets more challenging due to the decaying impact of one latent node, or increasingly concurring influences between the nodes on SP convective processes or CAM climate variables. To summarize, regardless of how the output data are normalized (see Appendix A Figure A.1), the VED performs better than the reference linear model and approaches the performance of the fully-connected reference ANN as the latent space width increases.

As a complementary metric to evaluate the performance of the VED, we use the Coefficient of Determination ($\mathbf{R^2}$) (Equation 3.5).

$$\mathbf{R^2} = 1 - \frac{\mathbf{MSE}}{\mathbf{Var}} \tag{3.5}$$

$$\mathbf{MSE} = \frac{1}{\mathbf{P}} \sum_{t=1}^{\mathbf{P}} (\mathbf{Y_t} - \mathbf{Y_t^{emul}})^2 \tag{3.6}$$

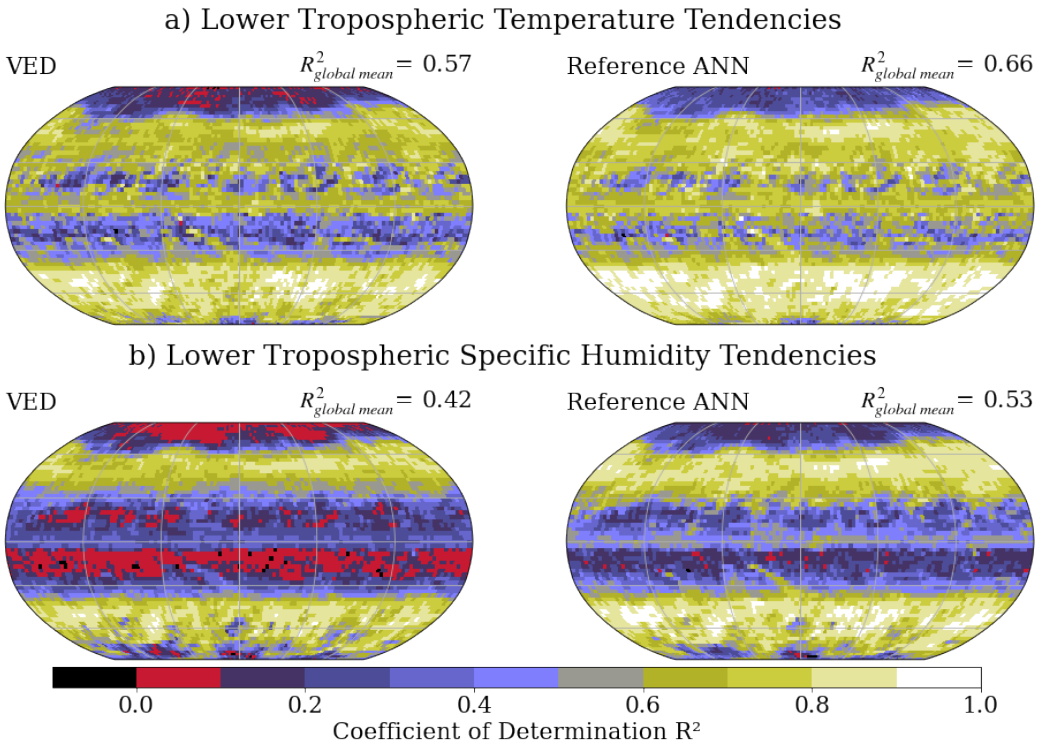

Figure 3.3.: Coefficient of Determination (R^2) of lower tropospheric temperature tendencies (a) and lower tropospheric specific humidity tendencies (b) at 700 hPa for the VED (left) and reference ANN (right column). The global mean $\mathbf{R^2}$ of each field is indicated in the upper right above every subplot. This plot was directly reproduced from **Behrens** et al. 2022

$$\mathbf{Var} = \frac{1}{\mathbf{P}} \sum_{t=1}^{\mathbf{P}} (\mathbf{Y_t} - \frac{1}{\mathbf{P}} \sum_{t=1}^{\mathbf{P}} \mathbf{Y_t})^2 \tag{3.7}$$

It is defined as the difference of 1 and the ratio between the MSE (equation 3.6) and the true variance (equation 3.7) of the data, where **P** is the length of the time series, **t** is the respective time step and $\mathbf{Y}$ / $\mathbf{Y}^{emul}$ are the true value of the test data / VED prediction. We constructed at first the time series of all output variables **O** from the test data set or predictions and computed the respective coefficients of determination in each grid cell (64 points in latitude × 128 points in longitude = 8192) of all layers. We selected the global subgrid heating and moistening fields at 700 hPa for the evaluation of the VED's $\mathbf{R^2}$ (Figure 3.3).

We choose **dq/dt** and **dT/dt** fields at this pressure level because of the limited skill in fitting lower tropospheric convective processes with neural nets that has been reported across multiple investigations, and which has been speculated to be associated with an underrepresentation of stochastic variability linked to shallow and deep convection (Gentine et al. 2018; Mooers et al. 2021; Rasp et al. 2018; Wang et al. 2022b). Both networks, VED and reference ANN, exhibit similar emulation skill patterns for heating and moistening tendencies, including the skill deficits for low-level moistening tendencies in the tropics, as seen in previous studies. Overall, we see a decreased reproduced variability with the VED ($\mathbf{R^2_{global\,mean}}$ = 0.57 / 0.42 for **dT/dt** / **dq/dt**; 35% and 22% of horizontal grid cells for temperature and specific humidity

tendencies with $\mathbf{R}^2 > 0.7$, respectively) compared to the reference ANN ($\mathbf{R}^2_{\text{global mean}} = 0.66$, 0.53 for **dT/dt**, **dq/dt**; 51% and 35% of horizontal grid cells with $\mathbf{R}^2 > 0.7$ for temperature and specific humidity tendencies, respectively). The VED shows regions of high reproduction skill for both, temperature and specific humidity tendencies along the mid-latitude storm tracks ($\sim 45°$ N / S, $\mathbf{R}^2 \sim 0.7$) and in the ITCZ region near the equator (ascending branch of Hadley Cell associated with deep convection, $\mathbf{R}^2 \sim 0.6$). Both networks exhibit weaker prediction skill of specific humidity and temperature tendencies near the descending branches of the Hadley Cell (subtropical highs $\sim 20°$ N / S) associated with an underestimation of (shallow) convective variability. Mooers et al. 2021 also found comparably weaker reproduction skill of their neural net in this region. Recently Wang et al. 2022a showed that the reproduction of moistening tendencies in the subtropics can be improved by using non-local features from adjacent grid cells as additional inputs of the neural net. Nevertheless, the VED shows good reproduction skill associated with convective processes in the lower troposphere compared to the reference ANN, despite its strongly reduced dimensionality in the latent space. This suggests that the information from large-scale climate variables $\mathbf{X}$ that is relevant for the prediction of subgrid-scale convective processes $\mathbf{Y}$ is closer to 5 (our latent space's dimensionality) than 64 (the input vector length). In other words, this means that the number of large-scale variables needed to skillfully emulate subgrid-scale processes is far smaller than the number of original input variables of the superparameterization. This is consistent with assumptions made by reduced-complexity models, such as the lower-dimensional multi-cloud model (Frenkel et al. 2012) or the quasi-equilibrium tropical circulation model (Neelin and Zeng 2000).

3.2.2. Tropical Variability

Current ESMs exhibit large biases in tropical precipitation and associated patterns (Bock et al. 2020). These regional uncertainties can be attributed to the fact that many ESMs struggle to reproduce tropical intra-seasonal variability like the Madden Julian Oscillation (MJO), an eastward propagating pattern of clustered deep convection in the Indo-Pacific Region (Zhang 2005). SPCAM yields a more realistic reproduction of the MJO compared to the traditional convective parametrization of CAM (Khairoutdinov et al. 2005). Furthermore, the governing tropical variability is largely reproducible with deep learning approaches (Rasp et al. 2018). Here, we investigate the ability of the VED to not distort the high-frequency tropical variability (15° N to 15° S) as simulated by SPCAM compared to the reference ANN. For this analysis, we use the entire second year of the SPCAM simulation to identify driving tropical variability with frequency lower than $\frac{1}{30}$ days^{-1}. This second SP year includes the 3-month sequence of the validation data set but has no overlap with the training data set.

Figure 3.4 shows the Wheeler-Kiladis diagrams, diagnosing the equatorial symmetric component (zonal wave numbers $\mathbf{k}$) of outgoing longwave radiation ($\mathbf{Q}_{lw\ top}$) with respect to its frequency ω for both SPCAM (Figure 3.4a) and VED (Figure 3.4b). I added arrows to assist the readers to navigate through Figure 3.4, which were not included in the original version presented in **Behrens** et al. 2022. Eastward propagating, non-dispersive Kelvin waves (ω^{-1}

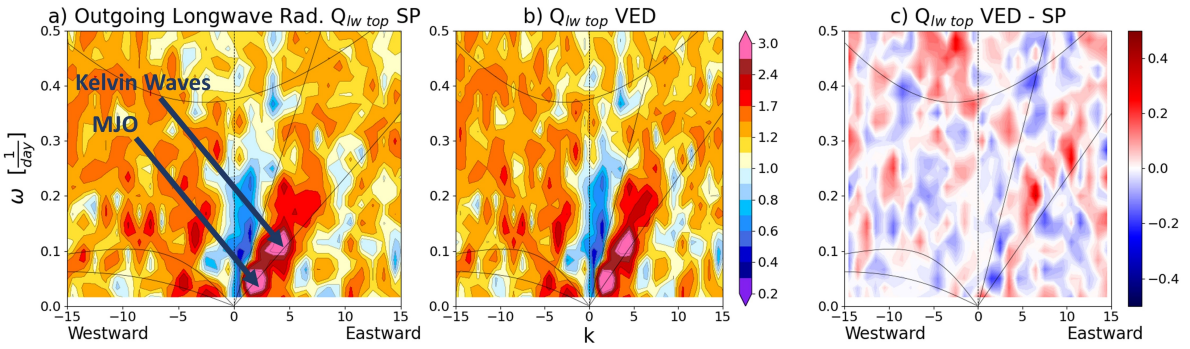

Figure 3.4.: Wheeler Kiladis diagram based on tropical outgoing longwave radiation [15° N-15° S] of SP (a), of VED predictions (b) and the absolute difference of spatio-temporal wave spectra VED - SP (c) for 1 year of SP simulations. The Figure is reproduced from **Behrens** et al. 2022 with small modifications (added arrows) to aid the reader with the interpretation.

~ 8 - 20 days, **k** ~ 2-5) and the MJO (ω^{-1} ~ 30 days, **k**=1) are not distorted by the VED. The resulting differences in the reproduced spatio-temporal variability with respect to SPCAM are generally confined within -0.2 to 0.2 (unit-less values) (Figure 3.4c), which amounts to a relative error of roughly 20%, and are not associated with a damping or absence of general features in ω-**k** space.

Although the reference ANN shows slightly better reproduction skill (see Figure A.3 in Appendix A), the VED and also ED (see Figure A.2 in Appendix A) can realistically reproduce not only mean regimes and characteristics of convective processes but also the associated variability even with its strongly reduced dimensionality on only 5 latent nodes.

Next, we evaluate our main interest – the physical interpretability of the VED with respect to convective processes – by exploring the information encapsulated in its latent space. We will show in the following sections that the representation of general convective processes is actually much lower dimensional than potentially envisioned.

3.2.3. Interpretability via Latent Space Exploration

In this section, we investigate convective processes and large-scale climate states captured in the latent space of the VED. This will give us a first impression of general drivers of convective predictability encapsulated in the latent manifold and will show the potential to study convective processes with only five latent nodes. Latent spaces of VAEs behave to some extent as a non-linear equivalent of a Principal Component Analysis (PCA), e.g., Rolinek et al. 2019, due to a skilful lower-dimensional encoding of information fed into the network. Therefore, we test whether the latent space of the VED retains a meaningful lower dimensional representation of convective processes like we would expect from a traditional PCA.

Human visualization of the full five latent dimensions (5 nodes, 5D) in a 2D schematic requires some additional dimensionality reduction. For visualization purposes, we therefore use a PCA to first compress the 5D manifold into a 2D lower-dimensional embedded space, which allows a visual inspection of the encapsulated information. The resulting 2D PCA

representation contains 82% of the total variance of the VED's latent space. Figure 3.5 shows the first (x-axis) and second leading Principal Component (PC) (y-axis) of the compressed latent space for 1 million randomly sampled points. The manifold, which is spanned by the two leading PCs, is then divided into a regular grid of size 50 (PC 1) × 50 (PC 2) cells. Tracking each selected sample allows us to characterize the embedded information for both convection and large-scale climate states. This permits us to compute conditional averages of these convection related variables in each grid cell of the 2D PCA compressed manifold.

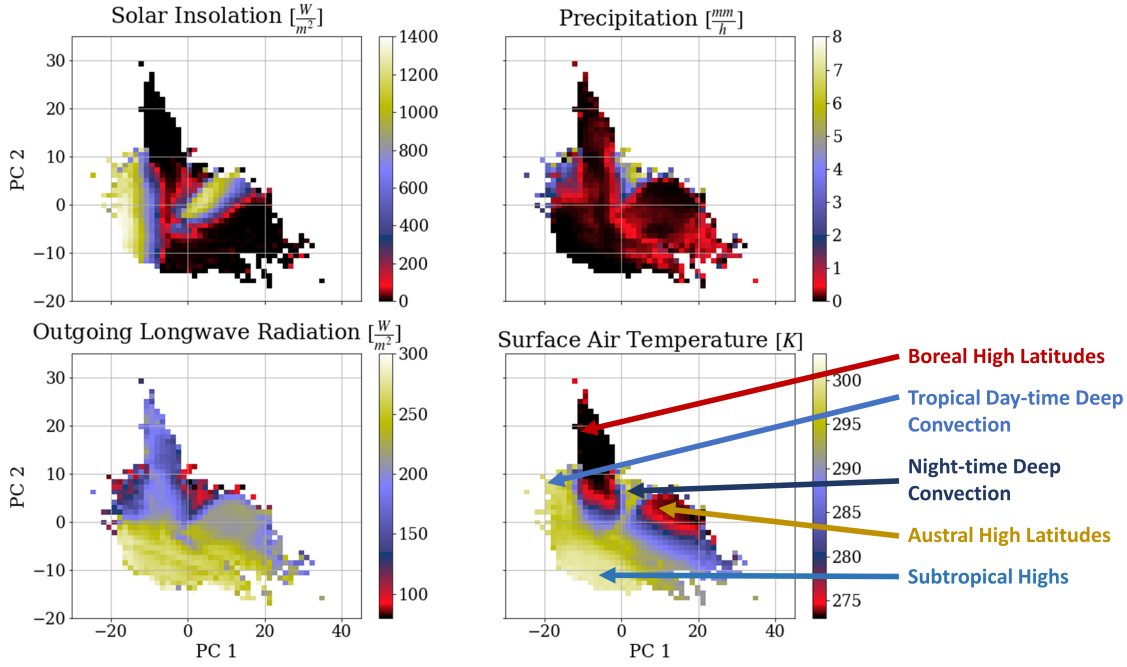

Figure 3.5.: The 2D Principal Component Analysis (PCA)-compressed latent space of the VED and associated conditional averages of solar insolation (upper left), precipitation (upper right), outgoing longwave radiation (lower left) and surface air temperature (lower right panel) of projected SP test data (see color scheme in each subplot). The x-axis / y-axis in all subplots indicates the 1^{st} / 2^{nd} leading Principal Component (PC) of the 5D latent space, which have a combined "explained variance" of around 0.82. The arrows in the lower right subplot indicate the position of characteristic samples from different geographic regions inside the 2D PCA-compressed latent space of the VED mentioned in the text. The Figure is directly reproduced from **Behrens** et al. 2022.

Figure 3.5 depicts the conditional averages of solar insolation ($\mathbf{Q}_{sol}$), precipitation (**precip**), outgoing longwave radiation ($\mathbf{Q}_{lw\ top}$), and surface air temperature ($\mathbf{T}_{surf}$) in the 2D PCA compressed latent space of the VED. Together the results show that distinct convective regimes are clearly separated in the latent space. More information on how the complex global superposition of distinct geographic convective regimes and large-scale processes in the latent space is contributed by separate latitudinal bands of the aquaplanet (tropics, boreal and austral mid latitudes) is provided in the Appendix A. Therein Figure A.4 shows the fixed Sea Surface Temperature (SST) field of the simulation and Figure A.6 the regional decomposition of patterns in the VED's latent space. These two figures can aid as a reference guide for the following latent space exploration. We start the analysis by investigating the impact of the insolation $\mathbf{Q}_{sol}$ on the latent space position, including whether the expected diurnal

cycle of convective processes in SPCAM (Khairoutdinov et al. 2005; Pritchard and Somerville 2009) is manifested in the latent space of the VED. Indeed, solar insolation $\mathbf{Q}_{sol}$ reveals 2 distinct maxima with day-time conditions and 2 minima with night-time conditions, which are separated by diurnal transition zones, as expected from diurnally varying input and output data of SP. Cross-evaluating the conditional averages of solar insolation with $\mathbf{T}_{surf}$, one can diagnose that the 2D PCA compressed latent space of VED stores information that can be used to infer the geographic location of a sample. As an example, we can focus on the 'fin-shaped' region (PC1 ~ -8, PC2 ~ 15) protruding from the top of the 2D PCA compressed latent space. Here the samples are characterized by anomalously cold (273 K < $\mathbf{T}_{surf}$ < 278 K) climate conditions without solar insolation. Based on the fixed SST forcing (see Figure A.4 in the Appendix A), the low surface air temperatures and the constant perpetual Austral Summer solar forcing, we can conclude that these samples originate from polar and subpolar latitudes in the Northern Hemisphere. Furthermore, we find a zone with day-time solar insolation ($\mathbf{Q}_{sol}$ > 700 $\frac{\mathrm{W}}{\mathrm{m}^2}$) and cold surface air temperature (273 K < $\mathbf{T}_{surf}$ < 280 K) in the upper-right part of the latent space (PC1 ~ 10, PC2 ~ 5), which represents large-scale climate conditions that can be only found in the austral polar and subpolar latitudes in SPCAM test data.

We also explore the fingerprinting of precipitation on the latent space as a proxy for the strength of convective processes, since it is closely connected to convective moistening and convective heating (Emanuel 1994; Lohmann et al. 2016). The 2D PCA compressed latent space of the VED reveals a good separation of samples with no or negligible precipitation, shallow convection with the formation of weak precipitation, and deep convective samples with intense precipitation (**precip** > 10 $\frac{mm}{h}$ in the tropics, see Figure A.6 in the Appendix A). We expect to see a clear separation between tropical deep convective samples and samples with no or negligible precipitation from the colder higher latitudes or the region of the subtropical highs in the 2D PCA compressed latent space due to the strong variation in the magnitude of convective processes with latitude as it is visible in Figure 3.5. If we now focus on the conditionally-averaged plot of precipitation, two maxima are evident. The first precipitation maximum (PC1 ~ -15, PC2 ~ 5) is associated with day-time solar forcing, a minimum of outgoing long-wave radiation ($\mathbf{Q}_{lw\ top}$ < 150 $\frac{W}{m^2}$, which suggests high cloud tops in the upper half of the troposphere) and tropical surface air temperatures ($\mathbf{T}_{surf}$ > 295 K). Therefore, this maximum originates from tropical day-time deep-convective samples in SPCAM. The second maximum (PC1 ~ 5, PC2 ~ 5) exhibits slightly colder surface air temperatures, night-time conditions, decreased outgoing longwave radiation ($\mathbf{Q}_{lw\ top}$ ~ 100 $\frac{W}{m^2}$) and precipitation formation of more than 3 $\frac{mm}{h}$. It can be shown that this maximum originates from night-time deep convection from the tropics in its center and predominantly strong precipitating samples from the Northern and Southern extratropics along the left and right boundary, respectively.

Outgoing longwave radiation ($\mathbf{Q}_{lw\ top}$) is a good estimator for both the height of cloud tops based on the inferred brightness temperatures for convective samples or surface temperatures for non - or negligibly - convective samples. Based on the combination of high $\mathbf{Q}_{lw\ top}$ (no or negligible convection), no precipitation formation and anomalous warm surface temperatures ($\mathbf{T}_{surf}$ ~ 300 K), one can conclude that samples from subtropical highs (the descending branch

of the Hadley cell, with limited deep-convective processes with large vertical extent in the free troposphere) are concentrated in the lower left part of the PCA compressed latent space (PC1 ~ -10, PC2 ~ -10) of the VED.

These results demonstrate how large-scale climate conditions and convective processes are connected and physically interpretable in the latent space (e.g., equivalence of **precip** maxima and $\mathbf{Q}_{lw\ top}$ minima), which illustrates the encoding power of the VED. Furthermore, the evaluated mean statistics support that the VED realistically reproduces convective processes and the associated variability despite a strong dimensionality reduction down to only five nodes in the latent space, which shows the decoding power of the network.

Similar reproduction abilities can be investigated for ED, but the physical interpretability of the resulting latent space is reduced compared to VED. The KL divergence used for the VED ensures an improved separation of latent modes. The effect can be seen in a larger number of centers of action in the ED's latent space and weaker gradients in the conditional average plots with respect to subgrid-scale and climate variables, as can be seen in the Appendix A Figure A.5 (ED vs VED latent spaces) and Figure A.6 for the VED conditional average plot or Figure A.7 for ED conditional average plot. Additionally, we tested the interpretability of the 2D PCA compressed latent space of a VED trained on **X** to emulate **Y**, in other words mirroring the input data and output data of SP (see subsection A.4.1 in the Appendix A). In this case the latent space strongly focuses on the magnitude of heating or moistening tendencies, resembling a weak gradient from negligible convective processes towards strongly precipitating deep convective samples (see Figure A.16 in Appendix A). For large-scale climate variables like surface air temperature, the 2D PCA compressed latent space of a $\text{VED}_{X\rightarrow Y}$ mostly distinguishes between warm conditions and cold conditions sorting samples from both poles close together in one minimum (see Figure A.17 in Appendix A), which makes the visual separation of austral and boreal polar latitudes nearly impossible. In contrast, VED shows a pronounced separation of austral and boreal polar samples and reveals distinct regimes of convective processes in its 2D PCA compressed latent space as seen in Figure 3.5, which is a clear advantage in interpretability of this network compared to $\text{VED}_{X\rightarrow Y}$.

We further compared the interpretability of the 2D PCA compressed latent space of the VED against a traditional PCA on the large-scale input features **X**, as an unsupervised linear reference method. The first two leading PC's with respect to **X** show overall weak gradients in its lower-dimensional space for the conditional averages of solar insolation, outgoing longwave radiation and surface air temperature (Figure A.8 in Appendix A). The 'centers of action' are less pronounced for the PCA on **X** compared to its equivalent on the latent space of VED seen in Figure 3.5. Especially the identification of deep convective samples is hardly possible inside the submanifold spanned by the two leading PC's of the large-scale variables as can be seen in Figure A.8 in Appendix A. In latitude - longitude plots (Figure A.9 in Appendix A) these two leading PC's resemble large-scale patterns with meridional gradients that show similarities with temperature or radiation fields but barely with subgrid-scale variables. In contrast, the latent space of VED focuses on both large-scale and subgrid-scale patterns. The first two latent variables are characterised by large-scale patterns connected to geographic variability

and solar insolation (see Figure A.10 in Appendix A). The remaining three latent variables describe mostly subgrid-scale convective processes, as can be seen in Figure A.10 in Appendix A.

The concept of the computing conditional averages can be repeated also on 2D projections spanned by the original latent variables of the VED without a PCA as postprocessing step. An example of this more detailed latent space inspection can be found in the Appendix A in section A.3 (Figure A.12 for precipitation, A.13 for solar insolation and A.14 for surface air temperature).

As a next step, we will combine the reproduction skill and the encapsulated information in the latent space of the VED to investigate convective processes by identifying distinct large-scale drivers, associated convective regimes and geographic variability in detail.

3.3. Unveiling Drivers of Convective Processes in SPCAM Using Generative Modeling

This section is directly reproduced from **Behrens** et al. 2022 and based on the generative modeling section of this paper.

In this section, we discuss the dominant drivers of convective processes encapsulated in the latent space of the VED using a generative modeling approach. We compute the marginal distributions of all 5 latent variables $\mathbf{z}$. We focus on the 10^{th}, 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles of the marginal distributions of the latent variables. Since most of these distributions are bi-modal (see Figures 3.6 – 3.10a), we select their median values as estimators for the intersect (origin) of the 5-dimensional $\mathbf{z}$, instead of the mean. For all latent variables, the median is close to the mode value (peak value) of the marginal distributions. To generate the 'median' climate conditions and associated convective processes from the 'median' values of the latent variables, we construct a reference state $\mathbf{z}_{median}$ (Equation 3.8). $\mathbf{z}_{median}$ contains the median values for all five latent variables. This reference state is fed into the decoder of the VED to generate vertical heating, moistening, specific humidity, and temperature profiles. These vertical profiles represent the 'median' state of convective processes and associated climate conditions.

$$\mathbf{z_{median}} = [\text{median}\,(z_1)\,, \text{median}\,(z_2)\,, \text{median}\,(z_3)\,, \text{median}\,(z_4)\,, \text{median}\,(z_5)] \tag{3.8}$$

To investigate encapsulated convective regimes and large-scale climate states in the latent space of VED via generative modeling, we replace the median value with the different percentiles ($\text{perc}\,(z_1)$ in Equation 3.9) along one specific marginal distribution. This analysis identifies how each latent node drives a variation of convective processes and large-scale climate states generated by the decoder and manifests in well-known convective regimes. The modified $\mathbf{z}_{translation}$ (Equation 3.9) can be seen as a latent forcing on the decoder, acting as a knob which amplifies or damps the associated convective features. Furthermore, $\mathbf{z}_{translation}$ influences the geographic variability of generated samples, allowing an interpolation from a

tropical to a polar 'background' climate state like a knob for the general volume of generated large-scale profiles. A clear separation between geographic versus convective modulation with a distinct $\mathbf{z}_{translation}$ is challenging and not the primary goal of our VED's decoder setup. The evaluation whether a distinct latent node drives more geographic than convective modulation necessarily involves an analysis of all generated variables - an interesting analysis trade-off revealed by this latent space exploration. $\mathbf{z}_{translation}$ can be geometrically interpreted as a translation along one distinct latent dimension in the 5-dimensional latent space. For instance, $\mathbf{z}_{translation}$ is applied as an example to latent node 1 perturbing the 'median' conditions along this latent dimension, while keeping the median values for the 4 other dimensions:

$$\mathbf{z}_{\textbf{translation node 1}} = \left[\text{perc}\left(z_1\right), \text{median}\left(z_2\right), \text{median}\left(z_3\right), \text{median}\left(z_4\right), \text{median}\left(z_5\right)\right] \tag{3.9}$$

Applying a translation along one latent dimension while keeping the other latent variables fixed to their median values implicitly assumes that latent variables do not overly depend on each other. To test this independence, we calculate the Pearson correlation between all five latent variables using the entire test data set. The mean correlation coefficients between the latent dimensions are confined within ±0.35, except for a mean correlation of -0.74 between latent variables 2 and 5. The relatively large linear connection between latent variables 2 and 5 can be further explored by density plots using the 2D projection spanned by these latent variables, see Figure A.11 in Appendix A. While Latent Node 2 separates moist and warm from cold and dry tropospheric conditions, Latent Node 5 represents deep convection samples, which rely on anomalous wet and warm conditions in the troposphere. Therefore it is not surprising to see a pronounced anti-correlation between these nodes. This is a further evidence of the interpretability and meaningfulness of the VED's latent space, i.e., learning physical processes in the lower-order manifold.

In the following we will use $\mathbf{z}_{translation}$ along all five latent dimensions to identify large-scale drivers of convective processes and different convective regimes in SPCAM. We use the notation 'high $\mathbf{z}_{translation}$' to describe the cases when $\mathbf{z}_{translation} > \mathbf{z}_{median}$ and 'low $\mathbf{z}_{translation}$' if $\mathbf{z}_{translation} < \mathbf{z}_{median}$. Figures 3.6 – 3.10 illustrate the marginal distribution along the respective latent dimensions (Panels a, where the dashed black line indicates the median value of each dimension, Equation 3.8). The other subplots of these figures show the generated vertical moistening, heating, specific humidity and temperature profiles (Panels b-e) of the decoder with respect to $\mathbf{z}_{median}$ (Equation 3.8) or $\mathbf{z}_{translation}$ (Equation 3.9 along a distinct latent dimension). Additionally, two subgrid-scale and climate variables (Panels f), which are strongly affected by the applied latent forcing, are displayed as a function of $\mathbf{z}_{translation}$ for illustrative purposes. The marker-edge-color in the respective Panels f reveal the chosen percentiles. All other generated subgrid-scale and large-scale climate variables are shown in Tables A.7-A.11 in the Appendix A. We investigate in the following that latent node 1 and latent node 2 focus on the large-scale climate (geographic) variability in $\mathbf{X}$ rather than on subgrid-scale convective processes in $\mathbf{Y}$. In contrast, latent nodes 3, 4 and 5 exhibit main characteristics of dominant convective regimes captured in $\mathbf{Y}$.

3.3.1. Large-Scale Climate Variability Nodes

In this first part we demonstrate that latent nodes 1 and 2 capture mostly large-scale climate variability in **X**.

Latent Node 1: Global Temperature Variations

Global temperatures in the troposphere are dominated by the large meridional gradients from equatorial to polar latitudes mainly related to solar insolation differences between the tropics and extratropics.

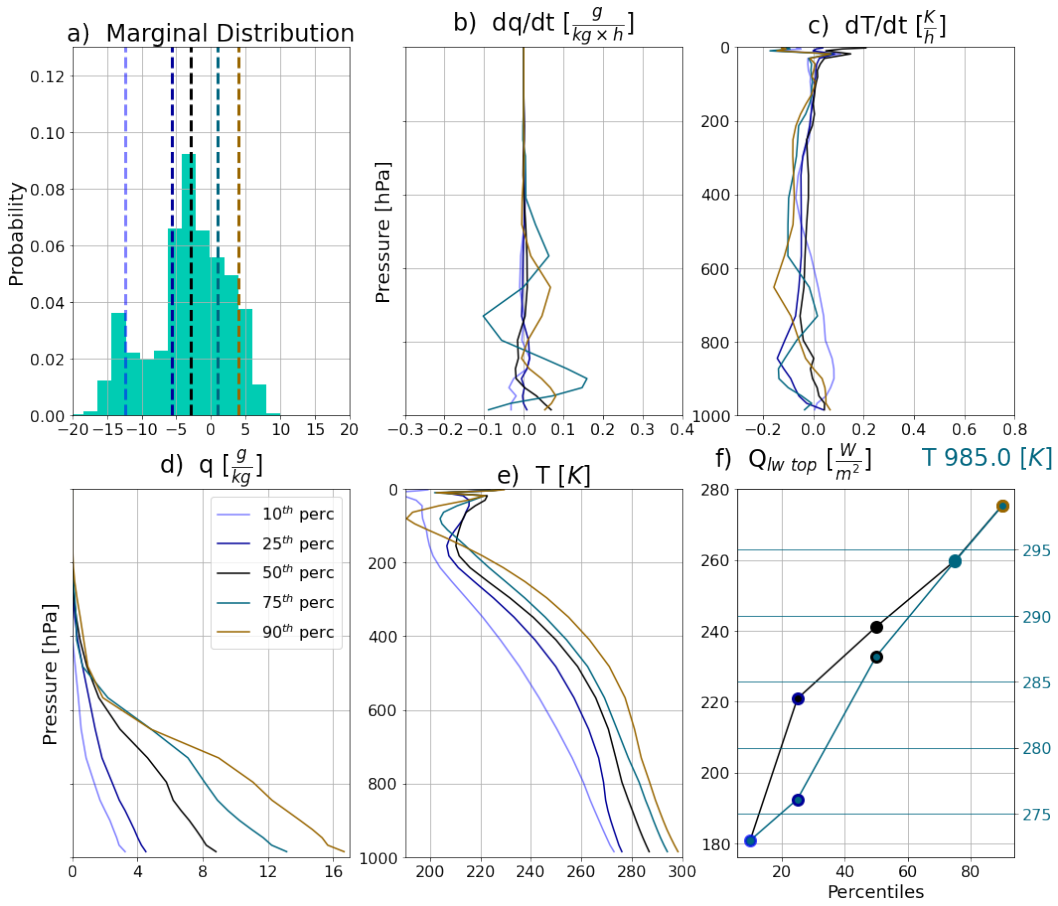

Figure 3.6.: Marginal distribution of latent node 1 (a) and the resulting generated vertical profiles of specific humidity tendencies **dq/dt** (b), temperature tendencies **dT/dt** (c), specific humidity **q** (d) and temperatures **T** (e). The dashed lines in the marginal distribution plot represent the chosen percentiles (see legend in subplot d) and the resulting effect of the respective translation $\mathbf{z}_{translation}$ on the profiles is shown in the subplots. Furthermore, the longwave heat flux at the model top ($\mathbf{Q}_{lw\ top}$) and the surface air temperature ($\mathbf{T}_{surf}$ / **T 985.0**) (f) are illustrated as function of the translation $\mathbf{z}_{translation}$ along the latent dimension 1. The marker-edge-color in panel f symbolise the respective percentiles of $\mathbf{z}_{translation}$. The black lines in subplots b-e indicate the generated reference state with $\mathbf{z}_{median}$. This Figure is directly reproduced from **Behrens** et al. 2022.

The first latent node (Node 1) captures these global meridional temperature variations (Figure 3.6e), as suggested by the large spread of the surface temperature response to $\mathbf{z}_{translation}$,

encompassing the tropics ($\mathbf{T}_{surf} \sim 298$ K, high $\mathbf{z}_{translation}$) and polar regions ($\mathbf{T}_{surf} \sim 273$ K, low $\mathbf{z}_{translation}$). Tropical regions are characterized by very moist conditions in the boundary layer ($\mathbf{q} > 10 \frac{g}{kg}$), while being extremely dry at the poles ($\mathbf{q} \sim 1.5 - 3.5 \frac{g}{kg}$), see Figure 3.6d. The strong connection between tropospheric temperatures or specific humidity and Node 1 can be shown with a linear correlation of globally concatenated temperature space-time series (of horizontal grid cells and time, featuring the large meridional gradients) and respective node space-time series. The resulting "linear explained variance" of temperature space-time series on Node 1 exceeds 0.5 (Figure A.18 in Appendix A), while the "linear explained variance" vanishes if the analysis is repeated on the time series for each horizontal grid cell (Figure A.19 in Appendix A, without the large meridional gradients). A detailed description how these two correlations metrics were computed can be found in the Appendix A in section A.5.

A physical interpretation of this response on the $\mathbf{z}_{translation}$ can be given based on the Clausius-Clapeyron relationship. A warmer atmosphere results in a near-exponentially higher saturation water vapor pressure, which in turn allows higher specific humidity content. Therefore, we see strongly coupled variations of temperature and specific humidity between the equator and the poles. In short, the first latent node represents these overarching large-scale meridional variations in tropospheric temperatures, influencing specific humidity, but is not necessarily linked to convective processes **Y**, but rather to large-scale conditions **X**, which are also part of the VED reconstruction.

Latent Node 2: Large-Scale Variability along the Mid-Latitude Storm Tracks

Latent node 2 characterizes more the large-scale climate (and thus geographic) variability in **X** than focuses on a distinct convective regime. Latent dimension 2 (Node 2, Figure 3.7) clearly captures temperature and specific humidity variations in the troposphere, as can be seen in Figure 3.7d and 3.7e. Warmer and moister tropospheric conditions are associated with high $\mathbf{z}_{translation}$.

Low $\mathbf{z}_{translation}$ characterizes cold and stable conditions during day-time ($\mathbf{Q}_{sw\ top} \sim 1000 \frac{W}{m^2}$). These anomalous cold and dry conditions in the upper troposphere are associated with negligible convective processes, as diagnosed with a large outgoing longwave heat flux at the model top ($\mathbf{Q}_{lw\ top} \sim 240 \frac{W}{m^2}$) and the formation of no precipitation (Table A.8 in Appendix A). Due to the large shortwave heat flux at the model top, the perpetual austral summer solar forcing and the low surface air temperatures ($\mathbf{T}_{surf} \sim 281$ K), low $\mathbf{z}_{translation}$ can be traced back to the austral mid-latitudes. Whereas high $\mathbf{z}_{translation}$ is linked to night-time conditions ($\mathbf{Q}_{sw\ top} < 200 \frac{W}{m^2}$) with a warm, moist troposphere ($\mathbf{T}_{surf} \sim 291$ K). High $\mathbf{z}_{translation}$ is further characterized by mid-level convection ($\mathbf{Q}_{lw\ top} \sim 180$ - $200 \frac{W}{m^2}$) with intermediate precipitation formation ($\mathbf{precip} \sim 0.12$ to $0.15 \frac{mm}{h}$, Table A.8 in Appendix A) associated with a warmer and moister upper troposphere and can be found in the subtropics on both hemispheres.

Our approach allows us to identify the main patterns of the large-scale climate state in **X**, which are main drivers of the general circulation and convection, besides convective processes in **Y** in the latent space. These convective processes are heavily modulated by **X**. Node 2

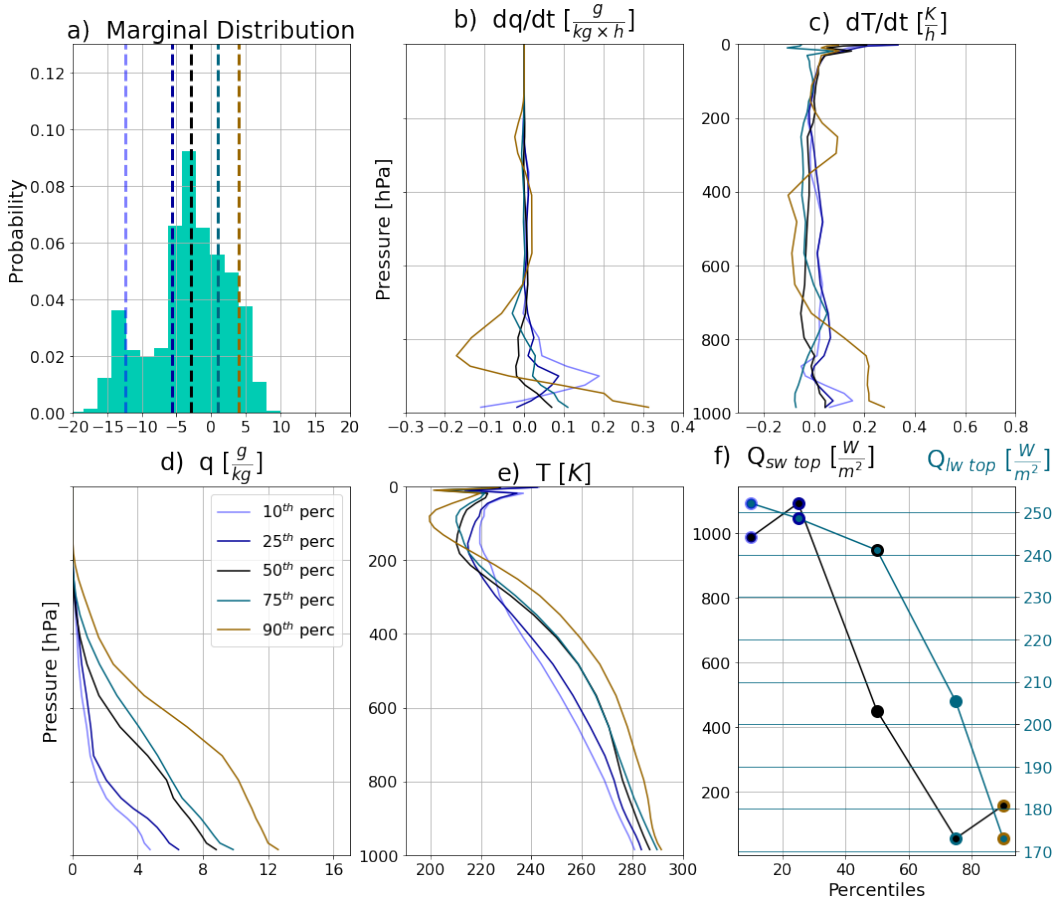

Figure 3.7.: Marginal distribution of latent node 2 (a) and the resulting generated vertical profiles of specific humidity tendencies **dq/dt** (b), temperature tendencies **dT/dt** (c), specific humidity **q** (d) and temperatures **T** (e). The dashed lines in the marginal distribution plot represent the chosen percentiles (see legend in subplot d) and the resulting effect of the respective translation $\mathbf{z}_{translation}$ on the profiles is shown in the subplots. Furthermore, the shortwave heat flux at the model top ($\mathbf{Q}_{sw\ top}$) and the outgoing longwave heat flux ($\mathbf{Q}_{lw\ top}$) (f) are illustrated as function of the translation $\mathbf{z}_{translation}$ along the latent dimension 2. The marker-edge-color in panel f symbolise the respective percentiles of $\mathbf{z}_{translation}$. The black lines in subplots b-e indicate the generated reference state with $\mathbf{z}_{median}$. This Figure is directly reproduced from **Behrens** et al. 2022.

captures characterising features of the large-scale meridional variability of specific humidity and temperatures between the mid latitudes and the subtropics (e.g., an essential driver of mid-latitude storm track dynamics; Bony et al. 2015). Latent dimension 2 is further influenced by the solar forcing. The clear separation between austral mid latitude temperature profiles on one side and samples from subtropical regions on the other side of the $\mathbf{z}_{translation}$ are further evidence that latent node 2 encapsulates a part of the geographic variability inside the latent space seen in Figure 3.5.

3.3.2. Convective Regime Nodes

Next, we will show that latent node 3, 4, 5 usefully characterize mostly distinct convective regimes in the subgrid-scale process rate variables **Y**.

Latent Node 3: Shallow Convection

Shallow convective processes are one of the dominant cloud regimes investigated in observational studies (e.g., Huaman and Schumacher 2018). Latent node 3 characterizes some of the main characteristics of shallow convective processes as revealed by its vertical profiles of specific humidity and temperature tendencies influenced by large-scale specific humidity and surface diabatic fluxes.

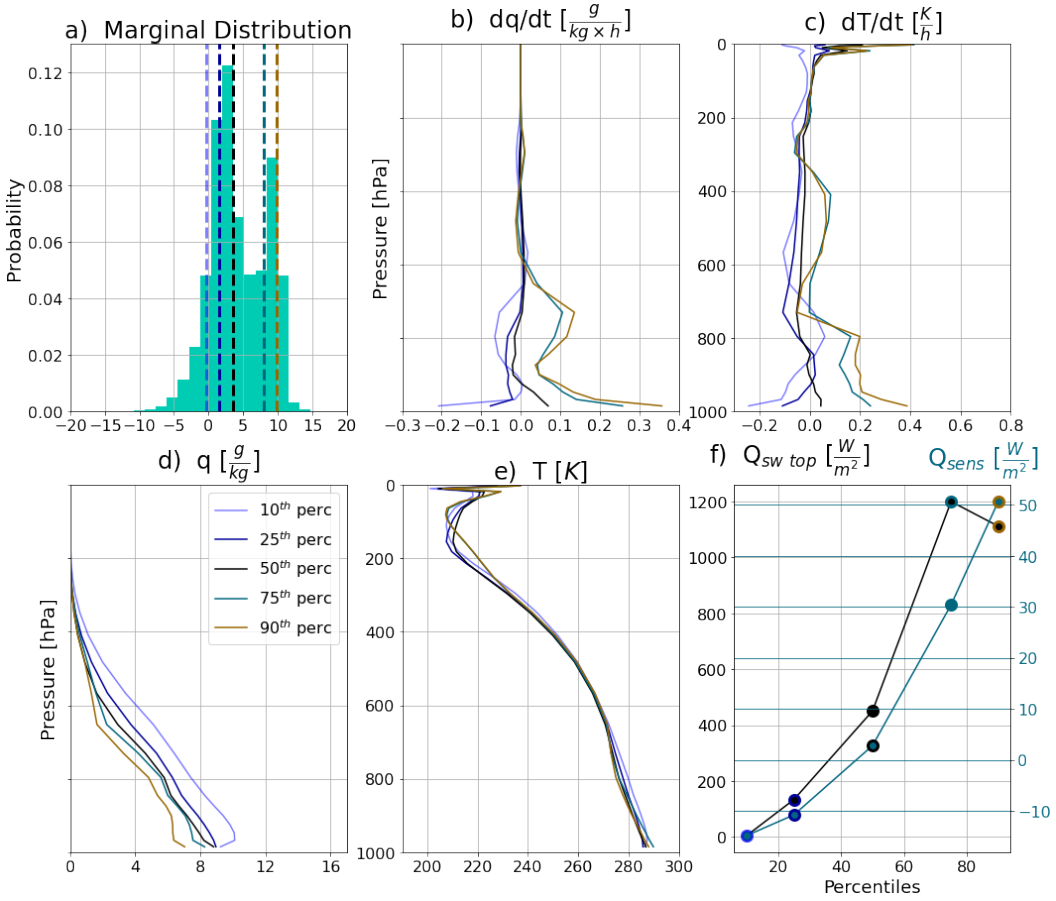

Figure 3.8.: Marginal distribution of latent node 3 (a) and the resulting generated vertical profiles of specific humidity tendencies **dq/dt** (b), temperature tendencies **dT/dt** (c), specific humidity **q** (d) and temperatures **T** (e). The dashed lines in the marginal distribution plot represent the chosen percentiles (see legend in subplot d) and the resulting effect of the respective translation $\mathbf{z}_{translation}$ on the profiles is shown in the subplots. Furthermore, the shortwave heat flux at the model top ($\mathbf{Q}_{sw\ top}$) and the surface sensible heat flux ($\mathbf{Q}_{sens}$) (f) are illustrated as function of the translation $\mathbf{z}_{translation}$ along the latent dimension 3. The marker-edge-color in panel f symbolise the respective percentiles of $\mathbf{z}_{translation}$. The black lines in subplots b-e indicate the generated reference state with $\mathbf{z}_{median}$. This Figure was directly reproduced from **Behrens** et al. 2022.

Figure 3.8 shows the marginal distribution of latent node 3 (Node 3), the generated vertical specific humidity and temperature tendencies, and the large-scale specific humidity and temperature profiles of the Decoder for $\mathbf{z}_{median}$, as well as the applied $\mathbf{z}_{translation}$. Furthermore, the generated shortwave heat flux ($\mathbf{Q}_{sw\ top}$) and surface sensible heat flux ($\mathbf{Q}_{sens}$) are displayed as a function of $\mathbf{z}_{translation}$. Along latent dimension 3, the specific humidity (**q**) decreases through-

out the entire troposphere for increasing $\mathbf{z}_{translation}$, while surface diabatic fluxes (sensible heat flux $\mathbf{Q}_{sens}$ and latent heat flux $\mathbf{Q}_{lat}$, Table A.9 in Appendix A) increase. Likewise, the outgoing longwave radiation $\mathbf{Q}_{lw\ top}$ increases with increasing $\mathbf{z}_{translation}$ suggesting higher cloud tops and stronger convective processes for low $\mathbf{z}_{translation}$ (Table A.9 in Appendix A). In contrast, the intensity of shallow convection and outgoing longwave radiation decreases when $\mathbf{z}_{translation}$ increases (high $\mathbf{z}_{translation}$). Specific humidity tendencies (**dq/dt**) in the lower troposphere (p $>$ 600 hPa) react to $\mathbf{z}_{translation}$ in a bimodal way. They moisten, in combination with a strong positive surface diabatic forcing, the relatively dry ambient air in the lower troposphere above the reference conditions (high $\mathbf{z}_{translation}$), whereas the opposite is true for low $\mathbf{z}_{translation}$. In this case, negative **dq/dt** in combination with negative diabatic forcing lead to a drying of moist conditions in the lower troposphere. Precipitation is insensitive to $\mathbf{z}_{translation}$ due to the small vertical extent of convective moistening, confined below 600 hPa; this latent node evidently avoids deep convective regimes. The generated temperature profiles of $\mathbf{z}_{translation}$ along latent dimension 3 are characteristic of the subtropics and mid-latitudes in the SP simulation. The **dT/dt** profiles show slight variations near the surface due to $\mathbf{z}_{translation}$, while being insensitive in the middle troposphere. The fixed SST field (Figure A.4 in Appendix A) or the conditional averages of surface air temperatures (Figure A.6 in Appendix A) in certain regions can be used to gain a first visual orientation of the geographic origin of a generated sample. This first impression is complemented with a detailed search for such conditions in the SP test data. Furthermore, night-time conditions with small shortwave radiative heat flux at the model top $\mathbf{Q}_{sw\ top}$ and day-time conditions with high values of $\mathbf{Q}_{sw\ top}$ ($\mathbf{Q}_{sw\ top} \sim 1000\ \frac{W}{m^2}$) can be distinguished for low $\mathbf{z}_{translation}$ and high $\mathbf{z}_{translation}$, respectively.

Interestingly, the generated profiles and variables suggest that latent node 3 is mostly sorting information about subgrid-scale processes **Y** within one sub-regime of **X**, rather than focusing on sorting the large-scale geographic variability in **X**. The strong response of **dq/dt** in the planetary boundary layer and adjacent layers, negligible precipitation formation and the characteristic temperature range between the subtropics and mid-latitudes, are key evidence that the latent node 3 encapsulates shallow convective processes. Shallow convection is influenced by the diurnal cycle, leading to a strengthening of shallow convective processes during the day and a weakening of these processes accompanied with a drying of the planetary boundary layer during the night, as it is supported by Figure 3.8.

Latent Node 4: Mid Latitude Frontal Systems

Mid-latitude frontal systems are characterized by a large variety of convective regimes associated with the warm or cold front of these systems (Bony et al. 2015). On latent node 4 we discover certain characteristic features in subgrid-scale profiles **X** and associated large-scale fields **Y**. These features allow us to draw links to distinctive convective regimes of mid-latitude cyclones based on their fingerprint in **X** and **Y**. Unlike the previous latent nodes, the response of the latent node 4 (Node 4, Figure 3.9) to the translation $\mathbf{z}_{translation}$ results in nearly constant

solar insolation ($\mathbf{Q}_{sw\ top} \sim 440$ - $450\ \frac{W}{m^2}$, see Table A.10 in Appendix A) and a narrow meridional band.

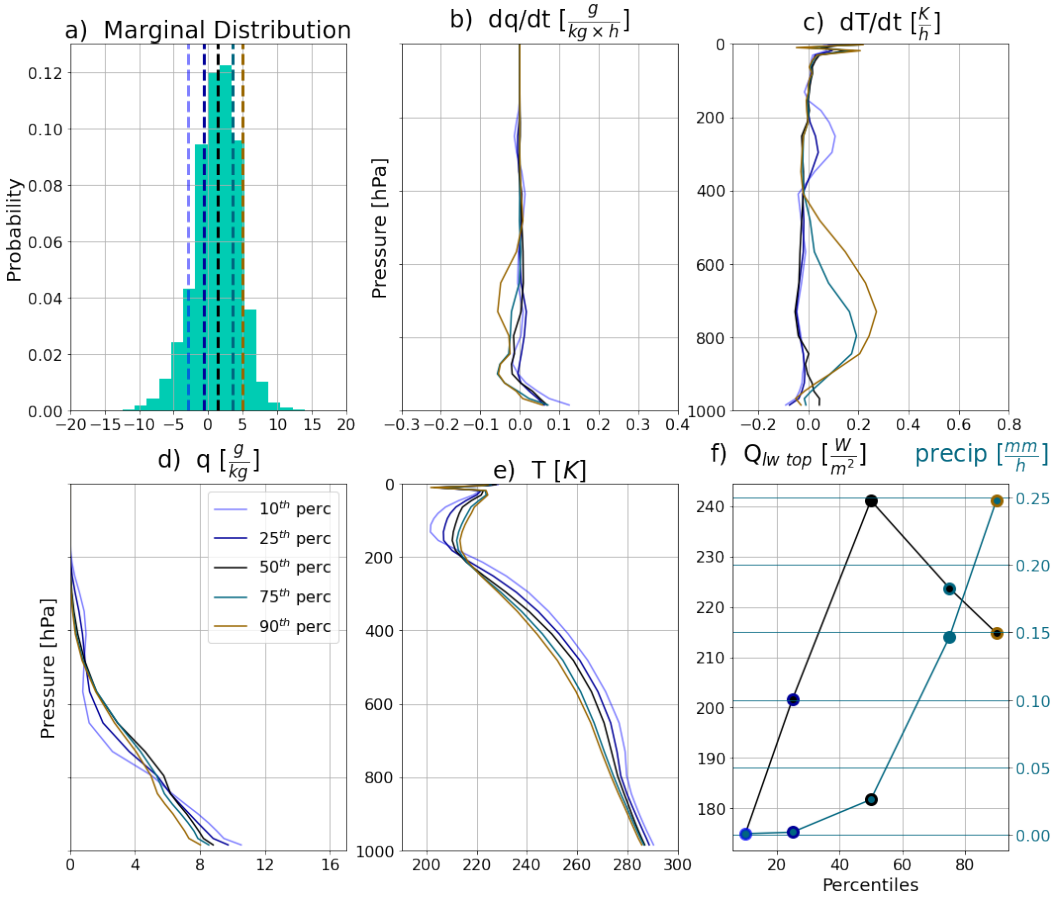

Figure 3.9.: Marginal distribution of latent node 4 (a) and the resulting generated vertical profiles of specific humidity tendencies **dq/dt** (b), temperature tendencies **dT/dt** (c), specific humidity **q** (d) and temperatures **T** (e). The dashed lines in the marginal distribution plot represent the chosen percentiles (see legend in subplot d) and the resulting effect of the respective translation $\mathbf{z}_{translation}$ on the profiles is shown in the subplots. Furthermore, the longwave heat flux at the model top ($\mathbf{Q}_{lw\ top}$) and the precipitation rate (**precip**) (f) are illustrated as function of the translation $\mathbf{z}_{translation}$ along the latent dimension 4. The marker-edge-color in panel f symbolise the respective percentiles of $\mathbf{z}_{translation}$. The black lines in subplots b-e indicate the generated reference state with $\mathbf{z}_{median}$. This Figure is directly reproduced from **Behrens** et al. 2022

The generated surface temperature ranges from 286 K to 290 K with varying $\mathbf{z}_{translation}$. This temperature range is common to mid-latitudes or the subtropics (e.g., see Figure A.6 in Appendix A) and can be found in the SPCAM simulations between 45° N / S and 25° N / S. Low $\mathbf{z}_{translation}$ corresponds to warmer and drier conditions in the free mid-troposphere between 800 hPa and 400 hPa, while moister conditions are found above and below. The anomalous moist conditions in the upper free troposphere are connected to a heating peak at 300 hPa (**dT/dt** $\sim 0.1\ \frac{K}{h}$, Figure 3.9c). Likewise, the difference between the shortwave heat flux at the model top and the surface is relatively small ($\mathbf{Q}_{sw\ top} - \mathbf{Q}_{sw\ surf} \sim 120 - 130\ \frac{W}{m^2}$, Table A.10 in Appendix A), which suggests optically thin clouds. Additionally, the outgoing long

wave radiation is small ($\mathbf{Q}_{lw\ top} < 200\ \frac{W}{m^2}$) and no precipitation is formed. These conditions are characteristic of high cirrus-like convection.

On the other side, high $\mathbf{z}_{translation}$ shows relatively strong heating tendencies in the free troposphere (**dT/dt** > 0.2 $\frac{K}{h}$, see Figure 3.9c) and drying conditions below 600 hPa down to the surface (**dq/dt** ~ -0.1 $\frac{g}{kg \times h}$, Figure 3.9b). These conditions, along with moderate precipitation (**precip** ~ 0.15 - 0.25 $\frac{mm}{h}$), higher outgoing longwave heat flux ($\mathbf{Q}_{lw\ top} > 200\ \frac{W}{m^2}$) and lower shortwave transmissivity ($\mathbf{Q}_{sw\ top} - \mathbf{Q}_{sw\ surf} \sim 170\ \frac{W}{m^2}$, Table A.10 in Appendix A) characterize mid-level cumulus convection.

Based on this evidence, we were able to show that latent node 4 focuses on subgrid-scale convective processes in **Y**. The generated large-scale conditions exhibited by Node 4 are well-suited for these cirrus-like or cumulus convection regimes. In detail, latent node 4 shows a clear transition from a cirrus type convective regime (low $\mathbf{z}_{translation}$) to a cumulus type precipitating convective regime (high $\mathbf{z}_{translation}$) in mid-latitudes. This response is associated with frontal systems, which consist of high cirrus clouds in the surroundings of the warm front and cumulus convection along the cold front (Bony et al. 2015).

Latent Node 5: Deep Convection

Deep convection is the cloud regime with the largest vertical extent. It is characterized by especially strong convective heating and drying throughout almost the entire troposphere, as can be seen in Frenkel et al. 2015 and accompanied by anomalous intense precipitation (see Figure 3.5). The first mode of latent node 5 reveals general characteristics of a deep convective regime captured in generated subgrid-scale variables **Y**. The response of latent dimension 5 (Node 5) to $\mathbf{z}_{translation}$ shows either strong deep convection (first mode in Figure 3.10a) or stable conditions (second mode in Figure 3.10a) in the troposphere. A surface temperature of 293 K for low $\mathbf{z}_{translation}$ indicates subtropical regions (e.g., the surface temperature in the tropics is at least 3 K warmer in this SPCAM simulation). The warmer and moister troposphere for low $\mathbf{z}_{translation}$ is accompanied with strong heating and drying tendencies peaking at around 500 hPa of 0.5 to 0.7 $\frac{K}{h}$ and -0.15 to -0.2 $\frac{g}{kg \times h}$ respectively. In this case, we observe intense precipitation formation up to 0.6 $\frac{mm}{h}$ and low outgoing longwave radiation ($\mathbf{Q}_{lw\ top} < 201\ \frac{W}{m^2}$). All these conditions are characteristics of subtropical deep convective events.

In contrast, high $\mathbf{z}_{translation}$ is associated with a mid latitude surface air temperature ($\mathbf{T}_{surf} \sim$ 5 K colder than in the subtropics). A night-time (Table A.11 in Appendix A), dryer troposphere with very small or negligible heating and moistening tendencies (manifestation of stable conditions) throughout the troposphere is accompanied with relatively large outgoing long wave radiation ($\mathbf{Q}_{lw\ top} > 250\ \frac{W}{m^2}$) and no precipitation. Similar to latent node 3 and 4, latent node 5 comprises dominantly information about subgrid-scale convective processes rather than large-scale geographic variability. Latent node 5 represents both deep convective events originating from the subtropics and mid-latitude stable conditions as can be already seen in the strong bimodality along the marginal distribution in Figure 3.10a.

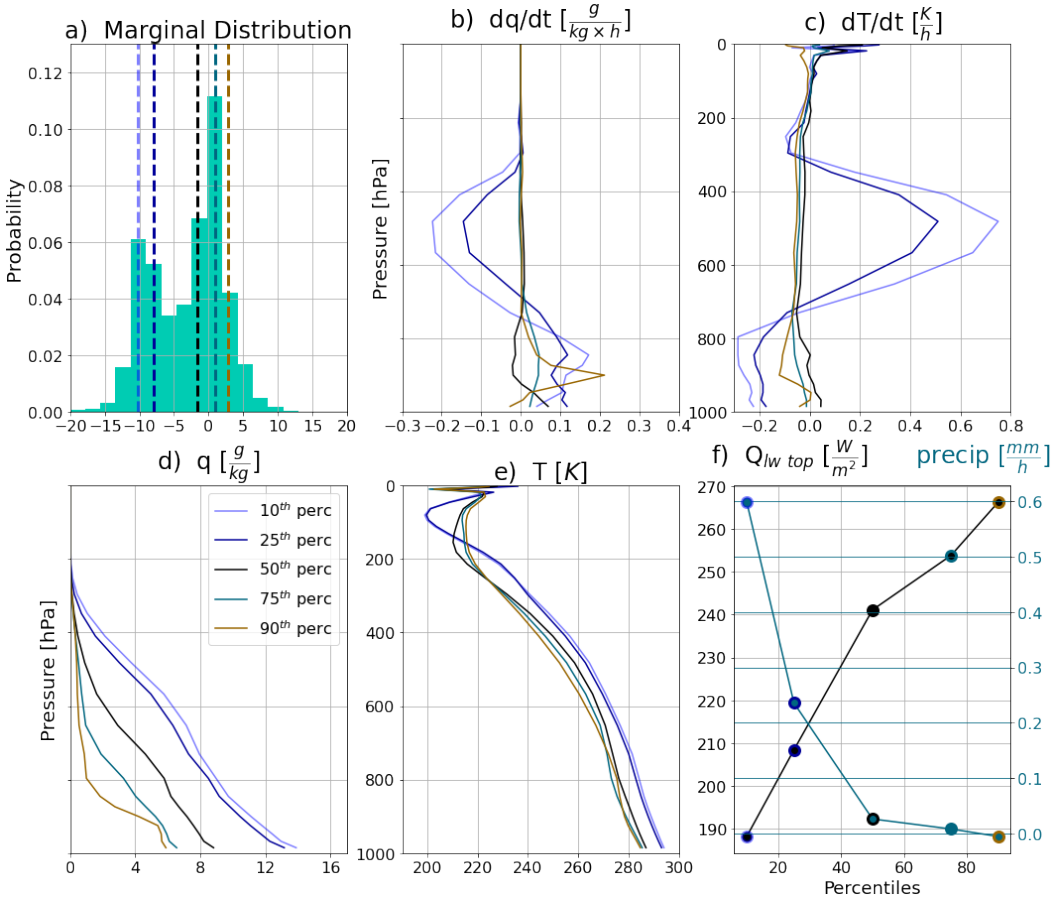

Figure 3.10.: Marginal distribution of latent node 5 (a) and the resulting generated vertical profiles of specific humidity tendencies **dq/dt** (b), temperature tendencies **dT/dt** (c), specific humidity **q** (d) and temperatures **T** (e). The dashed lines in the marginal distribution plot represent the chosen percentiles (see legend in subplot d) and the resulting effect of the respective translation $\mathbf{z}_{translation}$ on the profiles is shown in the subplots. Furthermore, the longwave heat flux at the model top ($\mathbf{Q}_{lw\ top}$) and the precipitation rate (**precip**) (f) are illustrated as function of the translation $\mathbf{z}_{translation}$ along the latent dimension 5. The marker-edge-color in panel f symbolise the respective percentiles of $\mathbf{z}_{translation}$. The black lines in subplots b-e indicate the generated reference state with $\mathbf{z}_{median}$. This Figure is directly reproduced from **Behrens** et al. 2022

3.4. A VED to Unveil and Understand Convective Processes, Convective Drivers and Convective Regimes in a Climate Model

This section condenses the main findings of my paper **Behrens** et al. 2022 and is directly based on the related Summary and Conclusion section.

This study has shown how a Variational Encoder Decoder (VED) can successfully machine learn a convective parameterization with considerable input compression while simultaneously enhancing the interpretability of deep learning methods, and enable better understanding of convective processes in climate models. We first showed that the VED is able to realistically reconstruct convective processes simulated by a superparameterized climate model, similar to previous studies based on a regular Artificial Neural Network (ANN) ar-

chitecture (Gentine et al. 2018; Rasp et al. 2018), but using automatically compressed input data. Furthermore, we demonstrated that the VED also enhances the interpretability of the relationship between large-scale climate fields and sub-grid-scale convective variables via its latent manifold, which is unfeasible via ANNs without attribution methods due to ANNs' large dimensionality (large number of hidden layers and nodes per layer). Our analysis is based on 9 months (equally split into training, validation and test data) of an aquaplanet simulation of the Super Parameterized Community Atmosphere Model (SPCAM). As shown in Figure 3.11a, the input variables of the VED resembled the large-scale climate fields (temperature, specific humidity and other thermodynamic drivers) from the general circulation model (CAM) passed onto the embedded cloud resolving model (SP). The latent space (lower dimensional manifold inside the network) of the VED had a dimensionality of five nodes, which is a small fraction of the dimensionality of the original input nodes information. To create an interpretable latent space, our optimal network reconstructed a combination of sub-grid-scale convective variables related to the SP component and large-scale climate variables associated with CAM. In comparison, as we have shown in the supplemental material, VEDs that attempt the traditional mapping from $\mathbf{X}$ to $\mathbf{Y}$ alone turn out to be less amenable to latent space exploration.

As a first step, we evaluated the reproduction performance of convective processes of the VED against a reference ANN (Rasp et al. 2018). The VED was capable of reconstructing the mean statistics of sub-grid-scale convective variables with an overall comparable, though slightly decreased, skill than the reference ANN despite the strong dimensionality reduction down to five latent nodes. This speaks to the dimensionality of information content required for a convective parameterization, and associated trade-offs. We found that compressing the input information did not overly distort the tropical wave spectrum. We showed that the choice of the latent space width is a critical hyperparameter for reproduction skills. Larger latent space widths ($\sim$ 8 nodes) yielded a reproduction performance of convective processes with almost the skill of the reference ANN, while smaller latent space widths ($\sim$ 2 nodes) still enabled an improved reproduction compared to a multi-dimensional linear regression baseline. We chose a latent space of five nodes as a sensible compromise between reproduction abilities of convective regimes and sensitivities separable in the latent manifold.

We began the analysis towards our main interest - latent space exploration with respect to physical interpretability – using traditional methods visualizing physical properties in a 2D projection of its leading PCs. This revealed that the VED distinguished day- and night-time conditions and varying strength of convective processes using the precipitation rate and outgoing longwave radiation as a proxy in its latent space (which was 2D compressed with a PCA for the purpose of visualization). The VED separated different global climate conditions and associated convective regimes from the poles to the equator in its latent space. The realistic reproduction of convective processes and climate conditions, along with the encapsulated information on geographic variability in an interpretable latent manifold, allowed a detailed analysis of governing drivers of convection and convective regimes with a VED.

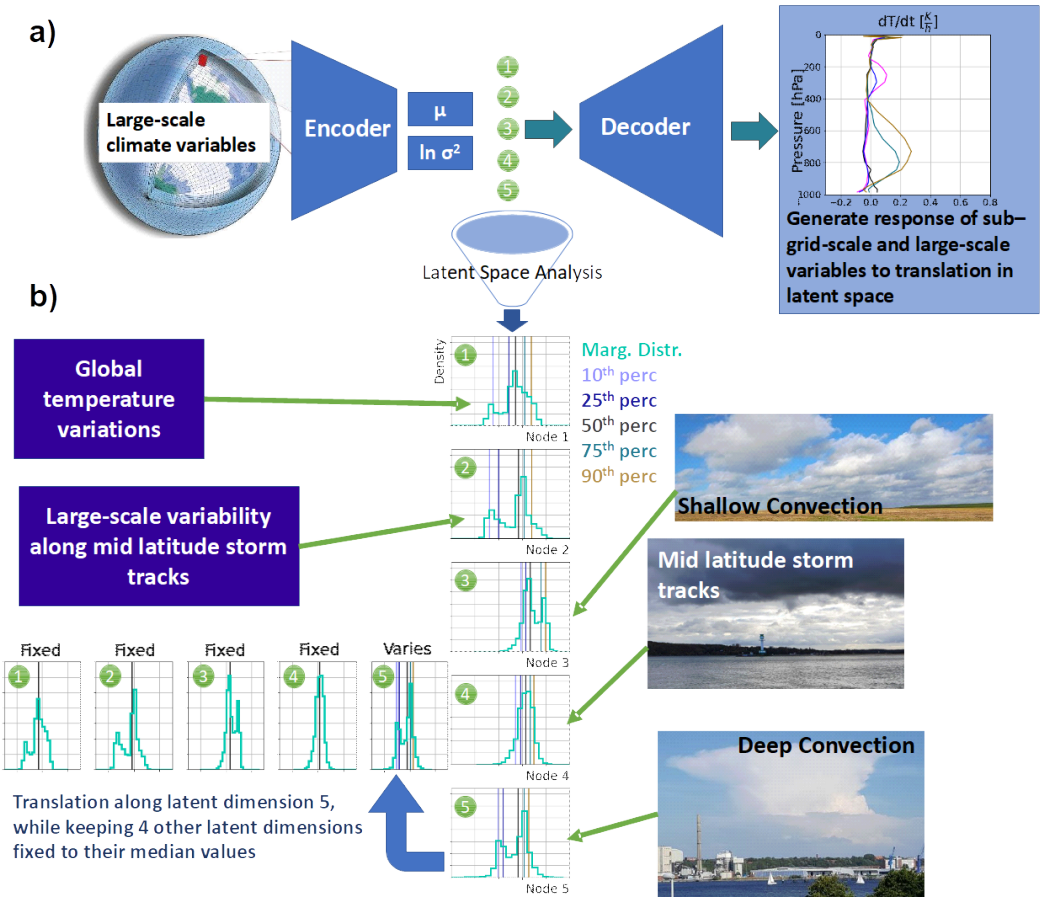

Figure 3.11.: Schematic of the VED setup **(a)** the investigated convective regimes and drivers of convective processes in the latent space of VED for each node **(b)**. The translation along each latent dimension is shown in this example for Latent Node 5. The schematic of the large-scale atmospheric grid in **(a)** was adapted from Schneider et al. 2017. This summary schematic is directly reproduced from **Behrens et al. 2022**

Our latent exploration was then deepened by investigating convective processes and related drivers via a generative modeling approach, i.e., forcing the decoder with the variability encapsulated along each latent dimension. The resulting temperature, specific humidity, heating, and moistening profiles successfully separated well-known large-scale driving climate conditions and convective regimes. Figure 3.11b summarizes the main results of this generative modeling approach. Overall, convective processes are controlled by large meridional gradients in temperature and specific humidity, from the equator to the poles, which were captured by the VED's Node 1 (Figure 3.11b). We identified the large-scale climate variability in specific humidity and temperatures along the mid-latitude storm tracks (Node 2, Figure 3.11b) as the other major driver of convective processes. Daytime stable, cold and dry tropospheric conditions suppress convective processes in the entire troposphere, whereas night-time unstable, warm and moist conditions in the troposphere drive precipitating mid-level convection. Apart from these large-scale nodes, the VED further reveals characteristics of distinct convective regimes on the remaining 3 latent nodes. The VED confined shallow

convective processes below 600 hPa within its Node 3 (Figure 3.11b); these processes are generally driven by surface diabatic fluxes and are predominantly originating from mid-latitudes and the subtropics. In anomalous dry conditions, positive surface diabatic heat fluxes during day-time enhance shallow convective processes associated with a convective moistening of the lower troposphere. The opposite is true in anomalous wet conditions during night-time. The mid-latitude storm tracks show large variability with respect to convective regimes associated with the eastward migrating frontal systems, features that were captured in the VED's latent space (Node 4, Figure 3.11b). In the surroundings of the warm front high, optically thin, non-precipitating cirrus-like convection is found. In contrast lower, optically thick cumulus-like convection with intermediate precipitation formation is predominant near the cold front. Furthermore, deep convective regimes in the subtropics were clearly captured by the VED (Node 5, Figure 3.11b). In this case, convective processes extend in the entire troposphere with a pronounced convective heating and drying near 500 hPa and are associated with intense precipitation. Opposing this extreme convective case, we found night-time, stable, cold and dry conditions in the free troposphere, which suppress convective processes on the other side of Node 5. Finally, while the interpretation of these convective regimes always required domain knowledge, the generative modeling approach simplified the analysis in comparison to other statistical analysis tools (e.g., correlations, clustering, attribution methods).

Repeating this analysis with an Encoder Decoder (ED) yielded almost identical reproduction capabilities compared to the VED, but the ED's latent space was significantly harder to interpret, with less pronounced center of actions for a given variable (see Figure A.5 and A.7 in Appendix A). This hindered the identification of convective regimes or large-scale drivers of convective predictability within the latent space of ED. For example, although the ED captured a cirrus-like regime, no cumulus or deep convective regimes could be found with the generative modeling method. Likewise, the connection between large-scale climate variables was often less pronounced for the ED, which resulted in larger uncertainties of the geographic origin of a specific sample compared to the VED.

We discovered convective regimes with the VED that are in general agreement with existing work focused on tropical convection (Frenkel et al. 2012, 2013; Frenkel et al. 2015; Huaman and Schumacher 2018). The specific humidity profile of the shallow convective regime of the VED was largely similar to the observed shallow convective latent heating profile in Huaman and Schumacher 2018 with a heating peak around 800 hPa. Furthermore, the heating profile of the mid-latitude cirrus-like regime of the VED compared well with that of the tropical stratiform regime shown in Frenkel et al. 2015 despite strong differences in the ambient conditions that led to their formation. Also the heating profiles of the mid-latitude cumulus regime of the VED and their tropical congestus expressed similarities in the lower troposphere with a pronounced convective heating peak above the boundary layer. Likewise, the VED's subtropical and tropical deep convection regime of Frenkel et al. 2015 were characterized by similar heating profiles. In our case, we identified these regimes solely based on SPCAM data in the latent space of the VED, where we did not prescribe the characteristics of each convective regime like it was done in the multi-cloud approach presented in Frenkel et al.

2012 and adapted from Khouider and Majda 2006. Furthermore, our approach was not based on inferred heating profiles via subclassing precipitation regimes (Stratiform, Convective, Shallow) as it was done for observational satellite products in Huaman and Schumacher 2018.

This work presented how convective processes, convective regimes, and large-scale drivers of convection in climate models can be investigated by leveraging generative machine learning (ML) approaches. Our approach enhanced the understanding of acting convective processes and the corresponding large-scale environment in which they form. As a next step, one could study cirrus-like or cumulus convection in detail by, for example, separating specific humidity and moistening tendencies related to the ice phase, linking how microphysical processes influence convection and are, in turn, affected by climate conditions (i.e., formation of ice phase, mixed phase or liquid phase clouds). Likewise, the development of regime-oriented ML-based convection parameterizations appears to be achievable with generative deep learning methods. Finally, VEDs could play an essential role in constructing new stochastic convection parameterizations, which could improve the representation of clouds and convection in Earth System Models. Our results suggest that VED representations of climate processes can effectively combine statistical prediction with data-driven analysis, paving the way towards machine learning-based Earth System Models that remain interpretable, albeit through the yet mostly unfamiliar eccentricities of latent space exploration.

In the next Chapter, I will build up on this analysis and will investigate how deep learning ensembles could improve the reproduction of convective processes in a realistic global Earth System Model using ideas from **Behrens** et al. 2022.

4. Reproducing convective processes of an Earth System Model with deterministic and stochastic deep learning ensembles

The following chapter is reproduced from **Behrens** et al. 2024 with small modifications of the nomenclature of the different deep learning models to facilitate reading and to homogenize this thesis. In this chapter I will present ways to investigate convective regimes and large-scale drivers of convective processes based on the latent space of one Variational Encoder Decoder using generative modeling. This chapter of the thesis is based on my paper that is currently in review (**Behrens** et al. 2024). It will assess the reproduction capabilities of stochastic and deterministic ensemble deep learning in comparison to individual deep learning models based on realistic global Earth System Model data. It is structured as follows: Section 4.1 gives an overview on the Earth System Model that is used. Section 4.2 briefly explains the neural network architectures that are used to construct the stochastic and deterministic ensembles. It is followed by section 4.3 that illustrates suitable ways how stochastic deep learning ensembles and deterministic counterparts can be built. The reproduction skill of subgrid convective processes of the stochastic and deterministic ensembles is evaluated in section 4.4. This chapter is concluded with a summary of the first part of the summary section focusing on reproduction skill presented in **Behrens** et al. 2024.

For the version of **Behrens** et al. 2024, that is currently in review, I, as the author of the thesis, contributed all figures, tables and large parts of the code to produce them. Furthermore I let the writing and the analysis of the paper.

4.1. Climate Modeling Setup

This section is directly based on the section with the identical name in my publication that is currently in review (**Behrens** et al. 2024).

In this study we use the Super Parameterized Earth System Model (SPCESM) Version 2.1.3 (SPCESM2, Danabasoglu et al. 2020) for the construction of our stochastic and deterministic parameterizations. The atmospheric component of the Community Earth System Model (CESM) version 2 (CESM2) is the Community Atmosphere Model (CAM) version 6 (CAM6). In our configuration CAM6 is run without interactive chemistry, and thus radiatively-active aerosols and gases are prescribed. CAM6 has a horizontal grid size of approximately $2^\circ \times 2^\circ$ (144×96 grid cells). The vertical axis consists of 26 levels on a hybrid-sigma grid with

14 tropospheric levels (p > 200 hPa). CAM6 has a timestep of 1800 s. To represent subgrid processes (convection, subgrid radiative effects, and fine-scale eddies) in each grid cell of CAM6, we use a Superparameterization (SP) (Grabowski 2001; Khairoutdinov and Randall 2001). SP, also known as multiscale modeling framework (MMF) (i.e., Yu et al. 2023), consists of 32 nested two-dimensional grid columns with a finer horizontal resolution of 4 km, which partially resolves deep convection and associated gravity waves. These grid columns are meridionally oriented (north to south) as described in Pritchard et al. 2014. SP and CAM6 share the same vertical discretization after an initial interpolation at the beginning of each SP time step (20 s), from the 24 levels of SP to the CAM6 vertical axis. Furthermore the configuration of SP we employ uses a Smagorinsky 1.5-order turbulence scheme to parameterize fine-scale turbulence and a one-moment microphysics scheme (Grabowski 2001; Khairoutdinov and Randall 2001). The microphysics scheme allows the separation into cloud ice and liquid water phase and respective phase tendencies. Horizontal advection of high resolution convection related fields (momentum, cloud condensates) from the nested SP to the neighbouring CAM6 cells' nested SP is neglected. Instead the advection of these convection-related fields is handled via the dynamical core of the coarse CAM6 model with known limitations (Jansson et al. 2022).

The atmosphere is coupled to the land component, Community Land Model version 5 (CLM5), which includes realistic topographic boundary conditions. We use prescribed sea surface temperatures and sea ice fields Merged Hadley - National Oceanic and Atmospheric Administration / Optimum Interpolation Sea Surface Temperature and Sea Ice Concentration data set (MH-NOAA/OI-SST-SIC) (Hurrell et al. 2008). Our simulations are driven by observed solar spectral irradiance and concentrations of aerosols and atmospheric trace gases (e.g., ozone). For a more detailed description of CESM2, we point the interested reader to Danabasoglu et al. 2020, and for SP to Khairoutdinov and Randall 2001. The SPCESM2 version used here can be found on GitHub (https://github.com/SciPritchardLab/CESM2-ML-coupler).

The next section explains the deep learning approaches we developed to build a stochastic or ensemble, data-driven emulator of SP.

4.2. Deep Learning Parameterizations

This section is directly based on the section with the identical name in my publication that is currently in review (**Behrens** et al. 2024).

In this section, we first describe the general approach the training of the deep learning subgrid processes in SPCESM2 (section 4.2.1). We then describe the deep learning (DL) algorithms (section 4.2.2), before constructing stochastic and deterministic DL parameterizations in the next section (section 4.3). Table 4.1 gives an overview of our developed stochastic and deterministic parameterizations. Moreover it helps the reader to understand the acronyms of the different models that we will use in the following.

4.2.1. General Approach

DL parameterizations aim to represent the *aggregate* effect of subgrid processes, as simulated by the SP component of SPCESM. To achieve this, DL algorithms predict a grid-averaged subset of SP's subgrid variables based on the large-scale atmospheric conditions modeled by CAM6, hereafter referred to as "CAM variables". During the neural network-coupled climate model simulations, these predicted subgrid variables (i.e., vertical profiles of subgrid specific humidity and temperature) are used to couple the atmospheric model with the other components at the surface (e.g., CLM5 land model and boundary conditions from the ocean model). The application of DL for reproducing SP variables speeds up the emulation of the fine-scale convection resolution in the corresponding climate simulations while maintaining the high-quality representation of subgrid processes provided by the superparameterization (Rasp et al. 2018).

The input data closely follows the CAM variables except for one additional variable (Prec_{t-dt}) that was helpful for the performance of the DL algorithms. The input $\boldsymbol{X}$ (Figure 4.1) is a stacked vector of size 109 and is given by:

$$\boldsymbol{X} = \left[\boldsymbol{q}(\boldsymbol{p}) \quad \boldsymbol{T}(\boldsymbol{p}) \quad \boldsymbol{q_{cl}}(\boldsymbol{p}) \quad \boldsymbol{q_{ci}}(\boldsymbol{p}) \quad p_{\text{surf}} \quad Q_{\text{sol}} \quad Q_{\text{sens}} \quad Q_{\text{lat}} \quad \text{Prec}_{t-dt} \right]^T, \tag{4.1}$$

where $\boldsymbol{X}$ includes the 4 vertical profiles (with 26 vertical levels) of specific humidity $\boldsymbol{q}(\boldsymbol{p})$ [g/kg], temperature $\boldsymbol{T}(\boldsymbol{p})$ [K], cloud liquid water content $\boldsymbol{q_{cl}}(\boldsymbol{p})$ [g/kg], and cloud ice water content $\boldsymbol{q_{ci}}(\boldsymbol{p})$ [g/kg]. Additionally, $\boldsymbol{X}$ comprises the scalar values of surface pressure p_{surf} [hPa], solar insolation Q_{sol} [W/m^2], surface sensible Q_{sens} [W/m^2] and latent heat flux Q_{sens} [W/m^2] from the current timestep. Additionally we use the previous timestep's precipitation Prec_{t-dt} [mm/h] as input to complement the other CAM variables. Including Prec_{t-dt} strongly improves the prediction of near-surface heating and moistening tendencies that are of great importance for the coupling to the CLM5 land model, which is aligned with the findings of previous studies (Han et al. 2020; Han et al. 2023).

The output vector ($\boldsymbol{Y}$, predictants or target) of our data-driven parameterization has a length of 112 (Fig. 4.1) and is given by:

$$\boldsymbol{Y} = \left[\dot{\boldsymbol{q}}(\boldsymbol{p}) \quad \dot{\boldsymbol{T}}(\boldsymbol{p}) \quad \dot{\boldsymbol{q}}_{\boldsymbol{cl}}(\boldsymbol{p}) \quad \dot{\boldsymbol{q}}_{\boldsymbol{ci}}(\boldsymbol{p}) \quad \text{Snow}_{\text{CRM}} \quad \text{Prec}_{\text{CRM}} \quad \boldsymbol{Y_{rad}} \right]^T, \tag{4.2}$$

where $\boldsymbol{Y}$ includes the 4 vertical profiles of specific humidity tendency $\dot{\boldsymbol{q}}(\boldsymbol{p})$ $[\frac{g}{kg \times h}]$, temperature tendency $\dot{\boldsymbol{T}}(\boldsymbol{p})$ [K/h], cloud liquid water tendency $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}(\boldsymbol{p})$ $[\frac{g}{kg \times h}]$, and cloud ice water tendency $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}(\boldsymbol{p})$ $[\frac{g}{kg \times h}]$. Here, we use "tendency" and the notation $\dot{y}$ as a shorthand for the difference between the values of state variables before and after the SP call, normalized by the CAM6 timestep ($dt = 1800\,\text{s}$, see e.g., Appendix B Equation B.1). Note that this call precedes and does not include the calculations for surface coupling. $\boldsymbol{Y}$ further includes the cloud-resolving precipitation (Prec_{CRM}) and snow rates (Snow_{CRM}), both simulated by SP and expressed in units mm/h. To facilitate reading, we grouped all radiative outputs required for coupling to the surface in $\boldsymbol{Y_{rad}}$:

$$\boldsymbol{Y_{rad}} = \left[Q_{\text{lw surf}} \quad Q_{\text{sw surf}} \quad Q_{\text{sol lw}} \quad Q_{\text{sol lw, diff}} \quad Q_{\text{sol sw}} \quad Q_{\text{sol sw, diff}} \right]^T, \tag{4.3}$$

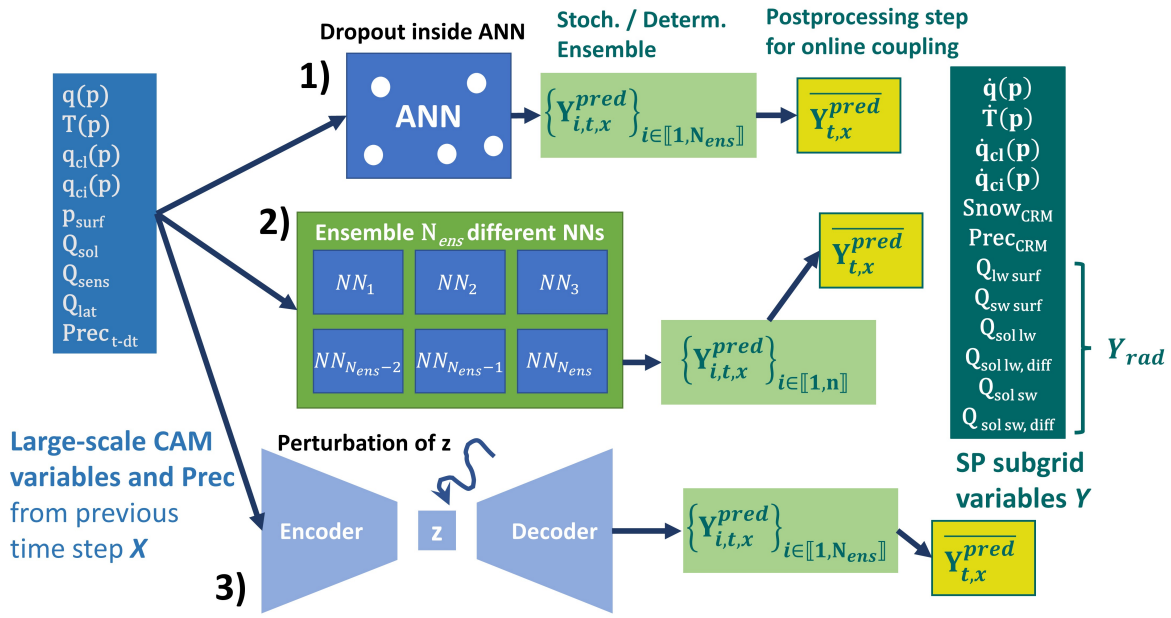

Figure 4.1.: We compare three stochastic parameterization strategies for reproducing the superparameterization (SP), which simulates SP subgrid variables ($\boldsymbol{Y}$) based on large-scale Community Atmosphere Model (CAM) variables ($\boldsymbol{X}$): 1) Applying Monte-Carlo dropout to a single artificial neural network (ANN) to generate a prediction based on the mean of N_{ens} draws. 2) Employing a subset of n neural networks randomly drawn from a set of N_{ens} deterministic neural networks to generate n predictions that can be averaged for the final prediction. 3) Perturbing the latent space of a Variational Encoder-Decoder network N_{ens} times to produce N_{ens} predictions that are subsequently averaged. In addition, Table 4.1 gives an overview of our developed deep learning ensemble parameterizations. This Figure is reproduced with minor modifications from **Behrens** et al. 2024

where $\boldsymbol{Y}_{\mathbf{rad}}$ includes the downward surface longwave heat flux $Q_{\text{lw surf}}$, the downward surface shortwave heat flux $Q_{\text{sw surf}}$, the near-infrared part of the downward solar radiation at the surface, decomposed into its direct ($Q_{\text{sol lw}}$) and diffuse ($Q_{\text{sol lw, diff}}$) components, and the direct ($Q_{\text{sol sw}}$) and diffuse ($Q_{\text{sol sw, diff}}$) components of the solar radiation's visible wavelengths; all are in units of W/m^2. In the following, we couple the predictions of the surface radiative fluxes $\boldsymbol{Y}_{\mathbf{rad}}$ to CESM2 to investigate also the stability of CESM2 with such deep learned radiative fluxes. This contrasts our work to Han et al. 2023 that sidestepped the coupling of these crucial terms. Apart from these radiative fluxes, momentum and mass fluxes are also used to couple CAM6 to the surface.

For DL algorithms that involve multiple input and output variables with different physical units, a suitable normalization is important for both inputs ($\boldsymbol{X}$) and outputs ($\boldsymbol{Y}$), as normalization choices affect their relative importance during training. We normalize each of the inputs by subtracting its mean and dividing the resulting difference by the corresponding range, resulting in normalized inputs between -1 and 1. We normalize each output variable using a reference standard deviation as in **Behrens** et al. 2022 (see Appendix B B.2 for details).

To avoid spatiotemporal correlations, we extract 84 days per year, specifically 7 consecutive days from each month, for training (Year 2013), validation (2014), and testing (2015). These

data originate from a historical SPCESM2 run spanning 2003 to 2015, ensuring the exclusion of any model spin-up effects. Each dataset contains 55,572,480 samples, and is balanced with respect to the diurnal and seasonal cycles.

4.2.2. Machine Learning Algorithms

To map **X** to **Y**, we implement two different model types (Table 4.1): Artificial Neural Networks (ANNs) and Variational Encoder Decoder (VED, Kingma and Welling 2014) structures, which use a lower-dimensional manifold between the encoding and decoding part of the network, also known as "latent space" in data science. In the following we will briefly describe these two network types and the associated hyperparameter searches we conducted. We will use these two neural network types to build stochastic and deterministic ensemble DL parameterizations (see section 4.3).

Artificial Neural Networks

ANNs consist of a set of fully connected layers that contain nodes. These nodes perform a non-linear regression task, and their weights and biases are optimized to reduce a loss function. The linear sum of the previous layer is then passed through a non-linear function, referred to as an "activation function". Our ANNs have an input layer of 109 nodes (**X**'s length) and an output layer of $N_{outputs}$ = 112 nodes (**Y**'s length). To optimize the ANNs' weights and biases, we use the mean-squared error (MSE) between the predictions ($\mathbf{Y}^{pred}$) and the original data (**Y**) as our loss function (Equation 4.4).

$$\mathrm{MSE}\left(\boldsymbol{Y}, \boldsymbol{Y}^{\mathbf{pred}}\right) = \frac{1}{N_{\mathrm{outputs}} \times N_{\mathrm{batches}}} \sum_{k=1}^{N_{\mathrm{batch}}} \sum_{j=1}^{N_{\mathrm{outputs}}} \left(Y_{j,k} - Y_{j,k}^{\mathrm{pred}}\right)^2, \tag{4.4}$$

where N_{batch} is the batch size (i.e., the number of samples fed to the network per backpropagation step), $Y_{j,k}^{\mathrm{pred}}$ is the network's prediction of the j-th output for the k-th sample in the batch, and $Y_{j,k}$ the corresponding target value we aim to predict.

To optimize the overall setup of the ANNs we conducted an extensive hyperparameter search, including the batch size, the learning rate (i.e., the down-gradient step with respect to the loss function for the network optimization during training), the number of nodes per layer (integral parts of the network, which determines the number of weights and biases to be optimized during training), the number of hidden layers (network layers between the input and output layer), and the activation function (see Appendix B section B.2.1). We find that the performance of ANNs is most sensitive to changes in learning rate and batch size. Other predefined settings of our ANNs are the use of Adam (Kingma and Ba 2014) as optimizer (an algorithm that improves the network performance during training) and a predefined learning rate schedule (which decreases the initial learning rate after a certain epoch, see Appendix B section B.2.1 for details). The hyperparameters of the 7 best-performing ANNs

are summarized in Table B.3. We will use these ANNs as building blocks of our deterministic and stochastic parameterizations and compare them to VEDs, which we describe next.

Variational Encoder Decoders

Similar to ANNs, the VEDs comprise input and output layers and dense fully connected hidden layers. The main difference between the VEDs and ANNs is the dimensionality reduction within the network into a small latent space and the addition of stochasticity in the latent space. The encoding part (Encoder) compresses the information down to the latent space through hidden layers with successively smaller node numbers from layer to layer (see Figure 4.1). This latent space is a lower-order representation of the original information with a latent space width of size N_{latent}, which is the number of nodes in the latent space. Within the latent space, the mean μ and logarithmic variance $\log \sigma^2$ of the latent distributions are optimized. μ and $\log \sigma^2$ are then used in a reparameterization to generate the stochastic latent variables $\mathbf{z}$ (**Behrens** et al. 2022; Kingma and Welling 2014). Different techniques can be used to interpret the encoded information with respect to the input and output data (see **Behrens** et al. 2022; Mooers et al. 2023; Shamekh et al. 2023). The latent variables $\mathbf{z}$ are then the input to the decoding part of the network (Decoder), which maps the information back to generate predictions. The VED's loss function $\mathcal{L}_{\text{VED}}$ is the sum of the MSE loss function given by equation 4.4 and a Kullback-Leibler (KL) loss term, which can be interpreted as a regularizer of the latent distribution towards a normal decorrelated distribution for disentanglement (Kingma and Welling 2014):

$$\mathcal{L}_{\text{VED}}\left(\boldsymbol{Y}, \boldsymbol{Y}^{\mathbf{pred}}\right) = \text{MSE}\left(\boldsymbol{Y}, \boldsymbol{Y}^{\mathbf{pred}}\right) + \lambda \times \underbrace{\frac{1}{2N_{\text{batch}}} \sum_{k=1}^{N_{\text{batch}}} \sum_{z=1}^{N_{\text{latent}}} \left(\mu_{z,k}^2 + \sigma_{z,k}^2 - \ln \sigma_{z,k}^2 - 1\right)}_{\text{KL Loss}}, \tag{4.5}$$

where the regularization factor (λ) regulates the weight given to the MSE and KL losses during training. We push this balance towards an enhanced reconstruction (smaller MSE) to the expense of the KL loss term ($\lambda < 1$). In this study we use a static regularization factor, so a constant λ that can be used as an additional hyperparameter. Our approach to construct the VED deviates from the standard data science approach of a probabilistic Variational AutoEncoder (VAE) Decoder in two ways. First, we use the MSE (Equation 4.4) between $\mathbf{Y}^{pred}$ and $\mathbf{Y}$ to measure the reconstruction error, instead of the squared error between the prediction of $\mathbf{X}^{pred}$ and respective input variables $\mathbf{X}$ that is often used in the loss function of VAEs (see e.g., Mooers et al. 2023 for more details). This allows us to directly quantify how well the original convective processes in $\mathbf{Y}$ are reproduced. Secondly, the main focus of the training of our VED lies on an accurate reproduction and not on a perfect disentanglement inside the latent space of the VED, thus using a strong regularization of the KL loss. Such an enhanced determinism of the VEDs is beneficial to increase the general performance on the

	Acronym climate model	Acronym parameterization	No. Networks	Method	Stochastic parameter
deterministic	$\overline{\text{ANN}}$-CESM	$\overline{\text{ANN}}$	7	deterministic ensemble mean prediction	-
	-	$\overline{\text{VED}}$	6	deterministic ensemble mean prediction	-
stochastic	-	ANN-dropout[a]	1	dropout	dr=0.01
	ANN-ens-CESM	ANN-ensemble[b]	7	ensemble	randomly draw 5 out of 7 members for averaging
	-	VED-draws[c]	1	latent space reparameterization	7 draws
	-	VED-static[d]	1	latent space perturbation	7 draws with scalar $\alpha = 0.5$
	-	VED-varying[e]	1	latent space perturbation	7 draws with α-array

[a] dropout including 7 samples per prediction of ANN 1 (Supporting Information Table S4)
[b] based on all ANNs (Table S4), the 5 out 7 members are randomly drawn for each timestep
[c] based on 7 predictions of VED 1 (Table S5)
[d] based on 7 predictions of VED 1 (Table S5) with scalar $\alpha = 0.5$
[e] based on 7 predictions of VED 1 (Table S5) with α-array

Table 4.1.: Summary of the stochastic and deterministic parameterizations we developed. The 2^{nd} and 3^{rd} column indicate the acronyms of the respective parameterizations in the Community Earth System Model version 2 (CESM2; section 5.4) and in our offline evaluation (sections 4.4, 5.2, 5.3). The other columns show for each parameterization the number of DL networks used, the method used to generate the predictions, and key stochastic parameters for the stochastic parameterizations. This table is reproduced with minor modifications from **Behrens** et al. 2024.

complex multi-input, multi-output data set of the superparamaterization compared to a fully probabilistic setup (Yu et al. 2023).

The list of evaluated hyperparameters for the VEDs includes batch size, learning rate, number of nodes in the first or last hidden layer of the Encoder or Decoder, the number of hidden layers of the Encoder or Decoder, the latent space width and the regularization factor λ. We find that the VED's performance is most sensitive to the batch size, learning rate, latent space width and the regularization factor. Details about the conducted hyperparameter search and VED architecture can be found in the Appendix B (section B.2.1 and Table B.4).

The next subsection explains suitable ways, that I developed in **Behrens** et al. 2024, to construct stochastic and also deterministic ensemble deep learning ensemble parameterizations.

4.3. Stochastic and Deterministic Ensemble Deep Learning Parameterizations

This section is directly based on the subsection with the identical name in my publication that is currently in review (**Behrens** et al. 2024).

Here, we present three suitable approaches to develop a stochastic parameterization based on the machine learning algorithms introduced in the previous subsection (Figure 4.1): dropout inside an ANN as a source of stochasticity, ensemble prediction of a number of neural networks, and a latent space perturbation of a single VED, inspired by the enhanced interpretability gained with latent space perturbations shown in **Behrens** et al. 2022.

4.3.1. Dropout

Dropout, also known as Monte Carlo Dropout (MCD), is widely applied to reduce overfitting, which is characterized by an elevated training performance compared to validation or test performance (Hinton et al. 2012). In addition, MCD can be used to quantify the uncertainty of predictions, and therefore to estimate stochasticity. It has been shown that the resulting uncertainty quantification and stochastic predictions of MCD have substantial limitations in particular an underestimation of systematic spread and the inflation of deterministic errors compared to more complex methods to construct stochastic predictions (Haynes et al. 2023).

With these caveats in mind, we use MCD as a simple baseline for our stochastic parameterizations. We apply MCD to one of the best-performing ANNs (ANN-dropout in Table 4.1 and hereafter) by adding a dropout layer after the last hidden layer of the network directly in front of the output layer. We choose a dropout rate **dr** of 0.01, meaning that 1% of the input linkages to the dropout layer are randomly discarded for each sample. While this small dropout rate underestimates the spread, higher values of the dropout rate (e.g., 0.05) significantly deteriorate reconstruction quality.

To construct an ensemble with MCD (Figure 4.1) we repeat the sample-level prediction N_{ens} times (see equation 4.6), where N_{ens} is the ensemble size and i symbolizes the i-th sampling of

the deterministic ANN with active dropout **dr**. Due to the active dropout the resulting ensemble $\{\mathbf{Y}^{pred}_{i,t,x}\}$ is of stochastic nature and provides uncertainty quantification for each timestep t and grid cell x. We use the ensemble mean of the MCD ensemble and individual members to compare against other approaches to construct a stochastic and ensemble parameterization for CESM2.

$$\overline{\boldsymbol{Y}^{\text{pred}}} = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \boldsymbol{Y}_i^{\text{pred}} \quad , \quad \boldsymbol{Y}_i^{\text{pred}} = (\text{Best ANN}_{\text{dr}})_i \, (\boldsymbol{X}) \tag{4.6}$$

4.3.2. Ensemble Method

Ensemble predictions are one common way to provide uncertainty quantification of weather forecasts (Gneiting and Raftery 2005) and climate projections (Eyring et al. 2016), as climate and weather are governed by internal variability and stochasticity; some of them due to convective and turbulent processes (Berner et al. 2017). Inspired by these traditional climate modeling approaches, we develop ensemble based stochastic and deterministic parameterizations using ANNs and VEDs (stochastic: ANN-ensemble; deterministic: $\overline{\text{ANN}}$, $\overline{\text{VED}}$ in Table 4.1). These parameterizations will prove to have considerable advantages relative to a single deterministic prediction of an individual neural network. In the following we use the terminology "deterministic ensemble" for ensembles built without additional subsampling (n = N_{ens} in equation 4.7, where N_{ens} is the maximum number of ensemble members and n is the used ensemble size). To account for limitations when it comes to the computational overhead and the applicability of the ensemble method, we restrict the ensemble size n to 7 members. Here, we chose a similar number of ensemble members as Han et al. 2023, who used an ensemble size of 8. We acknowledge that this number of ensemble members is a critical hyperparameter for ensemble predictions and larger (more diverse) ensembles yield often better performance over smaller ones with decreased spread between the ensemble members. Yet, larger ensemble require higher cost and memory so that they might not be practical.

We generate either a deterministic (n = N_{ens}) or a stochastic ensemble (see equation 4.7) for each timestep t and grid cell x. In the stochastic case (n < N_{ens}) we randomly draw for each time step and grid cell a subset of members of size n < N_{ens} out of the deterministic ensemble. Equation 4.7 shows the computation of the ensemble mean that we use for our online coupling experiments (Figure 4.1),

$$\overline{\boldsymbol{Y}^{\text{pred}}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{Y}_i^{\text{pred}} \quad , \quad \boldsymbol{Y}_i^{\text{pred}} = \text{NN}_i \, (\boldsymbol{X}) \tag{4.7}$$

where n elements are randomly drawn out of the entire ensemble N_{ens} in the stochastic case, decreasing n towards 1 yields a larger degree of stochasticity. We tested the number of samples that are randomly drawn and found that 5 out 7 members is a good compromise between added stochasticity and the overall reproduction skill of convective processes. For ensemble sizes smaller than 5 the general reproduction skill is deteriorating. In the following we show the

results of an ANN based stochastic ensemble with 5 out of 7 members (ANN-ensemble, Table 4.1), which illustrates the applicability of such an approach to generate stochasticity (Figure 4.1). The added value of stochasticity for the offline performance is negligible based on the analysed offline metrics, but we see an improved reproduction of precipitation extremes with the ANN-ensemble in comparison to the deterministic $\overline{\text{ANN}}$ parameterization when coupled to CESM2 later on.

4.3.3. Latent Space Manipulation

This method is inspired by the interpretability and the potential of perturbing the latent space of the VED (**Behrens** et al. 2022).

We develop a two-step approach to build a stochastic parameterization via latent space perturbation. First, we train one of the best-performing VEDs (Table B.4) to achieve a realistic reproduction of convection related SP variables $\boldsymbol{Y}$. This particular VED is the base for the VED-static and VED-varying stochastic parameterizations (Table 4.1) that use latent space perturbation. We perturb the latent variables $\mathbf{z}_i$ via Gaussian Noise $\mathcal{N}(0, \alpha_i)$ with a mean 0 and standard deviation α along all dimensions $\mathbf{z}$ of the VED's latent space with width N_{latent} (see equation 4.8). α_i is a hyperparameter that controls the magnitude of the Gaussian Noise added to each latent dimension. The resulting perturbed samples for each timestep t and grid cell x are fed into the decoder of the VED to generate a stochastic ensemble parameterization (equation 4.8).

$$\overline{\boldsymbol{Y}^{\text{pred}}} = \frac{1}{N_{\text{ens}}} \sum_{i=1}^{N_{\text{ens}}} \boldsymbol{Y}_i^{\text{pred}} \quad , \quad \boldsymbol{Y}_i^{\text{pred}} = \text{VED}_{\boldsymbol{z}_i + \mathcal{N}(0, \boldsymbol{\alpha}_i)} (\boldsymbol{X}) \, . \tag{4.8}$$

In equation 4.8, we create a stochastic ensemble by perturbing a single VED's latent space in two different ways: Either by adding isotropic Gaussian noise to the latent variables ("VED-static", Table 4.1) with $\alpha_i = 0.5$ to all latent variables $\mathbf{z}_i$, or by adding anisotropic Gaussian noise whose standard deviation depends on the latent dimension ("VED-varying", Table 4.1). We evaluate the performance of these two stochastic parametrizations against a stochastic paramaterization of the identical VED without latent space perturbation ("VED-draws", Table 4.1). VED-draws uses the repetitive draw from the latent space distribution based on the reparameterization. We investigate that VED-draws has limitations in the reproduction of convective processes and the representation of robust uncertainty quantification of them (Figure B.6). In detail, the spread of VED-draws is considerably smaller compared to the two stochastic parameterizations with latent space perturbation. Therefore we do not show VED-draws in the following to simplify the visualization of our results.

Latent space perturbation has similarities to existing stochastic ensemble approaches in numerical weather forecasts or seasonal predictions, like the Stochastic Perturbed Parameter Tendencies scheme (SPPT) (Buizza et al. 1999; Christensen et al. 2015) where the subgrid source terms in the general equation of $\mathbf{T}$, $\mathbf{q}$ and the horizontal velocities are perturbed with multiplicative random noise. One difference is that our approach is acting on a lower

dimensional latent space and not directly on specific output variables or tendencies like SPPT. This ensures that the added noise is well conditioned on the large scale CAM variables **X** and the resulting perturbations represent realistic variations in large-scale climate variables.

We develop a thorough strategy for suitable α_i latent space perturbation. Its objective is to find a balance between the reproduction skill and the ensemble spread of output predictions $\mathbf{Y}^{pred}$ by adjusting α_i (see Appendix B section B.5 for details). In the following we show the applicability of the latent space perturbation approach tuned for one of the best-performing VEDs (VED 1, Table B.4) and compare it against the ensemble parameterizations.

In the next section I will evaluate the reproduction skill of subgrid convective processes of my developed ensemble deep learning parameterizations (Table 4.1) against individual deep learning architectures based on realistic global Earth System Model test data data from SPCESM.

4.4. Results: The added value of ensembles and stochasticity for the reproduction of subgrid convective processes

This section is based on the first part of the "offline" results section presented in **Behrens** et al. 2024. Herein I illustrate the reproduction capabilities of my developed parameterizations for subgrid convective processes with two standard machine learning metrics the Coefficient of Determination (R^2), see equation 3.5, and the Mean Absolute Error (MAE).

We start our offline benchmark analysis by evaluating the reproduction performance of the different stochastic parameterizations compared to the deterministic ensembles and individual networks with respect to SP test data (Table 4.1). Figure 4.2 shows the median coefficient of determination R^2 across all horizontal grid cells for the vertical profiles of $\dot{q}$ (Figure 4.2a), $\dot{T}$ (4.2b). The respective vertical profiles of median R^2 for $\dot{q}_{cl}$ and $\dot{q}_{ci}$ are displayed in Figure B.1. We compute R^2 for the entire hold-out test data set along the time dimension (= 4020 timesteps) in each of the grid cells and for all output variables **Y**.

All DL models in Figure 4.2 show an elevated reproduction skill for $\dot{T}$ compared to $\dot{q}$. The majority of models have a median $R^2 > 0.5$ for these two tendency fields. Corresponding plots of median R^2 for $\dot{q}_{cl}$ and $\dot{q}_{ci}$ can be found in Figure B.1. These condensate tendencies are more challenging to fit skillfully (Figure B.1), likely due to their small absolute magnitude as well as overall noisy and stochastic nature. For these vertical tendency profiles we see a median R^2 below 0.3 for all models. In Section 5.4 we will discuss this weaker offline performance associated with unstable CESM2 simulations, when condensate tendencies are included in the coupling. In general, DL models show a reproduction minimum in the lower troposphere and planetary boundary layer (> 800 hPa), due to the turbulent and chaotic nature of convective processes on these levels. The coefficient of determination indicates low reproduction skill above 200 hPa for the DL models for all variables except for $\dot{T}$ (Figure 4.2, B.1). However we see that the related mean absolute errors (MAEs) for $\dot{q}$, $\dot{q}_{cl}$, $\dot{q}_{ci}$ above 200 hPa are almost

zero (Figure B.5). This underlines that R^2 is not an optimal metric for the upper levels of the atmosphere with negligible specific humidity in the test data set (Yu et al. 2023).

The advantages of the ensemble methods are immediately clear. In general, the average of the deterministic ANN ensemble ($\overline{\text{ANN}}$, Table 4.1) and the average of the stochastic ANN ensemble (ANN-ensemble, Table 4.1) show in general an increased reproduction skill compared to single deterministic neural network predictions (grey lines in the background of Figure 4.2). $\overline{\text{ANN}}$ and ANN-ensemble show virtually an equivalent but improved performance, with their R^2 difference ($\overline{\text{ANN}}$ - ANNs) around 0.02. In general the performance difference for $\dot{\boldsymbol{T}}$ between the $\overline{\text{ANN}}$ and ANN-ensemble and single ANN predictions is negligible. In the lower troposphere one ANN has a slightly improved reproduction for $\dot{\boldsymbol{T}}$ compared to $\overline{\text{ANN}}$ and ANN-ensemble. The "quasi-deterministic" VED ensemble ($\overline{\text{VED}}$) and the dropout-based ANN ensemble (ANN-dropout) result in enhanced reproduction skill compared to single VEDs, but these approaches are within the performance range of single ANNs. A single VED with latent space perturbation (VED-static, VED-varying; Figure 4.2) shows reproduction capabilities in the range of other individual VEDs or VED-draws (not shown). However, we find that the median R^2 decays with increasing magnitude of the perturbation $\boldsymbol{\alpha_i}$ in initial experiments (Figure B.14). This points to the fact that the magnitude of the latent space perturbation has to be well chosen to reach a good balance between reproduction skill and the diversity (ensemble spread) of the ensemble. We will see in the following that the perturbation of the latent space strongly improves the ensemble spread and can be well conditioned for a variety of output variables $\boldsymbol{Y}$.

Ensemble methods based on multiple ANNs improve the skill within the planetary boundary layer, which is a known challenge of DL subgrid parameterizations (**Behrens** et al. 2022; Gentine et al. 2018; Mooers et al. 2021). This is shown in Figure 4.2a, in which the minimal median R^2 for subgrid moistening $\dot{\boldsymbol{q}}$ in the boundary layer increases by about 0.05 between individual ANNs and the deterministic ensemble $\overline{\text{ANN}}$ or the stochastic ensemble ANN-ensemble. To deepen the analysis, we focus on the pressure level of 956 hPa, where the differences between $\overline{\text{ANN}}$, ANN-ensemble and individual ANNs are largest. Therefore we compare the global maps of R^2 of $\dot{\boldsymbol{q}}$ on 956 hPa of the two ANN ensemble parameterizations with ANN-dropout and ANN 1, as an example of a skillful individual member of $\overline{\text{ANN}}$ and ANN-ensemble. We see that the increase in reproduction skill is attributable to an improved representation of convective processes in the planetary boundary layer over Antarctica, the adjacent Southern Ocean and also over the Arctic Ocean (Figure B.2). In contrast, ANN-dropout shows a poorer performance over these regions compared to ANN 1, which suggests that even a minimal dropout rate leads to generally weaker reproduction of shallow convective processes compared to individual ANNs.

There is no substantial added value of the ensemble approaches evident for precipitation rates and radiative fluxes. We see high reproduction capabilities (median $R^2 > 0.8$, see Figure B.3), comparable to reproduced 2D fields of single ANNs. This suggests that these variables can already be learned with high skill with single deterministic networks.

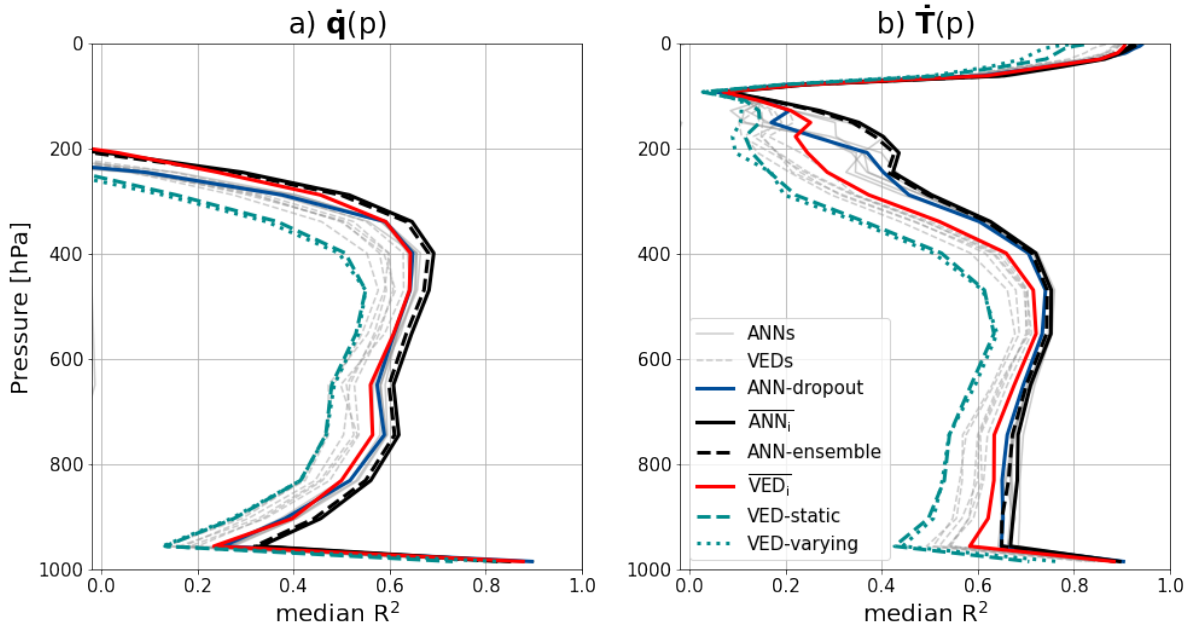

Figure 4.2.: Vertical profiles of median coefficient of determination R^2 for specific humidity tendency (a), $\dot{q}$), temperature tendency (b) $\dot{T}$) of ANN-dropout (solid navy blue); $\overline{\text{ANN}}$ and ANN-ensemble (solid and dashed black), $\overline{\text{VED}}$ (solid red); VED-static (dashed cyan) and VED-varying (dotted cyan line) and different individual ANNs and VEDs and (grey solid and dashed lines). The vertical profiles of median R^2 cloud liquid water tendency $\dot{q}_{cl}$ and cloud ice water tendency $\dot{q}_{ci}$ can be found in Figure B.1. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

Similar to our evaluation with R^2, we find an enhanced reproduction skill, as indicated by the mean absolute error (MAE), in both $\dot{q}_{ci}$ and $\dot{q}_{cl}$ with $\overline{\text{ANN}}$ and ANN-ensemble compared to a single ANNs (Figure B.4). For $\dot{T}$ and $\dot{q}$, the median MAEs of $\overline{\text{ANN}}$ and ANN-ensemble are either slightly higher or of similar magnitude like those of single ANNs in the troposphere (Figure B.4). A likely explanation is the overall skillful reproduction of $\dot{T}$ with individual ANNs (Figure 4.2 and B.4). Furthermore, one member of $\overline{\text{ANN}}$ and ANN-ensemble struggles to reproduce $\dot{T}$ and $\dot{q}$ with good skills, which decreases their performance. In contrast, we find an enhanced reproduction, based on MAEs, for $\dot{q}_{ci}$ above 400 hPa with $\overline{\text{ANN}}$ and ANN-ensemble in comparison to single ANNs (Figure B.4). For $\dot{q}_{cl}$ we see the same effect in the upper part of the planetary boundary layer (between 800 and 900 hPa, Figure B.4) for $\overline{\text{ANN}}$ and ANN-ensemble. In both cases, the enhanced performance of the ensemble methods are likely due to the larger contribution of stochasticity associated with turbulence in the planetary boundary layer and deep convection in the upper troposphere. Interestingly, however, $\overline{\text{VED}}$ and VED-draws have one of the best performance for $\dot{q}_{cl}$ and $\dot{q}_{ci}$ in the lower part of the planetary boundary layer compared to all other parameterizations (Figure B.4), despite their poorer reproduction skill for other vertical profiles. This suggests that the general characteristics of shallow convective processes can be well captured with the "quasi-deterministic" $\overline{\text{VED}}$ in comparison to ANNs. One possible explanation for this is the fact that VEDs are able to connect multiple convection related variables into a robust and more interpretable driver of convective predictability within their latent spaces (**Behrens** et al. 2022), e.g. forming a shallow convective

mode. As a consequence, we find a more realistic reproduction of $\dot{q}_{cl}$ (Figure B.4) and also surface radiative properties (Figures B.3 and B.5) with $\overline{\text{VED}}$ and VED-draws compared to ANNs that do not provide a lower dimensional representation of convective processes. ANN-dropout shows overall a good reproduction of convective processes with high R^2 and low MAE, but we will later show that this method strongly underestimates the uncertainty of convective processes. The opposite is true for VED-static and VED-varying, where we find the poorest performance in terms of R^2 and MAE. However we will see in the following that the uncertainty quantification via latent space perturbation reflects a well calibrated magnitude with respect to SPCESM2.

As a results of these conclusions I will evaluate in chapter 5 in detail the performance of the developed stochastic and deterministic deep learning parameterizations with respect to the quality of the respective uncertainty quantification. I will further use one metric that combines the reproduction skill with a term related to the spread of the ensembles. At a later stage I will couple the two best performing parameterizations based on this metric to CESM2 to investigate their stability in "hybrid" ESM simulations. But before that I will quickly summarize the main findings with respect to the reproduction skill of stochastic and deterministic deep learning parameterizations with respect to subgrid convective processes.

4.5. Summary Part I

This section is based on the first part of the concluding section in **Behrens** et al. 2024 and summarizes the key points from the general setup that I used to develop the stochastic and deterministic parameterizations and the evaluation of their reproduction capabilities of subgrid convective processes.

State-of-the-art deterministic deep learning algorithms based on one single model skilfully represent subgrid deep convective processes in climate models (Mooers et al. 2021; Rasp et al. 2018). However, reproducing the full complexity of convective processes, especially in the planetary boundary layer, remains challenging (**Behrens** et al. 2022; Gentine et al. 2018; Mooers et al. 2021). It has been speculated that this lower reproduction skill in the lower troposphere is largely related to the determinism of the used deep learning algorithms, neglecting the stochastic nature of convective processes (**Behrens** et al. 2022; Mooers et al. 2021). In this context, data-driven ensemble approaches that are scalable and can robustly overcome these issues would help improve ESMs.

This study presents and evaluates novel deep learning approaches to account for subgrid variability, due to stochasticity, to improve ESMs and projections. We demonstrate that the uncertainty and variability of such processes, as represented by the Superparameterized Community Earth System Model 2 (SPCESM2), can be correctly captured via an ensemble of predictions using Artificial Neural Networks (ANNs) or Variational Encoder Decoders (VEDs). This variability in unresolved convective processes is particularly relevant in the lower troposphere associated with turbulence and shallow convection, as well as in the upper

troposphere and lower stratosphere due to deep convection. There is, however, a trade-off between capturing the uncertainty of subgrid processes and their mean effect on the system, affecting the overall performance of the deep learned parameterization.

Randomly drawing an ensemble of predictions from different ANNs, ANN-ensemble (Table 4.1), enables skillful predictions as good as using the full deterministic ensemble of ANNs, $\overline{\text{ANN}}$ (Table 4.1). These two ensembles have an improved reproduction skill of convective processes compared to individual ANNs. We see the largest improvements for specific humidity and cloud liquid water tendencies within the planetary boundary layer. A similar improvement can be found for cloud ice water tendencies in the upper troposphere. An ANN with active dropout is not as accurate as the two ANN ensembles explored here. Perturbing the latent space of VEDs provides a good uncertainty range (see chapter 5) in their predictions, though accuracy in their predictions is substantially affected.

These findings show that ensemble deep learning parameterizations based on a number of networks with varying hyperparameters improve the presentation of subgrid convective processes on levels where these processes exhibit an elevated stochasticity related to subgrid turbulence like the planetary boundary layer or the upper troposphere due to deep convective activity compared to individual deterministic neural networks. In chapter 5 I will evaluate how robust the uncertainty quantification of the different stochastic and deterministic parameterizations is and how the two best of my developed parameterizations are performing in "hybrid model" simulations with CESM2.

5. Uncertainty quantification of ensemble deep learning parameterizations and hybrid simulations in an Earth System Model

This chapter is based on the second part of my paper that is currently in review (**Behrens** et al. 2024) with small modifications in the naming of distinct deep learning models and parameterizations. It is dedicated to a detailed evaluation of the uncertainty quantification of the different ensemble parameterizations and the results of hybrid simulations with them in an Earth System Model. It is structured as follows: Section 5.1 gives a quick overview over the used ensemble metrics and the technicalities in the background of the "hybrid" simulations, as both topics may potentially be unfamiliar to the reader and enhance the overall understanding of the thesis. Section 5.2 focuses on the uncertainty quantification of subgrid convective processes given by my developed parameterizations. Section 5.3 evaluates the deterministic and stochastic parameterizations with respect to the continuous rank probability score. Section 5.4 shows results of hybrid simulations with the best performing stochastic and deterministic ensemble coupled to Community Earth System Model (CESM). Section 5.5 summarizes the key findings of this chapter.

For this version of **Behrens** et al. 2024, that is currently in review, I, as the author of the thesis, contributed all figures, tables and large parts of the code to produce them. Furthermore I let the writing and the analysis of the published paper.

5.1. Ensemble Metrics and Online Coupling techniques

This section is directly reproduced from **Behrens** et al. 2024. It contains material presented in the methods section of my paper that is currently in review.

5.1.1. Ensemble Metrics

We evaluate the quality of the spread given by the different stochastic and deterministic approaches via uncertainty quantification with respect to the test data using three metrics. Specifically, we quantify the *aleatoric uncertainty* associated with the randomness aspect of the data-generation process, including the chaotic nature of convective processes in the atmosphere (Haynes et al. 2023). Firstly, we use the Continuous Rank Probability Score (CRPS), which is the difference between the MAE (first term) and the spread inside the ensemble (second term) in equation 5.1 (Haynes et al. 2023):

$$\mathrm{CRPS} = \frac{1}{\mathrm{N}_{ens}} \sum_{i=1}^{\mathrm{N}_{ens}} |\boldsymbol{Y}_i^{pred} - \boldsymbol{Y}| - \frac{1}{2\mathrm{N}_{ens}^2} \sum_{i=1}^{\mathrm{N}_{ens}} \sum_{l=1}^{\mathrm{N}_{ens}} |\boldsymbol{Y}_i^{pred} - \boldsymbol{Y}_l^{pred}| \tag{5.1}$$

CRPS is both sensitive to the deterministic quality of individual predictions $\boldsymbol{Y}_i^{pred}$ condensed in the MAE term and to the spread of the predictions inside the ensemble. This makes CRPS a suitable stochastic loss function for deep learning (Haynes et al. 2023). Moreover CRPS is a proper score (Gneiting and Raftery 2007) of negative orientation with a fixed lower bound of 0 (perfect skill) and upper bound 1 (no skill) in its probabilistic integral form.

Secondly, we use spread-skill diagrams to evaluate whether the skill of the stochastic and deterministic ensemble predictions (measured with the root mean square error (RMSE)) is correlated with the ensemble spread (Haynes et al. 2023). An ideal ensemble would have a pronounced correlation between spread and RMSE with a spread-skill ratio of one (Berner et al. 2017; Haynes et al. 2023). To sort the magnitude of the spread of the parameterizations for given $\boldsymbol{X}$ and selected output variables of interest $Y_{i,j}^{pred}$, we bin the spread into a number of classes $\mathrm{N}_{\mathrm{bins}}$ and compute the bin average for each class (equation 5.2). Then we calculate the conditionally averaged RMSE (equation 5.3) for each class $b \in [\![1, \mathrm{N}_{\mathrm{bins}}]\!]$:

$$\mathrm{Spread}_{j,b} = \frac{1}{\mathrm{N}_{\mathrm{counts},b}} \sum_{c=1}^{\mathrm{N}_{\mathrm{counts},b}} \sqrt{\frac{1}{\mathrm{N}_{ens}-1} \sum_{i=1}^{\mathrm{N}_{ens}} \left(\overline{Y_{c,j}^{pred}} - Y_{c,i,j}^{pred}\right)^2} \tag{5.2}$$

$$\mathrm{RMSE}_{j,b} = \sqrt{\frac{1}{\mathrm{N}_{\mathrm{counts},b}} \sum_{c=1}^{\mathrm{N}_{\mathrm{counts},b}} \left(Y_{c,j} - \overline{Y_{c,j}^{pred}}\right)^2}, \tag{5.3}$$

where $\mathrm{N}_{\mathrm{bins}}$ is the number of classes (bins) and $\mathrm{N}_{\mathrm{counts},b}$ represents the number of elements within a class $b \in [\![1, \mathrm{N}_{\mathrm{bins}}]\!]$.

Finally, we calculate the probability integral transform (PIT). This metric is similar to rank histograms, where the true value Y_j is ranked within the ensemble $\{Y_j^{pred}\}$ (i.e., the test data sample is situated between the (r-1)th and rth ensemble member and gets the rank r, where r is the rank ID). The PIT diagram is then obtained by computing the probability density function of all observed ranks $r \in [\![1, N_{ens}+1]\!]$ of Y_j (a probability value of each rank r; the y-axis) binned by the PIT values of each rank r (defined by the CDF of all ranks $N_{ens}+1$, x-axis). We use the PIT to evaluate whether the ensemble is overdispersive (which means that Y_j lies too frequently within the central percentiles of the ensemble) or underdispersive (Y_j is usually an outlier outside of the ensemble or lies in the lowest or highest percentiles of the ensemble). Ideally, the PIT curve is a horizontal line with an associated probability of $\frac{1}{\mathrm{N}_{ens}+1}$, which can be used to compute the PIT distance metric between the actual and ideal PIT case similar to the one shown in Haynes et al. 2023.

5.1.2. Online Coupling of the Ensemble Parameterizations

To couple our ensemble and stochastic parameterizations into CESM2 (replacing the SP component) we use the Fortran-Keras-Bridge (FKB) (Ott et al. 2020). To enforce the positivity of

precipitation and radiative fluxes as predictants, we add a "positivity layer" as a constraint layer (Beucler et al. 2021) to all DL models of the parameterizations. The "positivity layer" maps these variables with a ReLU activation to positive values.

We restrict our online coupling efforts to the deterministic and stochastic ensemble ANN parameterizations, which show superior offline performance compared to other developed parameterizations in section 4.4. First we transform the native weights and biases files into text files, which makes the files accessible for FKB and related Fortran compilers (Ott et al. 2020). Then we create a standalone repository that allows to couple individual ANNs, $\overline{\text{ANN}}$ and ANN-ensemble into CESM2. For ANN-ensemble we implement a random average function on grid cell level. In initial coupled experiments we find in some cases unrealistic simulated solar and shortwave radiative fluxes of more than 50 $\frac{W}{m^2}$ during night-time conditions on lower latitudes. To enhance the robustness of the online runs and the interpretability of simulated processes we enforce realistic radiative conditions for the coupling to the land and ocean surface by setting all solar fluxes and shortwave fluxes to zero $\frac{W}{m^2}$, if the cosine of the zenith angle of the incoming solar radiation in CESM2 at the current timestep and grid cell is zero (night-time conditions). Additionally, we implement a partial coupling scheme of our parameterization for certain variables, while other variables are simulated with the SP running aside. Our best performing setup that we present in section 5.4 relies on coupling all predicted variables from our parameterizations into CESM2 except for cloud ice water $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$ and cloud liquid water tendency $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$, which remain simulated by SP. Especially the partial coupling stabilizes online simulations, e.g., increasing the time until CESM2 crashes with our parameterization from the order of days or hours to more than five months (see section 5.4). For the online runs we use the predefined time stepping of SPCESM, with a native CESM2 timestep of 1800 s and an SP time step of 20 s. The subgrid source terms coming from SP and our parameterization are updated at every CESM2 time step. We perform CESM2 simulations based on initialisation files of January 2013 that included one month of SP spin-up, which is necessary for a realistic representation of global precipitation patterns. Our simulations start at the beginning of February. This coincides with the conditions that individual ANNs are optimized for during the training, as the respective data set contains the first seven days of each month of the year 2013. Nevertheless we test also three additional initialisation dates: the beginning of May, the beginning of August and the beginning of November, and investigate that the stability of our developed parameterizations is sensitive to the choice of the initialisation date.

As the next step, I will evaluate the quality of the uncertainty quantification of my developed stochastic and deterministic parameterizations with the in this section explained ensemble metrics.

5.2. Evaluating of Uncertainty Quantification

This section is directly based on the section with the identical name in my publication that is currently in review (**Behrens** et al. 2024).

Next, we evaluate the uncertainty quantification captured by the methods dealing with multiple predictions, meaning that prediction ranges from individual ensemble members will be assessed rather than their averages We focus on four vertical subregions with larger than average MAEs (Figure B.4): $\dot{\boldsymbol{q}}(\mathrm{p}_{surf})$; $\dot{\boldsymbol{T}}(\mathrm{p}_{surf})$; $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$(800-900 hPa); and $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$(200-400 hPa).

Figure 5.1 shows the spread-skill diagrams for these variables on specific pressure levels. An ideal spread-skill ratio of 1 is indicated by the grey dashed line (Berner et al. 2017). We randomly draw 500 timesteps from the test data set ($\sim 6.9 \times 10^6$ samples), and calculate the spread. Then we bin the spread arrays into 41 bins, based on the spread percentiles of VED-static, with bin width of 2.5^{th} percentiles. We finally calculate the conditional average of spread and RMSE for each bin (equation 5.3). The y-axis and x-axis represent the bin-averaged RMSE and spread, respectively. To put the magnitude of the shown maximum spread and RMSE values into perspective, their values are typically 10^2 to 10^3 larger than the MAEs (Figure B.4).

We find the best performance with respect to the spread-skill diagrams for VED-static followed by the ANN-ensemble and $\overline{\text{ANN}}$. As it is shown in Figure 5.1, for a spread smaller than 0.35 $\frac{g}{kg \times h}$ or $\frac{K}{h}$ for surface $\dot{\boldsymbol{q}}$ or $\dot{\boldsymbol{T}}$, these three parameterizations provide a considerably skillful uncertainty quantification. For larger spreads of surface $\dot{\boldsymbol{q}}$ and $\dot{\boldsymbol{T}}$ VED-static, DNN-ensemble and $\overline{\text{ANN}}$ illustrate an overdispersion. This means that the associated spread is larger than the RMSE and the respective spread-skill curves are situated below the ideal 1:1 ratio line. While for $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$ in the planetary boundary layer and $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$ in the upper troposphere, the underdispersion, when the spread is smaller than the RMSE, reduces with $\overline{\text{ANN}}$, ANN-ensemble and VED-static compared to all other developed parameterizations. This suggests an improved uncertainty calibration for these methods, which is also found in the respective PIT curves (Figure 5.2).

ANN-dropout and VED-draws (not shown) overall yield less well calibrated predictions, with larger deviations from the ideal 1:1 ratio, for the evaluated variables compared to all other developed parameterizations (Figure 5.1). Especially for $\dot{\boldsymbol{q}}$ at the surface, $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$ in the upper part of the planetary boundary layer and $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$ in the upper troposphere, we find a strong underdispersion with ANN-dropout and VED-draws. The pronounced underdispersion of ANN-dropout and VED-draws is also present in the associated probability integral transform (PIT) diagrams for $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$ in the planetary boundary layer (Figure 5.2). The ideal PIT curve is shown as the thick dashed grey line. For ANN-dropout, and VED-draws, almost all test data samples are situated in the tails of the distribution of the PIT curve. The same behaviour with too many outliers, situated in the tails of the distribution, is further visible for the PIT histograms of $\dot{\boldsymbol{q}}$ at the surface, $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$ in the upper troposphere or $\dot{\boldsymbol{T}}$ at the surface (Figures B.10 to B.12). In combination with the overall poor skill in the spread-skill diagrams (Figure 5.1), except for $\dot{\boldsymbol{T}}$ for ANN-dropout, this suggests that ANN-dropout and VED-draws yield uncertainty quantification that underestimates the variability in the test data for the

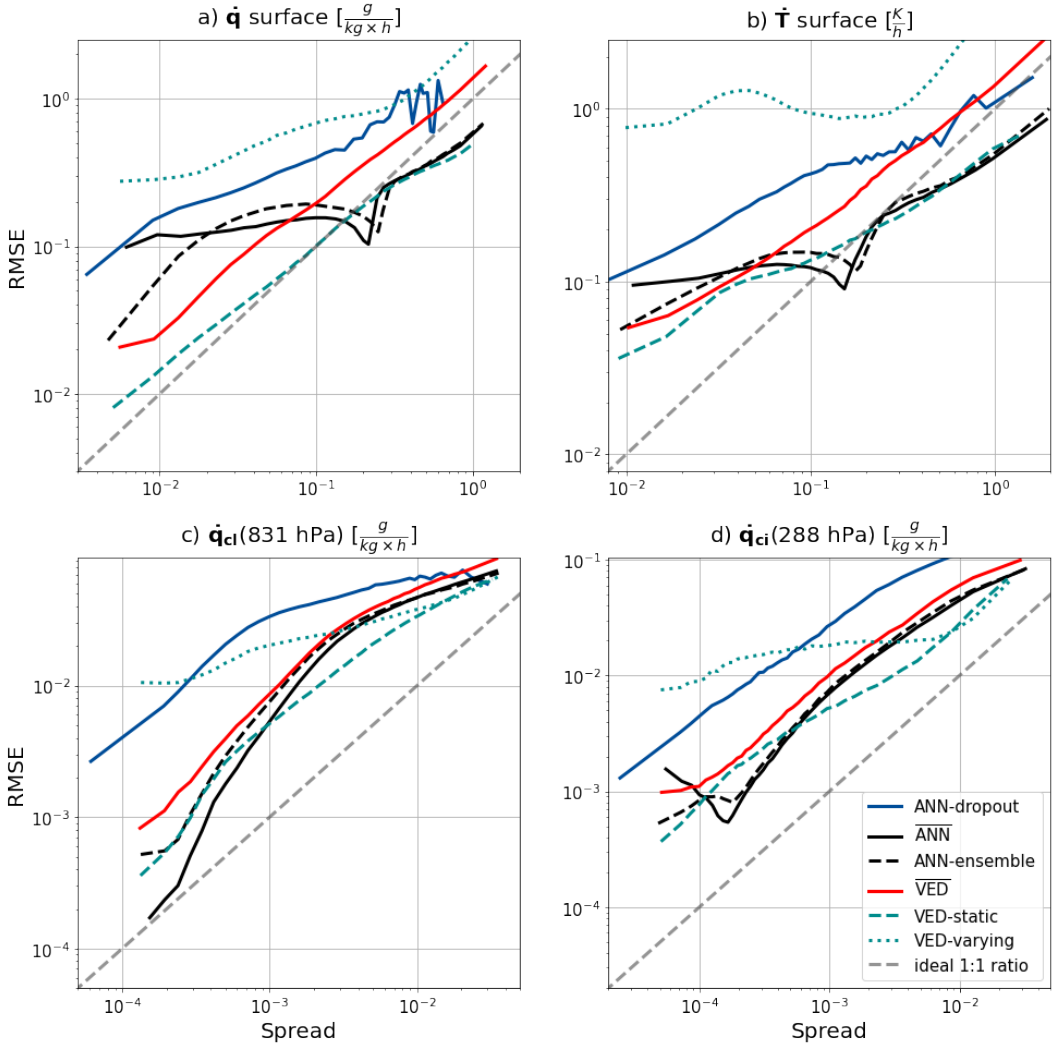

Figure 5.1.: Spread-Skill diagram between bin-averaged spread (x-axis) and Root Mean Square Error (RMSE, y-axis) based on the test data and predictions over 500 randomly drawn timesteps. Shown is the spread-skill diagram of surface specific humidity tendency $\dot{q}$ in a), surface temperature tendency $\dot{T}$ in b), cloud water tendency $\dot{q}_{cl}$ in the upper planetary boundary layer on 831 hPa in c) and cloud ice tendency $\dot{q}_{ci}$ in the upper troposphere on 288 hPa in d). The color-coding of the ensemble and stochastic parameterizations is identical to Fig. 2. Additionally we include the spread-skill ratio of 1:1 (dashed grey line) that symbolises the optimal calibration of the spread vs. skill based on literature (Berner et al. 2017; Haynes et al. 2023). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

evaluated variables on pressure levels where a large portion of convective processes is driven by turbulence. In the following, we will show how this translates into an elevated CRPS and poor skill for ANN-dropout and VED-draws.

Similarly, $\overline{\text{VED}}$ tends to be underdispersive for all evaluated variables (Figure 5.1), but with an improved spread-skill compared to ANN-dropout and VED-draws. Also we find that $\overline{\text{VED}}$ is competitive against all other parameterizations for smaller spread values (Figure 5.1). For $\dot{q}_{cl}$ in the planetary boundary layer, the probability that the SPCESM2 sample is situated within the ensemble slightly increases (Figure 5.2). However, it should be noted that

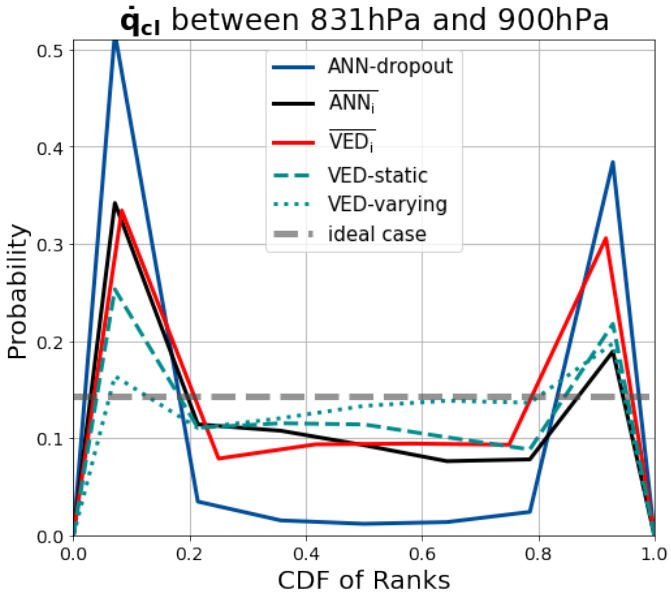

Figure 5.2.: Probability Integral Transform (PIT) histogram of $\dot{q}_{cl}$ in the planetary boundary layer between 831 and 900 hPa. The x-axis represents the Cumulative Distribution Function (CDF) of the ranks of the test sample from Super Parameterized Earth System Model (SPCESM) version 2 with respect to the number of ensemble members of the stochastic or ensemble parameterizations. The y-axis depicts the probability associated with each rank. The PIT histogram is based on 400 randomly drawn timesteps from the test data set. The thick dashed gray line in the subplot in horizontal direction symbolises the ideal shape of the PIT curve. The color coding is identical to Figure 4.2. Note that we exclude the curve of the stochastic ANN-ensemble due to the fact that it shares the majority of ensemble members with $\overline{\text{ANN}}$ and results in a similar PIT curve with fewer ranks. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

$\overline{\text{VED}}$ has one fewer member than the other parameterizations. The same is true for the PIT curves of $\overline{\text{VED}}$ with respect to $\dot{T}$, $\dot{q}$ at the surface or $\dot{q}_{ci}$ in the upper troposphere, suggesting that $\overline{\text{VED}}$ provides better calibrated uncertainty quantification compared to VED-draws or ANN-dropout. In the following the CRPS evaluation will further support this reasoning.

The spread-skill analysis reveals substantial differences in the quality of the uncertainty quantification between a latent space perturbation with isotropic Gaussian noise (VED-static) and anisotropic Gaussian noise (VED-varying). While VED-static is one of the best performing ensemble methods, we find a pronounced underdispersion for VED-varying for $\dot{q}$ and $\dot{T}$ at the surface (Figure 5.1), more so than for the ANN-dropout. This result of the spread-skill analysis is particularly interesting as it suggests that a latent space perturbation with an anisotropic Gaussian noise term (VED-varying) yields a decreased calibration of the uncertainty quantification of the surface moistening and heating compared to an isotropic Gaussian noise term (VED-static). However for $\dot{q}_{cl}$ in the planetary boundary layer and $\dot{q}_{ci}$ in the upper troposphere both VED-varying and VED-static show an improved calibration of the ensemble spread compared to all other developed ensemble and stochastic parameterizations for a bin averaged spread larger than 0.005 $\frac{g}{kg \times h}$. VED-varying shows a weaker prediction skill compared to VED-static for $\dot{q}_{cl}$ and $\dot{q}_{ci}$ for a spread smaller than 3 $\times 10^{-4}$ $\frac{g}{kg \times h}$ (Figure 5.1).

This results in an increased underdispersion of VED-varying compared to ANN-dropout for a spread smaller than $2 \times 10^{-4} \frac{g}{kg \times h}$. We could then cross-link the results from the spread-skill diagrams of VED-static and VED-varying with the respective PIT histograms (Figure 5.2 and Figures B.10 to B.12). Figure 5.2 shows that VED-static and VED-varying have strongly reduced outliers in their respective PIT histograms for $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$ in the planetary boundary layer. The calibration of the uncertainties for VED-varying is slightly improved compared to the one of VED-static. The probability that the true SPCESM2 sample is ranked at the outer edge of the PIT curves decreases for VED-varying, while the probabilities for the inner ranks for VED-varying is converging towards the ideal case for $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$ in the planetary boundary layer (Figure 5.2). The same improved quality of uncertainty quantification is also present for $\dot{\boldsymbol{q}}$, $\dot{\boldsymbol{T}}$ at the surface and $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$ in the upper troposphere with VED-varying compared to VED-static (Figures B.10 to B.12). However, we recall it came at the cost of worse predictive skill of convective processes (Figures 4.2, 5.1)

Overall we find that VED-static has the best uncertainty quantification based on the PIT curves and the spread-skill diagrams, followed by $\overline{\text{ANN}}$ and ANN-ensemble with a good calibration of the ensemble spread. These networks often indicate only a slight underdispersion or overdispersion compared to the ideal PIT curve. VED-varying provides calibrated uncertainty quantification in the PIT analysis but to the expense of a lower reproduction skill as can be seen in its relatively large RMSE for $\dot{\boldsymbol{q}}$, $\dot{\boldsymbol{T}}$ at the surface and condensate tendencies (Figure 5.1). Also $\overline{\text{VED}}$ represents the uncertainty of convective processes well. The uncertainty quantification of ANN-dropout and VED-draws is in general not well calibrated. Additionally, the PIT curves of ANN-dropout and VED-draws show the strongest underdispersion with most of the true SPCESM2 samples being sorted in the lowest or highest rank as outliers. This means that these two parameterizations strongly underestimate the simulated spread of key variables in SPCESM2, and could not represent variations in convective processes like all other parameterizations.

5.3. Proper Scoring

This section is directly based on the section with the identical name in the my publication that is currently in review (**Behrens** et al. 2024).

Here we provide a holistic evaluation of both the calibration of the ensemble spread and the quality of the reproduction error metrics (see Section 5.1.1). $\overline{\text{ANN}}$ and ANN-ensemble are the best-performing deterministic and stochastic parameterization based on CRPS (Figure 5.3). We start our CRPS analysis by focusing first on general statistics of CRPS calculated over all output variables $\boldsymbol{Y}$. Figure B.6 shows the mean, median, the 75^{th} and 90^{th} percentile of CRPS computed over all SP variables $\boldsymbol{Y}$. We find the lowest mean and median CRPS for $\overline{\text{ANN}}$ and ANN-ensemble over all subgrid SP variables $\boldsymbol{Y}$. This indicates that these two parameterizations are the best compromise between predictive skill on one side and uncertainty quantification on the other side. While $\overline{\text{VED}}$ and ANN-dropout perform considerably well,

VED-draws shows intermediate performance based on the mean and higher percentiles of CRPS calculated over $\boldsymbol{Y}$. Both VED-static and VED-varying have remarkably increased 75^{th} and 90^{th} percentiles compared to all other parameterizations (Figure B.6). However, we note that the respective median CRPS values decrease compared to VED-draws, which underscores that the latent space perturbation has the potential to improve the uncertainty quantification of convective processes.

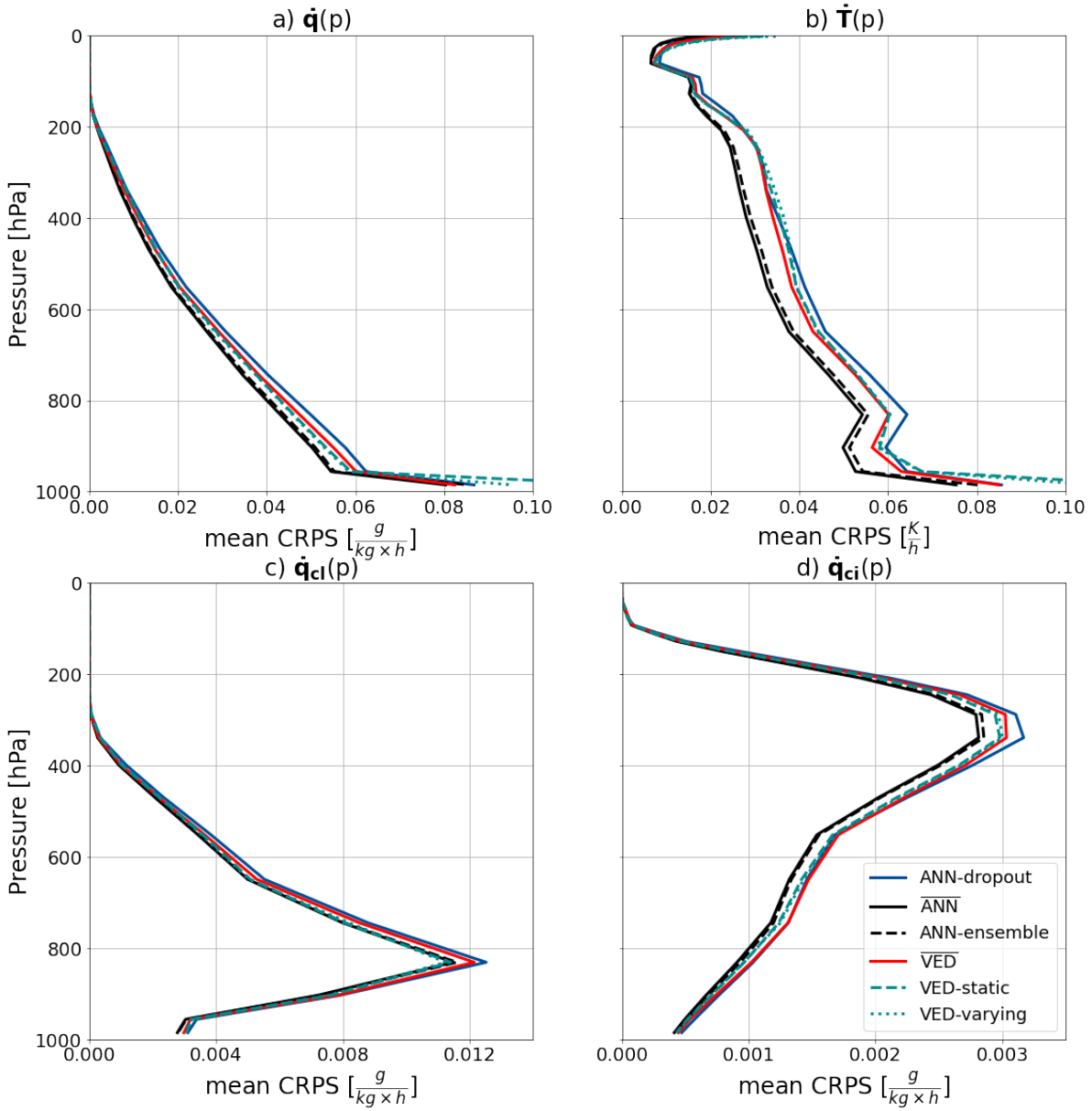

Figure 5.3.: Mean Continuous Rank Probability Score (CRPS) of the $\dot{\boldsymbol{q}}$ (a), $\dot{\boldsymbol{T}}$ (b), $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$ (c), $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$ (d) for the different ensembles based on 500 randomly drawn time step from the test data. The blue line indicates ANN-dropout. The solid and dashed black lines represent the deterministic $\overline{\text{ANN}}$ and stochastic ANN-ensemble parameterization alongside $\overline{\text{VED}}$ (red line). The dashed and dotted cyan lines represent VED-static and VED-varying. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

We extend our CRPS analysis to evaluate from which SP variables the differences between the parameterizations are arising. Figure 5.3 shows the vertical profiles of mean CRPS values for $\dot{q}(p)$, $\dot{T}(p)$, $\dot{q}_{cl}(p)$ and $\dot{q}_{ci}(p)$.

The similar performance of $\overline{\text{ANN}}$ and ANN-ensemble suggests that the latter does not exhibit a decline in reproduction skill of convective processes, as found with all other developed stochastic parameterizations, particularly in the upper planetary boundary layer and the upper troposphere. VED-static and VED-varying have a compatible performance to $\overline{\text{ANN}}$ and ANN-ensemble in the upper part of the planetary boundary layer for $\dot{q}_{cl}$ and in general a good skill for all vertical profiles (Figure 5.3). However VED-static and VED-varying have an elevated CRPS for surface $\dot{q}$ and $\dot{T}$ compared to other deterministic or stochastic ensemble parameterizations. These results suggest that the latent space perturbation yields well calibrated uncertainty quantification for convective processes in the troposphere like we saw already with the analysed uncertainty metrics. The shortcomings of VED-static and VED-varying on the surface levels mainly arise from the reduced reproduction skill with latent space perturbation while the calibration of the uncertainty quantification depicts high skill (Figures B.11 and B.12). $\overline{\text{VED}}$ shows in general a compatible performance in CRPS with intermediate scores, while ANN-dropout depicts the highest CRPS of all evaluated parameterizations for the vertical profiles of $\dot{q}$, $\dot{T}$, $\dot{q}_{cl}$, $\dot{q}_{ci}$ due to the shortcomings in the calibration of the ensemble spreads (Figures 5.1 and 5.3). The same shortcomings are visible for VED-draws.

Figure 5.4 shows the global map of the mean CRPS values of $\dot{q}_{ci}$ on 288 hPa for $\overline{\text{ANN}}$ based on 500 randomly drawn timesteps from the test data set. Moreover it depicts the differences of mean CRPS of all other developed parameterizations with respect to $\overline{\text{ANN}}$, excluding VED-draws due its overall weak performance in CRPS. In the supporting information similar maps for $\dot{q}_{cl}$, surface $\dot{q}$ and $\dot{T}$ (Figures B.7 to B.9) can be found. The CRPS structure shows the imprint of the atmospheric general circulation centers of action. In general, we find the largest mean CRPS, a decline in performance, associated with deep convective systems over the Maritime Continent, the tropical East Pacific offshore of Panama, the Congo basin, and the Amazonian and Parana regions. Especially over these regions $\overline{\text{ANN}}$ and ANN-ensemble have the best performance with respect to CRPS compared to the other parameterizations (Figure 5.4). $\overline{\text{ANN}}$ and ANN-ensemble have also the lowest global mean value of CRPS of $\dot{q}_{ci}$ on 288 hPa with 2.8×10^{-3} $\frac{g}{kg\times h}$, while the other parameterizations have a mean value larger than 3×10^{-3} $\frac{g}{kg\times h}$ except of VED-static. ANN-dropout has in general elevated CRPS over the deep convective regions for $\dot{q}_{ci}$ compared to the other developed parameterizations. As we already investigated, ANN-dropout is strongly underdispersive (Figure B.10) and does not provide robust uncertainty quantification for $\dot{q}_{ci}$ in the upper troposphere. In contrast, the VED-static and VED-varying parameterization yield the best calibration of the ensemble spread for the upper tropospheric $\dot{q}_{ci}$ (Figure B.10), which explains also the clear improvement of the mean CRPS visible compared to ANN-dropout.

For surface $\dot{q}$ and $\dot{T}$ or $\dot{q}_{cl}$ in the upper planetary boundary layer $\overline{\text{ANN}}$ and ANN-ensemble have the best performance compared to other parameterizations based on CRPS (Figures B.7 to B.12). The largest improvements with $\overline{\text{ANN}}$ and ANN-ensemble compared to the other

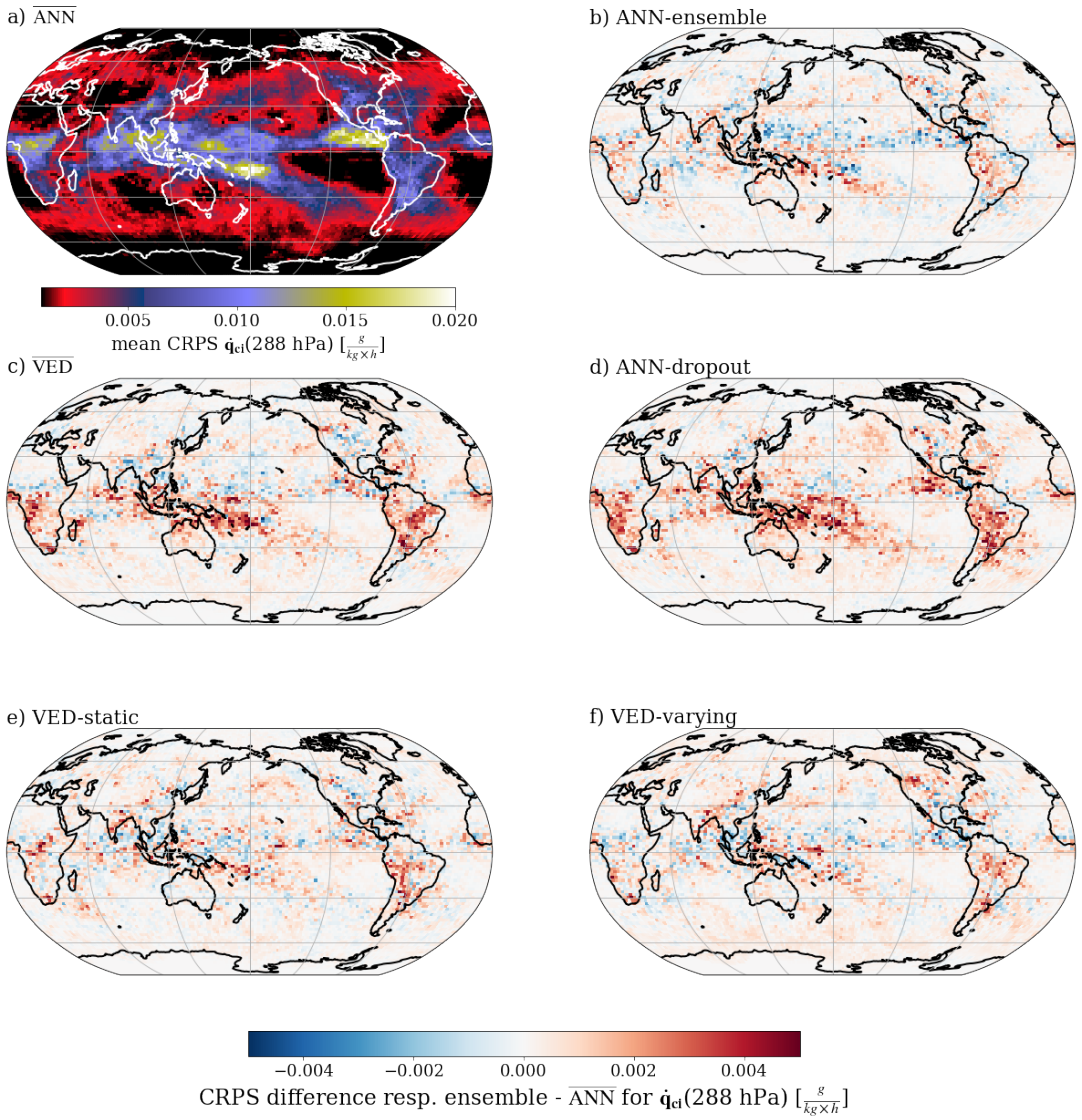

Figure 5.4.: Mean Continuous Rank Probability Score (CRPS) of $\dot{q}_{ci}$ on 288 hPa based on 500 randomly drawn timesteps from the test data set for (a) the deterministic $\overline{\text{ANN}}$ ensemble, the CRPS differences of (b) the stochastic ANN-ensemble, (c) the deterministic $\overline{\text{VED}}$, (d) ANN-dropout; VED-draws (e),VED-static (f), VED-varying (g) parameterizations to $\overline{\text{ANN}}$. This Figure is reproduced with minor modifications from **Behrens** et al. 2024

parameterizations are in the tropics in the regions with the highest CRPS for all evaluated variables. In contrast, the extra-tropical and especially regions with negligible deep convective activity, i.e., the upwelling regions offshore of the west coast of the Americas or Africa, are characterized by similar small CRPS across all parameterizations, as expected. In agreement with previous results, ANN-dropout often has elevated CRPS. For VED-static and VED-varying we find an improvement in CPRS compared to $\overline{\text{ANN}}$ and ANN-ensemble for $\dot{q}_{cl}$ on 831 hPa, but the largest CRPS for surface $\dot{q}$ and surface $\dot{T}$ as already expected from Figure 5.3.

In general, the latent space perturbation leads to an improvement in the calibration of the ensemble spread compared, for example, to ANN-dropout. Nevertheless, our CRPS and the

PIT analysis reveal that there is a trade-off between robust uncertainty quantification on one hand and reproduction skill on the other hand. Therefore we designed a hyperparameter tuning method to balance these two important factors for the development of a stochastic convection parameterization with latent space perturbation (see in Appendix B section B.5).

ANN-ensemble and $\overline{\text{ANN}}$ do not need such additional tuning steps and show a similarly good calibration of the uncertainty quantification of convective processes in combination with enhanced reproduction skill of convective processes compared to all other developed parameterizations.

In the next section, we will therefore focus on $\overline{\text{ANN}}$ and ANN-ensemble parameterizations coupled to CESM2, demonstrating the advantages of such parameterizations compared to single ANN parameterizations.

5.4. Online Results: Improved Stability and Precipitation Distributions

This section is directly based on the respective section in **Behrens** et al. 2024.

In this section, we first describe the challenges of coupling our parameterizations to CESM2. Second, we evaluate our prognostic runs against the high-resolution SPCESM2 model, the coarse CESM2 model with a traditional convection scheme, as well as a deep learning benchmark (Han et al. 2023).

5.4.1. Online Coupling Challenges

We couple $\overline{\text{ANN}}$ and ANN-ensemble, the two best-performing deterministic and stochastic parameterizations, into CESM2 using the Fortran-Keras-Bridge (Ott et al. 2020), resulting in $\overline{\text{ANN}}$-CESM and ANN-ens-CESM hybrid models. We follow the configuration detailed in Section 5.1.2 for our new hybrid model runs. Coupling the complete set of $\boldsymbol{Y}^{pred}$ to CESM2 led to unstable prognostic runs after few days. Note that running the hybrid model with individual ANN led to instabilities in only few time steps. This shows the stabilizing effect of ensembles consistent with Brenowitz et al. 2020. We identified one particular ANN with low performance of the parameterizations and retrained it. This allowed us to achieve few weeks long prognostic runs. While the stability of the prognostic runs depends to some extent on the initial conditions, the primary cause of the hybrid model instabilities were ice growth in the lower tropical stratosphere and subsequent radiative feedback. These anomalous signals manifested in rapidly increasing $\boldsymbol{q_{ci}}$ in the stratosphere, which led to unrealistic values of $\boldsymbol{Y}_{\boldsymbol{rad}}^{\boldsymbol{pred}}$ that are crucial for the coupling with the surface model components (e.g. land model), ultimately causing blow-ups of CESM2 with our ensemble and stochastic parameterizations.

Achieving a stable hybrid multi scale model is indeed a non trivial task (Yu et al. 2023). The deep learned representation of condensate tendencies is particularly challenging. To overcome this challenge, we performed a "perfect condensate" experiment, in which $\dot{\boldsymbol{q}}_{\boldsymbol{ci}}$ and $\dot{\boldsymbol{q}}_{\boldsymbol{cl}}$

variables are simulated by the SP component and the rest by our deep learned parameterizations. This partially-coupled setup, however, requires running the SP component alongside the predictions from the neural networks, with a clear drawback in terms of computational efficiency. Nevertheless, this configuration allowed us to achieve six months long stable hybrid runs for both, $\overline{\text{ANN}}$-CESM and ANN-ens-CESM. Specifically, $\overline{\text{ANN}}$-CESM ran stably from the beginning of February 2013 to the last third of July, and ANN-ens-CESM stopped at the beginning of July. Running the hybrid model with the "perfect condensate" setup but for individual ANNs, crashed in 6 out of 7 cases within the first five days of the simulation (see Figure B.16 and B.17). The ANN with the largest RMSE due to imperfect predictions representing average conditions (e.g. predicting constant drizzle conditions in all horizontal grid cells) survived until mid October. This suggests that model stability and the robustness or realism of the predicted convective and radiative fluxes are not associated with each other (Lin et al. 2023). Omitting the spurious ANN as an ensemble member destabilized $\overline{\text{ANN}}$-CESM and ANN-ens-CESM in test runs. Furthermore, we found that using fewer ensemble members (number of neural networks and larger stochasticity) for ANN-ens-CESM strongly affected its stability. This suggests that deep-learned ensemble parameterizations may require a trade-off between computational efficiency and the number of members.

5.4.2. Online Performance

We evaluate $\overline{\text{ANN}}$-CESM and ANN-ens-CESM prognostic runs over the period from February to June 2013 (before blow-up in mid-July), covering a total of 7200 timesteps. These simulations are evaluated against the original high-resolution SPCESM2 (abbreviated as SP-CESM), and against the coarse CESM2 (abbreviated as ZM-CESM) with the traditional convection parameterization (Zhang and McFarlane 1995), over the same period. We note that all coarse model runs, $\overline{\text{ANN}}$-CESM, ANN-ens-CESM and ZM-CESM, are based on one-month spin-up, while SP-CESM is based on a decade-long model run. Figure 5.5 shows zonal averages of the median precipitation (Figure 5.5a), as well as zonal averages of higher percentiles (Figure 5.5b,c). To investigate the influence of the internal variability on the zonal structures of the respective curves, we add uncertainty ranges based on 50 bootstrapped subsamples of 2000 random time steps ($\sim$ 41 days). Additionally, we show the precipitation probability distribution accumulated across all grid cells and timesteps and binned as a function of the baseline precipitation distribution simulated with SP-CESM (Fig. 5.5d).

$\overline{\text{ANN}}$-CESM and ANN-ens-CESM clearly outperform ZM-CESM reproducing not only median precipitation (Figure 5.5a), but also extreme rainfall (Figure 5.5d). Furthermore, they alleviate known overestimations of intermediate precipitation ($0.08\ \frac{mm}{h} < \text{Prec} < 0.3\ \frac{mm}{h}$) in coarse ESMs, such as ZM-CESM, compared to SP-CESM. These findings are in agreement with previous results with an idealized setup (Rasp et al. 2018). ANN-ens-CESM shows in general weaker reproduction of precipitation extremes compared to $\overline{\text{ANN}}$-CESM (Figure 5.5b,c), however this is not seen in the accumulated precipitation probability distribution (Figure 5.5d), which may be due to compensating errors in the large-scale thermodynamic fields (Figure

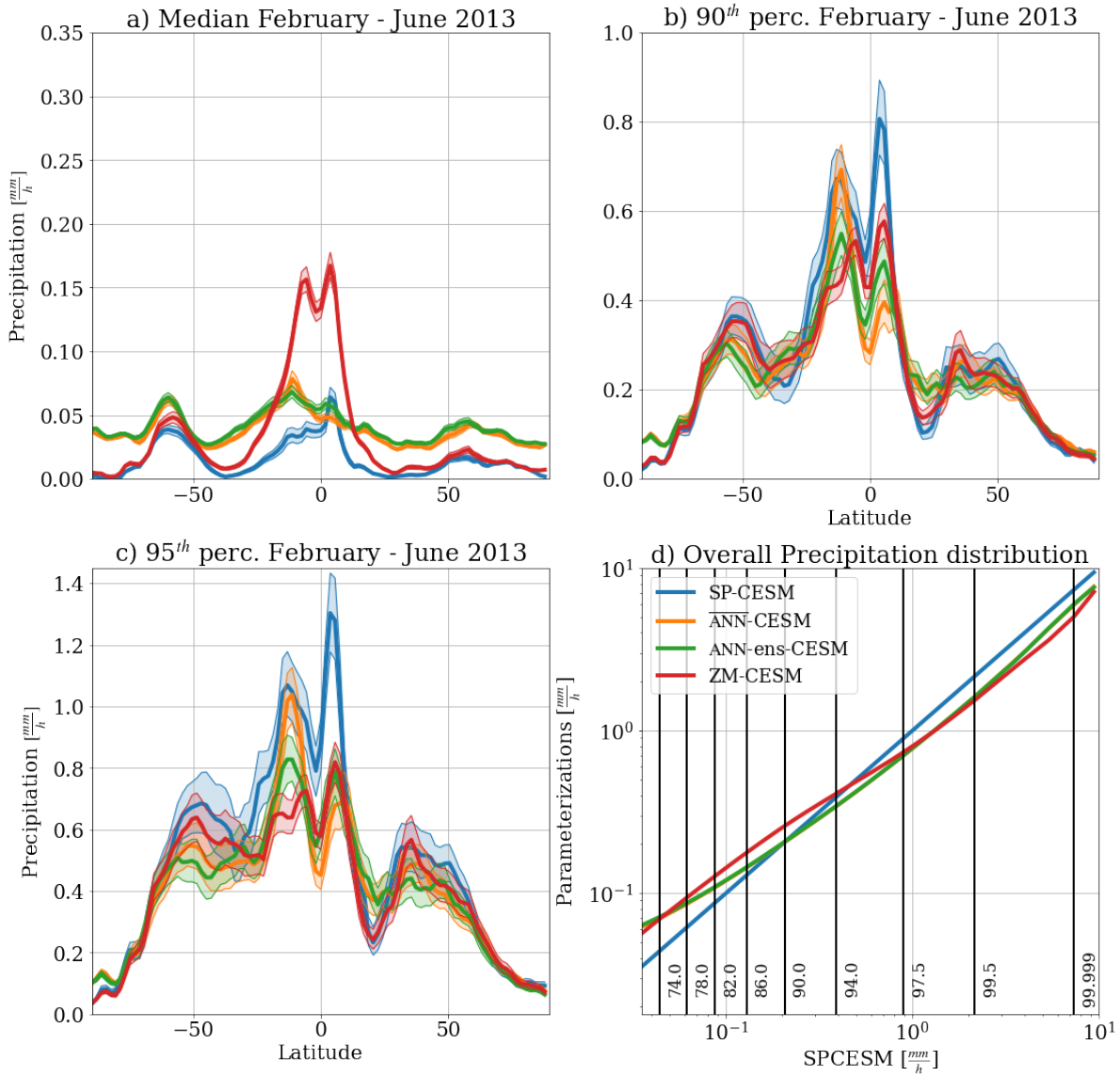

Figure 5.5.: Simulated zonal averages of median (a), 90th (b) and 95th percentiles (c) of total precipitation in the period February to June of CESM2 with a superparameterization (SP-CESM, blue), CESM2 coupled to the deterministic $\overline{\text{ANN}}$ parameterization ($\overline{\text{ANN}}$-CESM, orange), CESM2 coupled to the stochastic ANN-ensemble parameterization (ANN-ens-CESM, green) and CESM2 with the traditional Zhang-McFarlane scheme (ZM-CESM, red line). The uncertainty ranges indicate the span between minimum and maximum of the given metrics based on bootstrapping with 50 subsamples. Subplot d) shows the precipitation distribution of the different parameterizations (y-axis) as a function of the precipitation distribution simulated with the superparameterization (x-axis). The vertical lines in subplot d) represent distinct percentiles of the precipitation distribution in SPCESM. For subplot d) the entire simulated precipitation rates in all grid cells and all timesteps of the period February 2013 to the end of June 2013 are used. This plot is reproduced with minor modifications from **Behrens** et al. 2024.

B.18, B.19). $\overline{\text{ANN}}$-CESM and ANN-ens-CESM show a positive offset for small precipitation rates compared to SP-CESM (Figure 5.5d). Although we find positive precipitation biases in all models compared to SP-CESM in the tropics, near the ITCZ and along the midlatitude storm tracks ($\sim 40°$ to $60°$ N and S), these are alleviated to a good extent in the hybrid models. For

example, ZM-CESM strongly overestimates median precipitation over the tropical equatorial Pacific compared to SP-CESM (Figure B.21). Contrary, hybrid models show general median precipitation patterns that are more in agreement with SP-CESM in the tropics (Figure B.21). Nevertheless, $\overline{\text{ANN}}$-CESM and ANN-ens-CESM are not capturing the exact location of the ITCZ as simulated by SP-CESM at around 5° N (Figure 5.5a-c). The weaker representation of the first precipitation maximum at 5° N by the hybrid models compared to SP-CESM may be associated with a less developed ITCZ over the northern equatorial Pacific Ocean due to biases in the large-scale thermodynamic conditions (Figure B.19). We find that $\overline{\text{ANN}}$-CESM represents very well precipitation extremes at the second tropical precipitation maximum (10° S) compared to SP-CESM (Figure 5.5 b,c). This is, however, in general underestimated by ANN-ens-CESM and ZM-CESM (Figure B.20). Both $\overline{\text{ANN}}$-CESM and ANN-ens-CESM overestimate median precipitation at high-latitudes compared to SP-CESM (Figure 5.5a), partly due to one ANN with low performance (not shown).

To gain further insights about the biases highlighted above, we assess the mean state of large-scale thermodynamic fields. We find a pronounced warm bias in the stratosphere of up to 20 K in $\overline{\text{ANN}}$-CESM and ANN-ens-CESM compared to SP-CESM (Figure B.18). Furthermore, we find more than 10 K warmer conditions in near surface levels over Antarctica. These biases in ZM-CESM are considerably smaller in these two regions. With regard to $\boldsymbol{q}\,\boldsymbol{(p)}$, both hybrid models show drier conditions over the ITCZ region (Figure B.19), which partly explains the weaker precipitation patterns around 5° N compared to SP-CESM (Figures 5.5 and B.21). Moreover, the hybrid models show a moist bias in the subtropics above 800 hPa (Figure B.19). The corresponding biases of the specific humidity field in ZM-CESM are slightly weaker, however, but indicate comparable deficiencies in the simulation of the specific humidity field in the tropics.

Figure 5.6 shows global maps of diurnal precipitation peaks in Local Solar Time (LST). $\overline{\text{ANN}}$-CESM and ANN-ens-CESM yield an improved diurnal precipitation cycle compared to ZM-CESM, that is more in agreement with SP-CESM. Interestingly, both hybrid models capture the afternoon peak of precipitation over the Amazonian region (Figure B.23a), the Congo basin (Figure B.23b), and Europe (Figure B.23c), as represented in SP-CESM. In contrast, ZM-CESM simulates a peak precipitation around noon over these distinct regions. Furthermore, $\overline{\text{ANN}}$-CESM and ANN-ens-CESM show a too strong diurnal cycle over North Africa and the Arabian peninsula (Figure 5.6), which is related to a wet bias on the order of 0.015 $\frac{mm}{h}$ with respect to SP-CESM. We find the opposite case, a less pronounced diurnal cycle over maritime stratocumuli regions offshore of California, Peru and Angola in $\overline{\text{ANN}}$-CESM and ANN-ens-CESM compared to SP-CESM. Our results are in agreement with previous offline findings (Mooers et al. 2021), demonstrating the improvement of reproducing the diurnal precipitation cycle with deep learning schemes compared to conventional convection parameterizations.

Finally, we place our findings in context by comparing them against Han et al. 2023. The authors used deep convolutional residual neural networks to represent heating and moistening tendencies, as well as cloud liquid and ice water in the Community Atmosphere Model version 5 (CAM5) with real geography (Han et al. 2023). Moreover they successfully coupled one

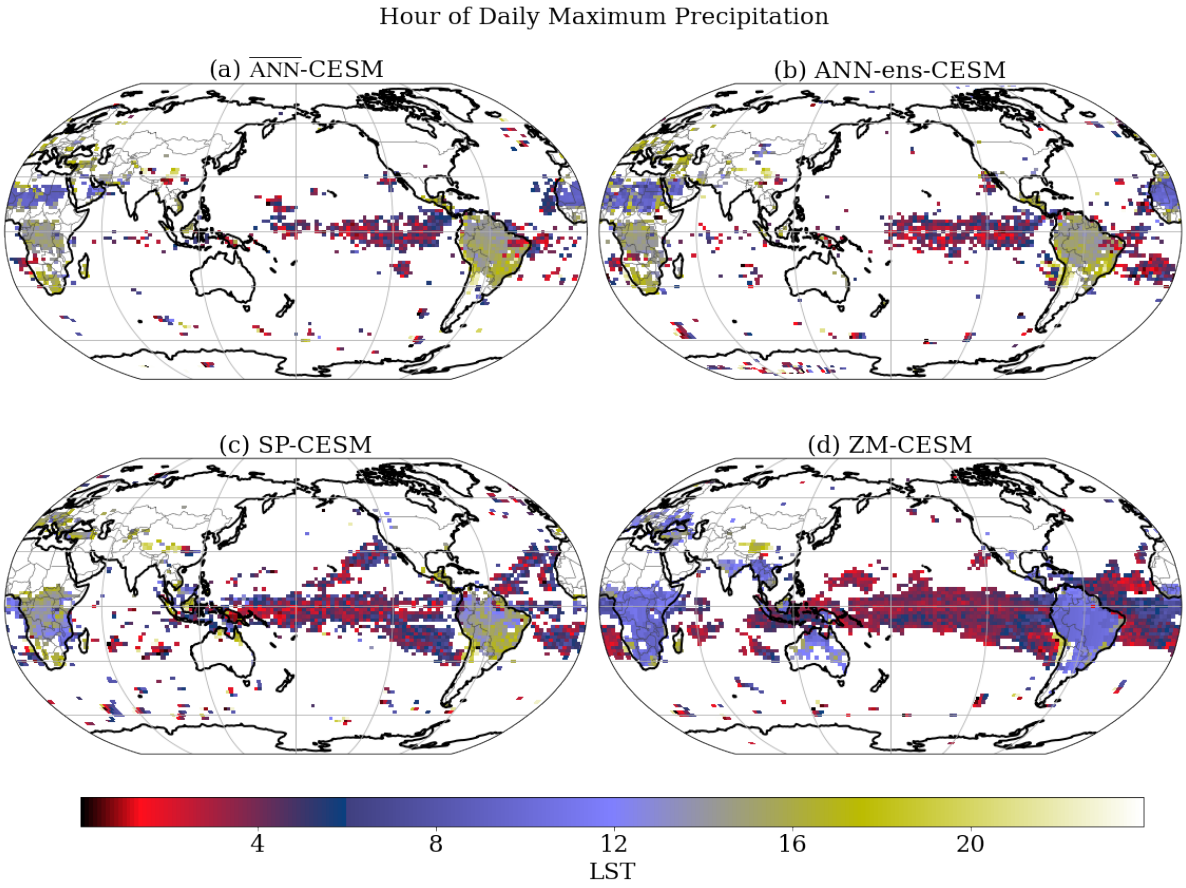

Figure 5.6.: Global Maps of the Hour of the Daily Maximum Precipitation in the CESM2 runs with the deterministic $\overline{\text{ANN}}$-CESM parameterization (panel a), the stochastic DNN-ens-CESM parameterization (b), the superparametrization SP-CESM (c) and the traditional Zhang-McFarlane scheme (d) analysed for the period from February to June 2013. The color-coding reveals the diurnal peak in precipitation in local solar time (LST) in areas with a pronounced diurnal cycle of precipitation with a magnitude over a certain threshold, similar to the one used in Mooers et al. 2021. This plot is reproduced with minor modifications from **Behrens** et al. 2024.

ensemble member to CAM5 and conducted a stable 5 year run with it. $\overline{\text{ANN}}$-CESM and ANN-ens-CESM show a considerably weaker ITCZ compared to Han et al. 2023. This might be related to larger biases in the large-scale specific humidity fields in this work, especially in the tropics (Figure B.19), compared to Han et al. 2023. However, our hybrid models tend to have a reduced temperature bias in the troposphere (Figure B.18) compared to Han et al. 2023, though this may be associated with their longer prognostic runs of 5 years. We also note that Han et al. 2023 sidestepped deep learning surface radiative fluxes (not coupled to the land component), whereas in our study, it is explicitly implemented and may well affect the stability of the hybrid models presented here.

In summary, $\overline{\text{ANN}}$ and ANN-ensemble have an enhanced stability compared to individual ANNs. Furthermore both ensemble parameterizations capture precipitation extremes and the underlying diurnal cycle better than existing convection schemes – despite the fact that there

are important distortions of the mean state rainfall compared to the original superparameterization related to biases in the large-scale thermodynamic fields.

The next section will summarize the main findings from our extensive offline evaluation with ensemble metrics and from our online test of the ensemble parameterizations coupled to CESM2.

5.5. Summary Part II

This section is directly based on the second part of the conclusion of **Behrens** et al. 2024.

As it was shown in Section 4.5 there are differences in the reproduction skill of the developed deterministic and stochastic ensemble deep learning parameterizations.

There is, however, a trade-off between capturing the uncertainty of subgrid processes and their mean effect on the system, affecting the overall performance of the deep learned parameterization. An ANN with active dropout neither fully captures the variability of unresolved processes nor is it as accurate as other deep learning algorithms explored here. Perturbing the latent space of VEDs provides a good uncertainty range in their predictions, though accuracy in their predictions is substantially affected. Randomly drawing an ensemble of predictions from different ANNs, ANN-ensemble (Table 4.1), enables us to achieve both a well-calibrated uncertainty compared to the superparameterized ESM and skillful predictions as good as using the full deterministic ensemble of ANNs, $\overline{\text{ANN}}$ (Table 4.1).

We, therefore, couple the best performing stochastic deep learned parameterization, ANN-ens-CESM, as well as its deterministic counterpart, $\overline{\text{ANN}}$-CESM, to the coarse ESM host model. The coupling of the entire set of output variables $\boldsymbol{Y}^{pred}$ remains challenging. The related hybrid runs with the deep learned ensemble parameterizations are stable over a few days. Therefore we designed "perfect condensate" experiments, where we partially coupled our developed parameterizations including key surface radiative fluxes for surface coupling. In this setup, condensate tendencies are simulated with the superparameterisation running alongside. With this pragmatic approach we conduct stable hybrid model runs for a duration on the order of six months with ANN-ens-CESM and $\overline{\text{ANN}}$-CESM. Our ensembles are stabilizing hybrid runs with CESM2, while simulations with individual ANNs fail within the first 5 days in most cases. ANN-ens-CESM and $\overline{\text{ANN}}$-CESM capture precipitation extremes and intermediate precipitation, clearly outperforming the traditional Zhang-McFarlane scheme with respect to a superparameterization. However, our ensemble parameterizations introduce biases in the large-scale thermodynamic structures that lead to a weakening of the ITCZ and a displacement of its position compared to the superparameterization. Despite these limitations, our developed parameterizations simulate in general the diurnal peaks of precipitation with higher accuracy than the traditional Zhang-McFarlane scheme (Zhang and McFarlane 1995), e.g., shifting the too early peaks of continental precipitation in the tropics of the traditional scheme around noon towards the afternoon like seen with a superparameterization.

Despite these encouraging results, there remain several open questions. First of all, in an ideal case an operational hybrid model, with deterministic or stochastic deep learning parameterizations, would run stably without blowups or climate drifts (systematic and increasing long-term errors). Han et al. 2023 and Wang et al. 2022a proved that this is possible with realistic boundary conditions in coupled simulations over several years. However, while Wang et al. 2022a used an atmosphere only configuration, Han et al. 2023 neglected radiative fluxes important for atmosphere-land coupling. Future work will aim to further develop deep learning parameterizations, including the stochastic approaches proposed here, to enable accurate long-term *hybrid model* simulations. Another open question is how to increase the reproduction skill of cloud water and cloud ice water tendencies with deep learning models. Potential approaches may include: substituting deterministic metrics in the loss function for proper scoring metrics such as the Continuous Rank Probability Score, using loss functions that maximize likelihood (Haynes et al. 2023), or applying novel probabilistic data-driven models. A community benchmark dataset has recently been released that should facilitate intercomparisons between future advances in machine learning parameterizations for ESMs with state-of-the-art algorithms (Yu et al. 2023). Likewise, the use of a more flexible Fortran-Python coupler might enable us to explore the potential of latent space perturbation with VEDs to obtain well calibrated uncertainty quantification of convective processes also in coupled simulations.

This work demonstrates that online runs of deterministic ensemble and of stochastic deep leaning ensemble parameterizations with a complete coupling of subgrid radiative fluxes to a comprehensive land model are stable over a period of more than five months, provided issues of emulating condensate tendencies are sidestepped. We show that deep learning ensemble parameterizations improve the representation of convective processes, especially within the planetary boundary layer, compared to individual neural networks. We further demonstrate that this translates also into a strongly enhanced online stability of ensemble deep learning parameterizations compared to individual networks. Such ensemble parameterizations further have the potential to add to each prediction and variable a related uncertainty quantification. These are key steps forward to increase the quality of simulated complex processes like convection and the trustworthiness of deep learning parameterizations in general that will be developed for the next generation of Earth System Models.

6. Conclusion

This chapter contains summarizing remarks of my thesis under the scope of "Understanding and Modelling Convection with Machine Learning". It further highlights the implications of the thesis for the climate modelling community and for the general understanding of non-linear processes like convection gained with deep learning. Section 6.1 summarizes the main findings of this thesis based on my published paper (**Behrens** et al. 2022) and the one currently in review (**Behrens** et al. 2024). Section 6.2 contextualizes these main results within the broader framework of Earth system modelling and understanding complex processes in the Earth system.

6.1. Summary

Modelling convective processes in an Earth System Model (ESM) remains challenging due to the representation of subgrid processes resulting in persistent systematic errors compared to observations (i.e., **Behrens** et al. 2022; Douville et al. 2021; Gentine et al. 2018; Mooers et al. 2021; Rasp et al. 2018). Traditional convection schemes are a key source of biases in ESMs compared to observations (Douville et al. 2021). For example, biases in cloud radiative effects are closely associated with deficiencies of convective processes in ESMs, leading to the simulation of a double Inter-Tropical Convergence Zone (ITCZ) or biases over the maritime statocumuli regions (Bock et al. 2020; Lauer et al. 2023). In turn, these deficiencies in ESMs result in large uncertainies in climate projections (Douville et al. 2021; Lee et al. 2021), as shown by key climate metrics like Equilibrium Climate Sensitivity (ECS) (Bock et al. 2020; Eyring et al. 2021). Storm Resolving Model (SRM) simulations, which permit the direct representation of a large fraction of convective processes (Hohenegger et al. 2020), remain a challenge from a computational standpoint with current high-performance infrastructures (Satoh et al. 2019; Stevens et al. 2019; Stevens et al. 2020). Thus, it is intuitive to deep learn the resolved high-resolution convective processes and use the resulting data-driven subgrid convective scheme in an ESM (Gentine et al. 2021). A simulation with a machine learning parameterization coupled to a host climate model is also called "hybrid model" or a "hybrid simulation". Trailblazing work with a hybrid model showed remarkable reproduction capabilities of convective processes (Rasp et al. 2018), albeit with a substantial lack of understanding of the fields driving such processes, largely due to the deep and complex nature of the Artificial Neural Network (ANN). Nevertheless, the deep learned convective processes had almost the quality of a superparameterization (Rasp et al. 2018), a set of nested high-resolution columns within each grid column of an ESM. A superparameterization is also used in this thesis as a high-resolution benchmark

scheme to evaluate the reproduction skill of the developed deep learning algorithms. Despite the reproduction capabilities shown in Rasp et al. 2018, the interpretability and evaluation of the simulated convection is difficult due to the complexities of the ANN (degrees of freedom).

Having this lack of interpretability and understanding of deep learned convective processes in mind, a novel (by construction) interpretable model architecture is developed in this thesis. To do so, I use a Variational Encoder Decoder (VED), building on theoretical proof-of-concept experiments (e.g., Krinitskiy et al. 2019; Tibau Alberdi et al. 2018) based on Variational Auto Encoder (VAE) architectures. Further, I develop a postprocessing tool to obtain better understanding of deep learned convective processes. This novel approach leverages a setup in which the ambient large-scale thermodynamic conditions (inputs) are encoded within the five nodes of the latent space (i.e., a lower dimensional space between encoder and decoder of the VED). The latent space is decoded and used to reproduce the associated subgrid convective processes together with the large-scale thermodynamic conditions. Despite the strong dimensionality reduction (compressing and retaining only the essential information), the VED accurately learns and reproduces subgrid convective processes. The reproduction skill of the VED is comparable to the original ANN developed in Rasp et al. 2018. On the one hand, this indicates that the VED is skillfully predicting convective processes using only a fraction of the information (**Behrens** et al. 2022). On the other hand, the resulting latent space of the VED reveals a close relationship between the large-scale and subgrid variables, enabling an enhanced understanding and interpretability of the deep learned convective processes and related large-scale drivers (**Behrens** et al. 2022). Combining these two capabilities of the VED, encoding and decoding, I apply a novel generative modelling approach based on the distribution of the latent nodes to better understand large-scale drivers of convection and convective regimes represented by the high-resolution model (**Behrens** et al. 2022). Specifically, two latent nodes represent well-known drivers of convection, namely: the meridional temperature gradient from the poles to the equator; and large-scale thermodynamic fields, such as temperature and specific humidity, along the mid-latitude storm tracks that separate moist and warm subtropic air masses from cold and dry subpolar air masses (**Behrens** et al. 2022). The three remaining latent nodes characterize distinct convective regimes (**Behrens** et al. 2022): shallow convection in which convective processes are limited to the planetary boundary layer and adjacent layers of the lower troposhere; a cumulus regime with lower tropospheric optically thick cumulus clouds; a cirrus-like optical thin regime near the tropopause; and a deep convection regime with cloud tops near the tropopause and related heating and moistening throughout the troposphere above the planetary boundary layer. These deep learned convective regimes are in good agreement with observations (Huaman and Schumacher 2018), as well as the prescribed regimes in idealized multi-convection regime model simulations (Frenkel et al. 2012, 2013; Frenkel et al. 2015; Khouider and Majda 2006). My newly developed generative modelling approach has some key advantages, as it is fully data-driven. There is no need to specify heating or moistening profiles related to distinct convective regimes for the analysis (e.g., Frenkel et al. 2012). Moreover, one does not need to infer convective regimes based on binning precipitation

(e.g., Huaman and Schumacher 2018), which is sensitive to predefined thresholds like the bin widths.

Let us come back to the first scientific question of this thesis: "Can deep learning enhance the understanding of convection and large-scale drivers of convection?" Overall, VEDs are powerful data-driven algorithms to disentangle complex convective processes via encoding-decoding essential information. They allow us to boost our understanding of convective processes by linking convective regimes with their respective large-scale drivers (**Behrens** et al. 2022). Therefore the first key scientific question of this thesis can be answered with "Yes, deep learning enhances our understanding about convective processes and related large-scale drivers".

Key limitations of individual deterministic deep learning models, however, are their over-confident predictions and low skills representing convective processes in the planetary boundary layer (i.e., **Behrens** et al. 2022; Mooers et al. 2021; Rasp et al. 2018). It is hypothesized that this may be related to the chaotic nature of convective processes in the planetary boundary layer that is difficult to reproduce with single deterministic deep learning algorithms (**Behrens** et al. 2022; Mooers et al. 2021). Likewise, multi-output deterministic models have well-known deficiencies in reproducing certain convection-related subgrid variables such as specific humidity tendencies in the planetary boundary layer, due to optimization challenges (e.g., Han et al. 2023). Therefore, I address these challenges by developing deep learning ensembles that take into account the stochastic nature of convective processes. Both deterministic and stochastic ensembles of deep learned models are used to investigate and improve reproduction skills with respect to "stochasticity", taking into account the variability related to the chaotic nature of convective processes. Furthermore, this work focuses on a realistic, thus more challenging, modelling framework without an aquaplanet setup by using an ESM with realistic topography and radiative forcing. In contrast to previous work (Han et al. 2023), the setup used in this thesis includes deep learning essential surface subgrid radiative fluxes enabling land-atmosphere interactions (**Behrens** et al. 2024). Based on large-scale atmospheric thermodynamic fields, the resulting deep learning parameterizations represent surface subgrid radiative fluxes for surface coupling. In addition, precipitation rates from the previous time step are included as additional input variable, which allows to investigate the effect of convective memory (e.g., Han et al. 2023). Including convective memory turns out to be especially beneficial for the reproduction of near surface heating and moistening rates in the atmosphere (**Behrens** et al. 2024). In this work, two different model architectures are applied, ANNs and VEDs, to create ensembles of predictions accounting for aleatoric uncertainty (i.e., the part of the uncertainty related to the randomness of the underlying data, **Behrens** et al. 2024). To construct an ensemble of predictions, I use three different approaches: 1) an ANN with a Monte Carlo Dropout Layer, in which active dropout is used to generate a stochastic ensemble (**Behrens** et al. 2024); 2) a stochastic and deterministic multi-neural network ensemble of predictions (**Behrens** et al. 2024); and 3) a stochastic ensemble based on a single VED either with or without latent space perturbation with Gaussian noise (**Behrens** et al. 2024). The ensemble

size, or the number of ensemble members, is a crucial hyperparameter, which trades-off the uncertainty spread against the computational overhead (**Behrens** et al. 2024).

The aforementioned deep learning stochastic approaches are evaluated offline based on their capabilities to reproduce subgrid convection. An enhanced reproduction skill of specific humidity tendencies and cloud liquid water tendencies is found for both, the stochastic and deterministic ANN ensemble over individual ANNs within the planetary boundary layer (**Behrens** et al. 2024). Similar improvements are present for the upper tropospheric cloud ice water tendencies (**Behrens** et al. 2024). Regarding specific humidity tendencies within the planetary boundary layer, the increase of prediction skill mainly originates from an enhanced representation of convective processes over the Southern Ocean, as well as in subpolar and polar latitudes in both hemispheres (**Behrens** et al. 2024). However, the stochastic ensemble based on the ANN with Monte Carlo Dropout lacks reproduction improvements of convective processes compared to individual ANNs (**Behrens** et al. 2024). The same is true for the multi-VED ensemble approach (**Behrens** et al. 2024). The weakest reproduction skill of convective processes using stochastic ensembles is found for the latent space perturbation of a single VED. Nevertheless, this is compensated to some extend by the realistic uncertainty quantification of these processes (see below; **Behrens** et al. 2024). Overall, the stochastic and deterministic ANN ensemble improves the representation of convective processes, especially within the planetary boundary layer and the upper troposphere, compared to individual models. Moreover, the advantages of these ensembles are found to be especially important for variables that are associated with a greater influence of subgrid turbulence, such as specific humidity tendency, cloud liquid water tendency, and cloud ice water tendency (**Behrens** et al. 2024). These variables are typically poorly represented by individual ANNs.

Let us come back to the second scientific question of this thesis: "Can stochastic and deterministic ensemble deep learning parameterizations that take into account the stochasticity improve the representation of subgrid convective processes "offline" based on ESM data?" The results presented in this thesis show that multi-network ensembles improve deep learned representations and more effectively capture the stochastic, turbulent nature of convective processes compared to individual models with overconfident predictions. This finding aligns with the deterministic findings of Han et al. 2023. As a result, the second key scientific question can be answered with: "Yes, stochastic and deterministic deep learning ensembles improve the representations of convective processes and reduce deficiencies of individual neural networks with respect to the reproduction of ESM data."

Building on the evaluation of the offline reproduction capabilities of the different stochastic and deterministic ensembles, I next focused on the uncertainty quantification of the ensembles given by the predictions of individual ensemble members. This detailed uncertainty analysis for a multi-variate prediction of subgrid convective processes is also a novel aspect in the field of deep learning for climate science. Quantifying the uncertainties captured by the ensembles is essential to assess them against the true variability represented by the high-resolution ESM. In other words, uncertainty quantification enables an evaluation of the physical plausibility of the captured stochasticty of the parameterizations. The ensemble based on the ANN with Monte

Carlo dropout results in a poorly captured uncertainty range, largely underestimating the spread represented by the high resolution model (i.e., an underdispersion of the Monte Carlo dropout based ensemble) (**Behrens** et al. 2024). This is in agreement with previously published literature (Haynes et al. 2023). In contrast, the stochastic VED via latent space perturbation shows the best calibrated aleatoric uncertainty for all evaluated variables (**Behrens** et al. 2024). The Continuous Rank Probability Score (CRPS) is used as a metric to assess the quality of the reproduction and the uncertainty quantification together. While the stochastic VED via latent space perturbation approach outperforms the ANN with Monte Carlo dropout method, the stochastic and deterministic ANN ensemble yields the best combined performance based on CRPS for all subgrid convective processes (**Behrens** et al. 2024). Therefore, these two multi-ANN based ensemble approaches are implemented into the ESM. All subgrid variables learned by the ANNs, except for cloud ice and cloud liquid water, are coupled to the dynamical core of the ESM, and partially replace the high-resolution superparameterization scheme (**Behrens** et al. 2024). The "online" implementation of the deep learned parameterizations is especially challenging due to the importance of surface subgrid radiative fluxes for the atmosphere-land interactions, and the stability of the hybrid model (**Behrens** et al. 2024). Hybrid model simulations with both, the stochastic and the deterministic ANN ensemble parameterizations, are stable over more than 5 months (**Behrens** et al. 2024). In contrast, hybrid model realizations with individual ANNs become unstable after a few simulated days (**Behrens** et al. 2024). This demonstrates the beneficial stabilizing effect of stochastic and deterministic ensemble parameterizations compared to standard deep learning approaches (**Behrens** et al. 2024; Han et al. 2023). Moreover, the resulting hybrid model simulations with stochastic and deterministic ANN parameterizations enhance the reproduction of extreme precipitation and the diurnal precipitation cycle compared to the simulations using the current Zhang-McFarlane scheme (Zhang and McFarlane 1995). Remarkably, these hybrid model simulations substantially alleviate also the drizzle precipitation bias compared to those with the traditional scheme, and are in good agreement with the original high-resolution ESM (**Behrens** et al. 2024).

Let us come back to the third scientific question of this thesis: "Do stochastic and deterministic ensemble parameterizations with calibrated uncertainty quantification of subgrid processes have an effect on the stability and improve the quality of hybrid ESM simulations?" The work presented in this thesis shows that stochastic and deterministic ensemble parameterizations with calibrated aleatoric uncertainty enhance the stability of hybrid models compared to realizations with individual ANNs and result in an improved reproduction of precipitation, both reducing apparent biases and its diurnal cycle (**Behrens** et al. 2024). Thus the third key scientific question can be answered with: "Yes, the constructed ensembles are capable to provide calibrated uncertainty quantifications for a multi-variate data set of an ESM. Moreover, the calibrated uncertainties of the ensembles enable a stabilization and improved quality of hybrid model simulations in an ESM."

In the next section I will illustrate the general implications of my thesis and my published work (**Behrens** et al. 2022; **Behrens** et al. 2024; Yu et al. 2023) and will give an outlook of my research in the context of enhancing climate modelling with machine learning.

6.2. Concluding Remarks and Outlook

Machine learning and deep learning has flourished for climate science applications in the last decade (Gentine et al. 2021). There has been significant advances in machine learning over this period that allow an improved simulation of convective processes with data-driven schemes in ESMs, previously parameterized via in quality limited convection schemes (Gentine et al. 2021). However, state-of-the-art machine learning solutions still exhibit limitations and challenges. Particularly, the low interpretability of their predictions, the lack of predictions taking into account the stochasticity of reproduced processes and the related aleatoric uncertainty, and common instability issues in hybrid simulations related to deficiencies of individual neural networks. This thesis illustrates ways forward to overcome or alleviate these limitations with broader implications in Earth system sciences and other fields. First, this thesis shows the applicability of VEDs to further enhance our understanding about convective processes and the large-scale environment in which they are forming. Second, capturing the aleatoric uncertainty via ensembles improves the general reproduction of convective processes on atmospheric levels with pronounced stochasticity e.g., in the planetary boundary layer. Moreover, calibrated aleatoric uncertainty enhances the evaluation of the physical plausibility of the results of machine learning schemes (Haynes et al. 2023). Calibrating aleatoric uncertainties for multi-variate ESM data is a novelty in climate science, like it is shown in this thesis. Third, this thesis complements the results of existing hybrid simulations in an ESM-like configuration with clear advantages over existing convection schemes (Han et al. 2023; Wang et al. 2022b). Using ensembles alleviates hybrid model instabilities of machine learning schemes coupled to ESM. This was previously hypothesized (Brenowitz et al. 2020) or could not be shown in hybrid simulations due to the required computational resources of the designed ensembles (Han et al. 2023) and is shown in this thesis in hybrid simulations with an ESM.

Apart from the work presented in this thesis, increasing the interpretability of machine learning algorithms e.g., with VEDs and other models with a latent space, has great potential to further enhance the understanding of complex non-linear processes in the atmosphere (Mooers et al. 2023; Shamekh et al. 2023). Models with a latent space may also enhance the identification of circulation regimes in the ocean similar to previous work based on traditional clustering algorithms (Sonnewald et al. 2019). Novel methods learning functional relationships directly from high resolution simulations via equation discovery may enable an improved understanding of complex non-linear processes in the Earth system and enhance their interpretability (Grundner et al. 2024). Regarding convective processes one suitable example may be finding an equation for the effect of a stratocumuli regime, that may be detected via the latent space of a VED in high-resolution simulations, on temperature or radiative

fluxes to alleviate longstanding biases of ESMs. Moreover causal discovery can be applied to help understand the key differences between the causal relationships and correlations, that result from common machine learning, in the Earth system (Iglesias-Suarez et al. 2024). In this context, a causal evaluation of the latent space and the generated convective processes of a VED may further broaden our understanding about convection and its connection to the large-scale thermodynamic fields. Likewise concepts from climate science, namely an intercomparison between different machine learning architectures based on protocols, show large potential to alleviate lacks of interpretability (Yu et al. 2023). Moreover, such intercomparisons are a starting point to investigate the effect of stochasticity on the reproduction of machine learning algorithms in an ESM (Yu et al. 2023). To improve the low reproduction of cloud water tendencies a complete stochastic framework with stochastic losses like CRPS or stochastic machine learning algorithms may help (**Behrens** et al. 2024; Haynes et al. 2023). There is large potential to enhance reconstructions of non-linear processes in the Earth system via stochastic machine learning with skillful uncertainty quantification in the future (Haynes et al. 2023), as traditional methods and initial machine learning experiments have proven (i.e., Berner et al. 2017; Sakradzija and Klocke 2018; Yu et al. 2023). The limitations of the parameterizations proposed in this thesis that lead to instabilities of hybrid simulations may be overcome with entire stochastic frameworks (Haynes et al. 2023) or constraints from the theoretical climate science that limit the necessary extrapolation of the machine learning ensembles (Beucler et al. 2021). In this context causal discovery can be used to identify non-physical correlations of machine algorithms that cause instabilities of hybrid simulations coupled to ESMs (Iglesias-Suarez et al. 2024).

This thesis shows steps forward alleviating existing limitations in machine learning algorithms, with the ultimate goal of improving Earth System Models and reducing uncertainties in their future climate projections. This thesis complements the advances in climate science over the recent years with a novel interpretable and robust data-driven approach that can help the scientific community to develop the next generation of Earth System Models and make them more explainable despite growing complexity of the simulated processes.

A. Supporting materials for Chapter 3: Understanding convective processes in a climate model with a Variational Encoder Decoder

Appendix A directly reproduces the supporting material of my published paper (**Behrens et al. 2022**). All Figures and Tables were produced from me as author of the thesis. Moreover I led the writing of the text for the supporting material that is shown in this appendix.

A.1. Introduction

The supporting information are structured as follows and each section can be read individually:

In section A.2 we show the hyperparameters of VED and explain how we conducted the search for a suitable set of hyperparameters of VED. Furthermore we discuss the used VED output normalization dictionary. In section A.3 we show additional figures for the general evaluation of VED and other reference networks. This section further describes differences in reproduction skill if either the VED or the output normalization of Rasp et al. 2018 is used. Furthermore we describe differences in the interpretability between the VED's latent space and a principal component analysis on the large-scale variables in this section. Also we show that the latent space exploration with conditional averages can be conducted on the five original latent dimension. Section A.4 shows one alternative VED and a conditional VAE structure and discusses their strengths and limitations. We describe in subsection A.4.1 the $\text{VED}_{X \rightarrow Y}$ and in subsection A.4.2 a conditional Variational AutoEncoder Decoder (cVAE). Section A.5 comprises the tables of all generated 2D SP or CAM variables with our generative modeling approach. Additionally the squared Pearson correlation coefficients R^2 between the latent nodes and vertical heating, moistening, specific humidity and temperature profiles in space-time and time are shown in this section respectively.

Hyperparameter of VED	**Values**
Learning Rate	0.00074594
Training / learning rate decrease	40 epochs, learning decrease every 7^{th} epoch by factor 5
Batch size	714
Latent Space Width	5 nodes
Node Size of Encoder	[64,463,463,232,116,58,29,5]
Node Size of Decoder	[5,29,58,116,232,463,463,129]
Activation Encoder	[Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, Lambda]
Activation Decoder	[Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, ELU]
KL Annealing	Linear annealing from 2^{nd} to 7^{th} epoch

Table A.1.: Hyperparameters and architecture of the final VED which uses large-scale CAM variables **X** to investigate simulated convective processes of SP **Y** together with driving climate conditions. This Table is directly reproduced from **Behrens** et al. 2022.

Hyperparameter range of VAE$_{X\rightarrow X}$	**Values**
Initial learning rate	10^{-5} to 5×10^{-4}
Batch size	200 to 8192
Latent Space Width	2 to 5 nodes
Node Size of first or last hidden layer of Encoder or Decoder	300 to 500
Depth of Encoder or Decoder in hidden layers	5 to 7 hidden layers

Table A.2.: Hyperparameter range of search for initial VAE, which reproduces large-scale climate variables **X** with **X** as input data set. The hyperparameter search was conducted over 120 trials and 30 epochs with a learning rate decrease after every 5^{th} epoch by a factor 5. This Table is directly reproduced from **Behrens** et al. 2022.

Hyperparameter range of VED	**Values**
Initial learning rate	5×10^{-5} to 5×10^{-3}
Batch size	200 to 8000

Table A.3.: Hyperparameter range of search for VED, the main model in this study. The hyperparameter search was conducted over 80 trials and 20 epochs with one learning rate decrease after the 10^{th} epoch. This Table is directly reproduced from **Behrens** et al. 2022.

Network	Training MSE	Validation MSE	Test MSE
VED	0.162	0.165	0.165
ED	0.162	0.165	0.165
LR	0.242	0.244	0.243
ANN	0.133	0.135	0.135

Table A.4.: Mean Squared Error (MSE) of predicted subgrid-scale SP variables **Y** of the VED, ED, LR, reference ANN on the training, validation and test data sets (3 month of SP data) using the VED output normalization. This Table is directly reproduced from **Behrens** et al. 2022.

A.2. VED Hyperparameters based on a Hyperparameter Search and Normalisation

In earlier experiments we found that the output normalisation used in Rasp et al. 2018 was not well-suited for the optimization of a VED during training. With their output normalisation dictionary, the VED focused solely on the reproduction of radiative fluxes in **Y** and lacked skill with respect to heating and moistening profiles. Therefore we had to re-scale the output normalisation dictionary for **Y** and implement a suitable scaling for the extended variable list **O**. The vertical profiles of temperature, specific humidity and specific humidity tendency are normalised by long-term (3 month) standard deviations of the near surface model level. In the case of temperature tendencies, the standard deviation on the 845 hPa level is used due to the dominant variability of convection related temporal temperature changes on this level near the upper limit of the planetary boundary layer in SP data. The remaining 2D variables of radiative properties, precipitation rates and surface pressure are standardised.

We initially performed a hyperparameter search (random search) with 120 trials for a $\text{VAE}_{X \rightarrow X}$,which was trained on large-scale climate variables **X** to reproduce X. Table A.2 shows the hyperparameter range for a hyperparameter search over a sequence of 1 month of SP data.

The best-performing encoder and decoder hyperparameter settings; 6 hidden layers, 463 nodes in the first and last hidden layer and a latent space width of 5 nodes; were fixed for the development of the VED presented in the paper. To account for shifts in suitable learning rates and batch size due to the additional subgrid-scale output variables **Y**, we conducted a second hyperparameter search (random search) for our main VED specifically over a sequence of 1 month of SP data, see Table A.3.

After that, we fixed the initial learning rate and batch size and conducted further sensitivity tests with respect to the latent space width (which are documented in Chapter 3) of VED. The choice of activation functions in the hidden layers is based on small initial experiments with VED, which showed enhanced emulation skill if the last hidden layer was elu-activated (exponential linear unit).

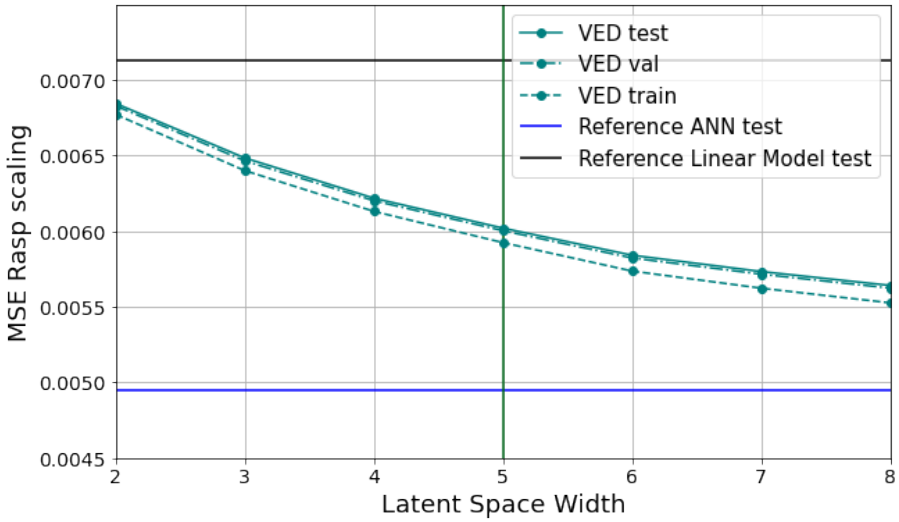

Figure A.1.: Similar to Figure 3.2, mean squared error (MSE) as a function of Latent Space Width of the VED for the test (solid cyan), validation (dashed-dotted cyan) and training data set (dashed cyan curve) using the output normalization of the reference ANN (Rasp et al. 2018) as y-axis. The horizontal solid blue / black line represents the MSE scores of the reference ANN Rasp et al. 2018 / a linear version of this network (Reference Linear Model) on test data with fixed layer width of 256 nodes in the 9 hidden layers. This Figure is directly reproduced from **Behrens** et al. 2022.

A.3. Evaluation of VED and the Reference Networks

If we use the output normalization of reference ANN to investigate the sensitivity of the VED performance as a function of latent space width, then we observe similar asymptotic behaviour as in Figure 2, see Figure A.1. The VED shows an improved emulation skill compared to the reference linear model with fixed layer widths of 256 nodes. The difference in performance between the VED and reference ANN increases if the output normalization of Rasp et al. 2018 is used, which points to the fact that the VED output scaling weights SP variables **Y** differently. The VED has a decreased performance compared to reference ANN, but is converging to a similar level of emulation with increasing latent space width.

The latent space of VED can be explored with the computation of conditional averages in a 2D PCA compressed submanifold as it is shown in Figure 5. However this analysis can be complemented with an inspection of the 5 latent dimensions itself. To visualize the five dimensional latent space, we projected two latent variables onto each other. This results in 20 spanned submanifolds of different latent variables and five projections of one latent variable onto itself oriented along the main diagonal in Figure A.11 to A.14. These 2D submanifolds of two different latent variables are often characterised by two or three centers of action with a strong concentration of samples (Figure A.11). In most cases there is a weak linear connection between the latent variables, except for latent variable 2 (Large-scale variations along mid latitude storm tracks) and latent variable 5 (Deep Convection), as can be seen in Figure A.11. The projection of these two latent variables is also characterised by a pronounced separation of samples with negligible convective processes (no precipitation, Figure A.12)

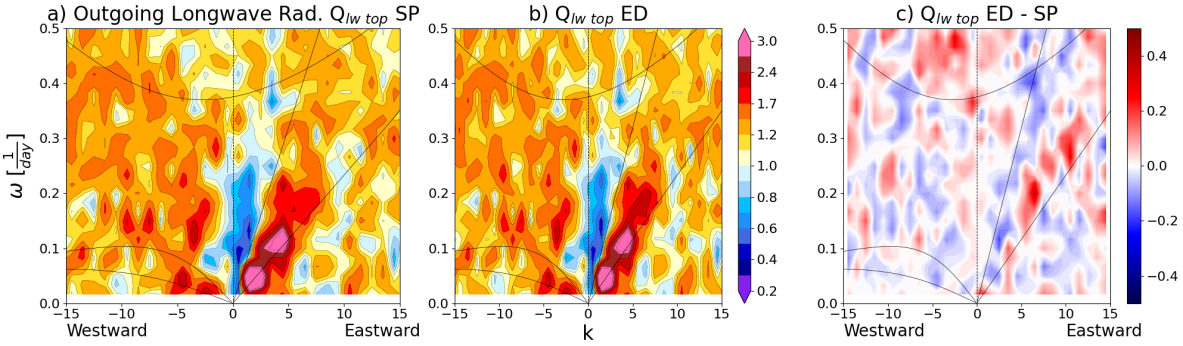

Figure A.2.: Wheeler Kiladis diagram based on tropical outgoing longwave radiation [15° N - 15° S] of SP (a), of ED predictions (b) and the absolute difference of spatio-temporal wave spectra ED - SP (c) for 1 year of SP data. This Figure is directly reproduced from **Behrens** et al. 2022.

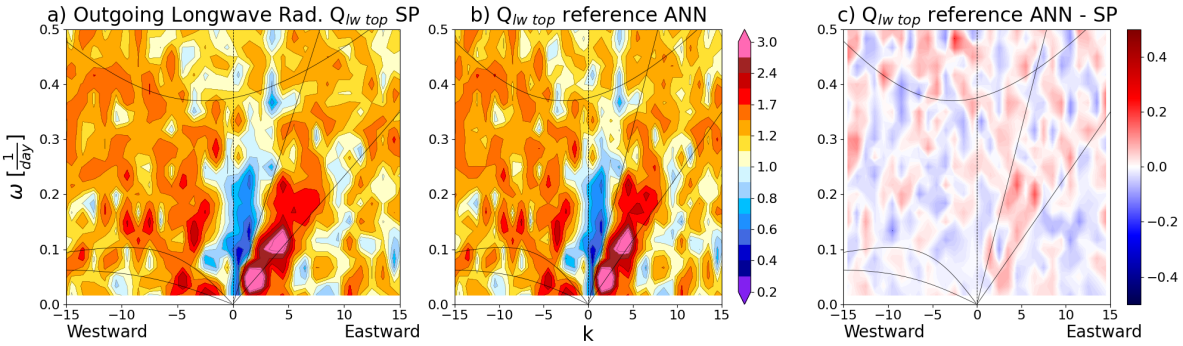

Figure A.3.: Wheeler Kiladis diagram based on tropical outgoing longwave radiation [15° N - 15° S] of SP (a), of reference ANN predictions (b) and the absolute difference of spatio-temporal wave spectra reference ANN - SP (c) for 1 year of SP data. This Figure is directly reproduced from **Behrens** et al. 2022.

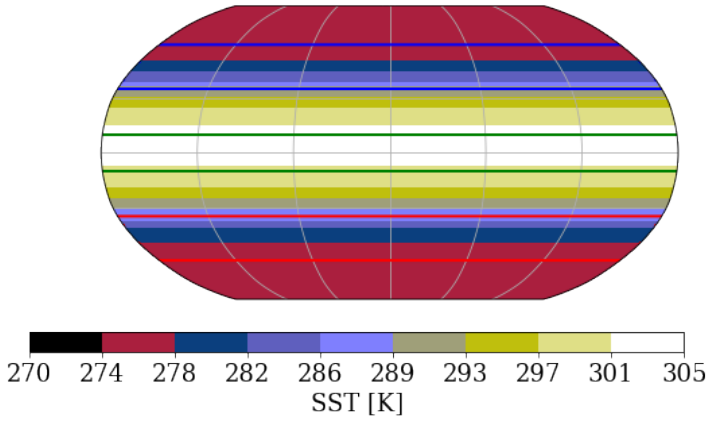

Figure A.4.: Fixed Sea Surface Temperature (SST) forcing of the Super Parameterized Community Atmosphere Model (SPCAM) simulation following (Andersen and Kuang 2012). The blue / red zonal lines indicate the region of Northern / Southern mid latitudes between 60° N/S and 35° N/S. The green lines indicate the deep tropics with the ITCZ between 10° S and 10° N. This Figure is directly reproduced from **Behrens** et al. 2022.

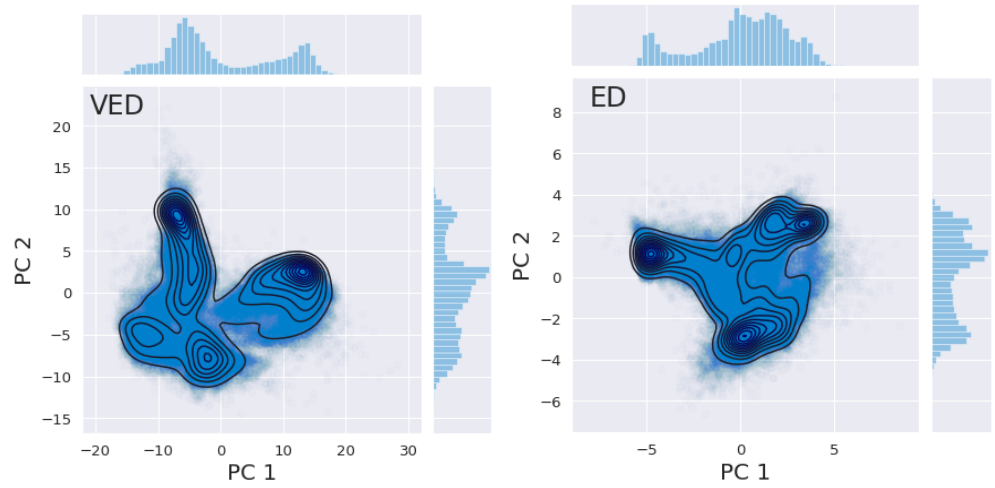

Figure A.5.: Scatter plot with isolines and histograms of the Joint and Conditional distributions of the PCA compressed latent space of VED (left) and ED (right). The plot is based on 100000 randomly picked samples from CAM test data. The 1^{st} / 2^{nd} PC of the resulting compressed latent space is the x-axis / y-axis in the respective subplot. This Figure is directly reproduced from **Behrens** et al. 2022.

and deep convective samples. This shows that the convective strength of the samples can be gauged with these two latent variables of VED and is not relying on a PCA as postprocessing step. Moreover the latent variables itself can be utilised to investigate large-scale geographic variability. One particular example for that, is the projection of latent variable 1 (Global temperature variations) and 2 (Large-scale variability along mid latitude storm tracks). In this submanifold we see two separated maxima of solar insolation (Figure A.13) and two areas with no solar insolation (night-time conditions). If we compare this distribution to the conditional averages of the surface air temperature (Figure A.14), we observe that one solar insolation maximum is associated with a minimum in surface air temperatures below 275 K, which can be only observed in polar latitudes. The combination of solar insolation with anomalous cold temperatures is a clear evidence that the respective samples are originating from austral polar or subpolar latitudes due to the austral summer solar forcing of the SPCAM simulations. In contrast the other minimum in surface air temperatures in this projection of latent variable 1 and 2 is associated with no solar insolation. This suggests that the corresponding samples are coming from the boreal high latitudes (due to constant polar night conditions). These two examples illustrate that the interpretation of convective processes and large-scale drivers of convective predictability is possible on the latent variables of VED itself and not relying on the PCA postprocessing step. Furthermore the latent space of VED can be used to investigate longstanding hypotheses of atmospheric science. As an example we can focus on the projection of latent variable 5 (Deep Convection) onto latent variable 1 (Global temperature variations). The strongest precipitating samples in Figure A.12 are situated in the middle of the conditional distribution of latent variable 1 (Figure A.11 and not in the right tail of the marginal distribution, which suggests that strong precipitation is not occurring in the regions with the highest surface air temperatures. This hypothesis can be evaluated with Figure A.14, where the region with the strongest precipitation is associated with conditional

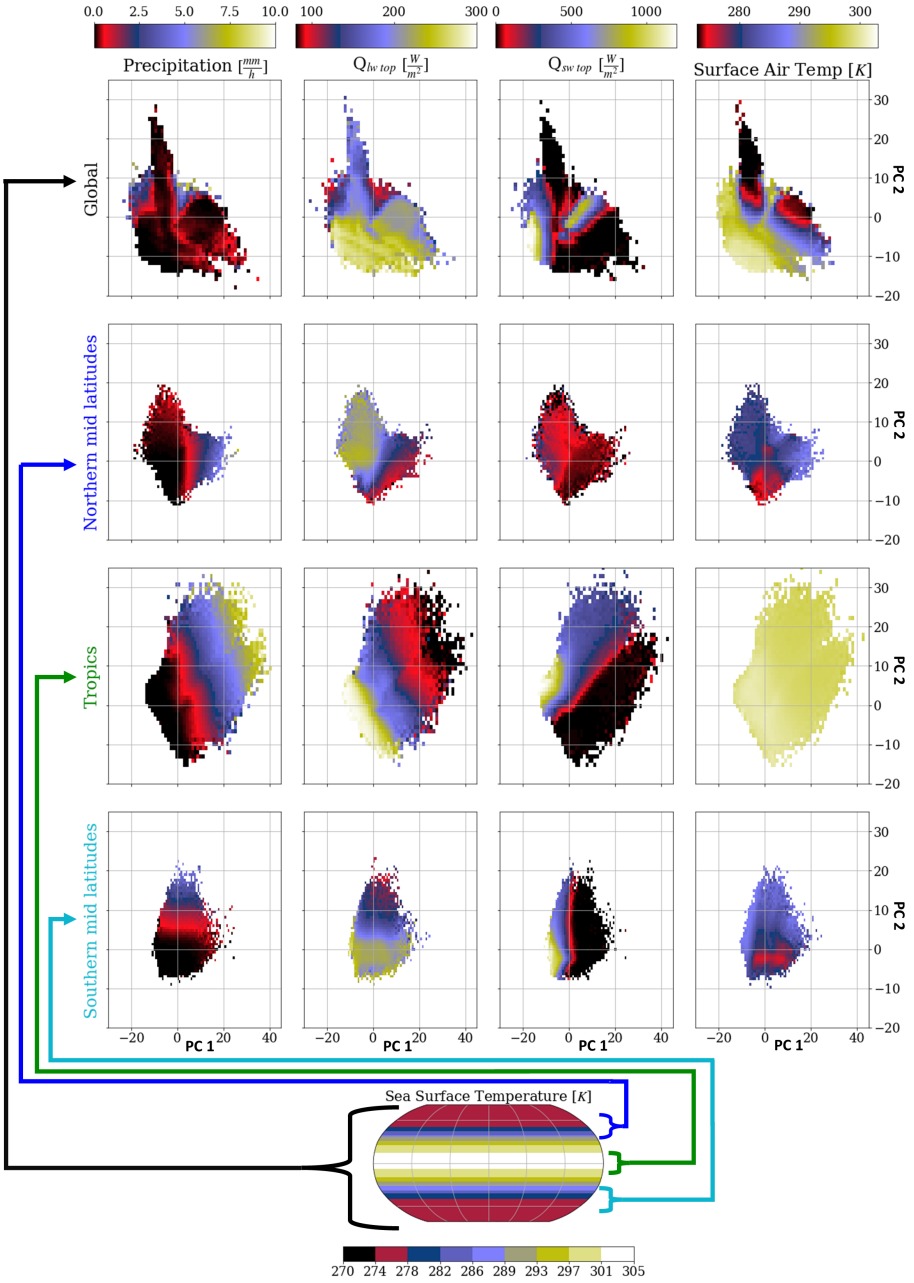

Figure A.6.: Latent Space clustering of VED for precipitation (left), outgoing longwave radiation ($\mathbf{Q}_{lw\ top}$) (left middle), shortwave heat flux at the model top ($\mathbf{Q}_{sw\ top}$) (right middle) and Surface Air Temperature ($\mathbf{T}_{surf}$) (right column). The first row illustrates the clustering in the PCA compressed latent space with respect to the SP / CAM variables on global scales (as seen in Figure 3.5). The lower rows depict the Latent Space clustering in the evaluated regions Northern Mid Latitudes (2^{nd} row), Tropics (3^{rd} row) and Southern Mid Latitudes (4^{th} row). The x-axis / y-axis represents the 1^{st} / 2^{nd} leading PC of the global / regional latent space. This Figure is directly reproduced from **Behrens** et al. 2022.

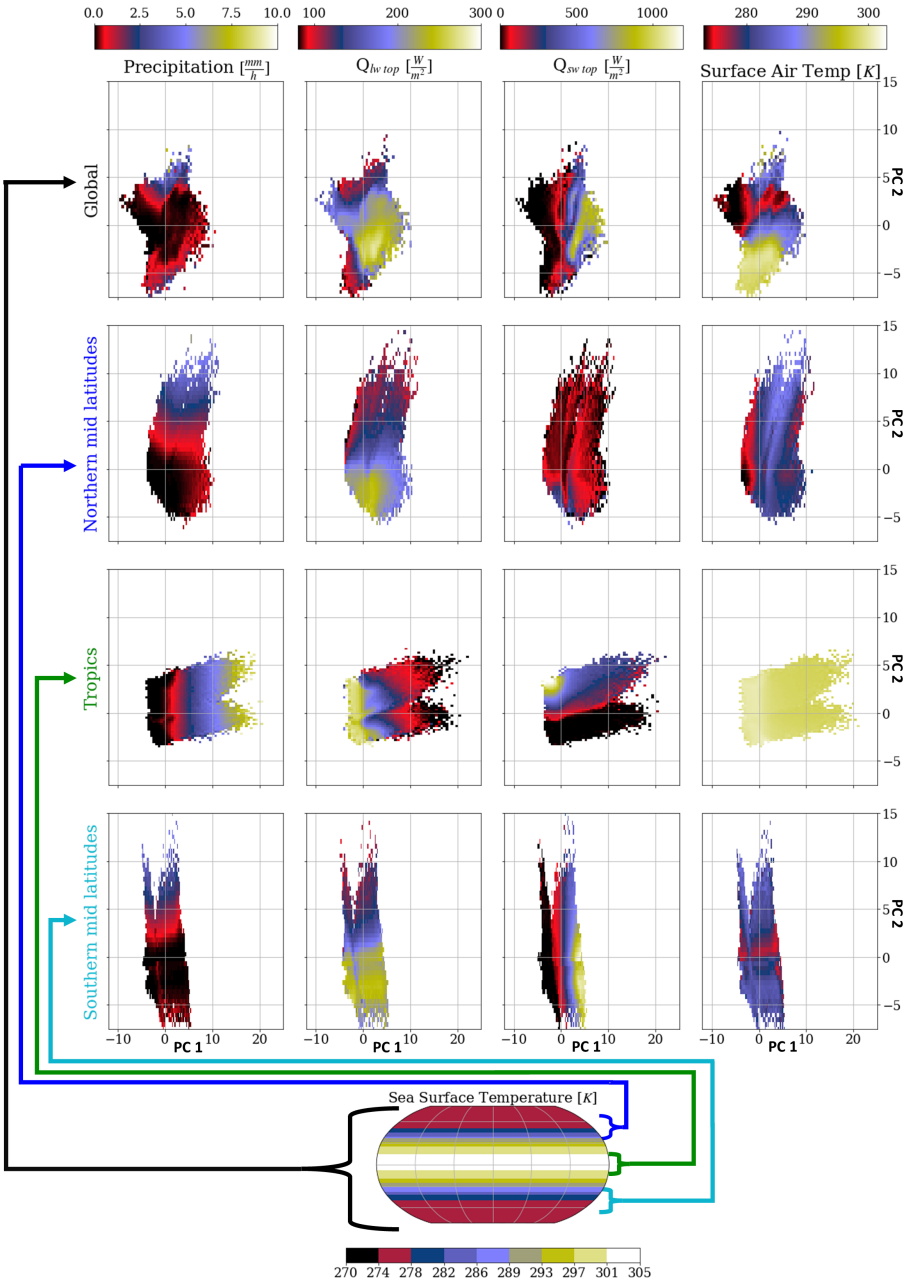

Figure A.7.: Latent Space clustering of ED for precipitation (left), outgoing longwave radiation ($\mathbf{Q}_{lw\ top}$) (left middle), shortwave heat flux at the model top ($\mathbf{Q}_{sw\ top}$) (right middle) and Surface Air Temperature ($\mathbf{T}_{surf}$) (right column). The first row illustrates the clustering in the PCA compressed latent space with respect to the SP / CAM variables on global scales (as seen in Figure 3.5). The lower rows depict the Latent Space clustering in the evaluated regions Northern Mid Latitudes (2^{nd} row), Tropics (3^{rd} row) and Southern Mid Latitudes (4^{th} row). The x-axis / y-axis represents the 1^{st} / 2^{nd} leading PC of the global / regional latent space. This Figure is directly reproduced from **Behrens** et al. 2022.

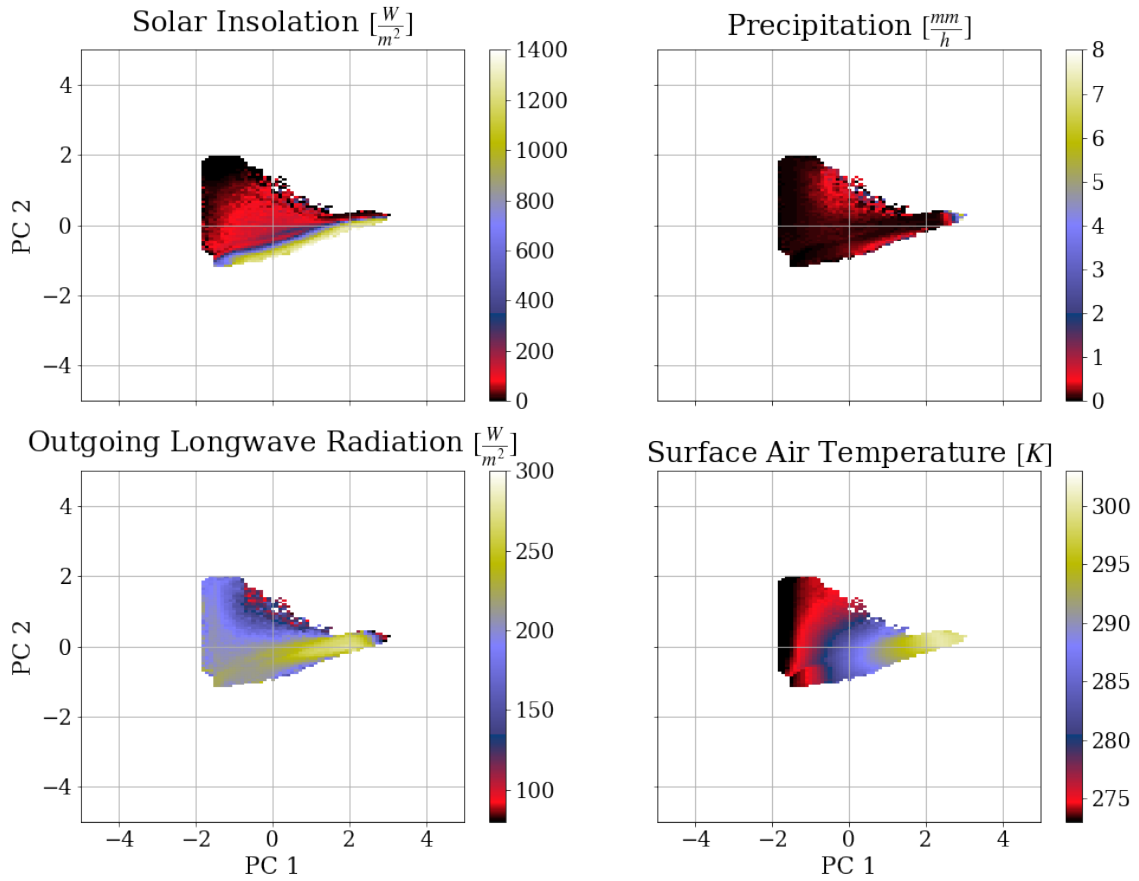

Figure A.8.: Conditional averages of solar insolation (upper left), precipitation (upper right), outgoing longwave radiation (lower left) and surface air temperature (lower right panel) in the submanifold spanned by the first two leading principle component's of the large-scale variables **X**. Similar to Figure 3.5 the conditional averages are computed based on 1000000 randomly selected samples from the test data set. This Figure is directly reproduced from **Behrens** et al. 2022.

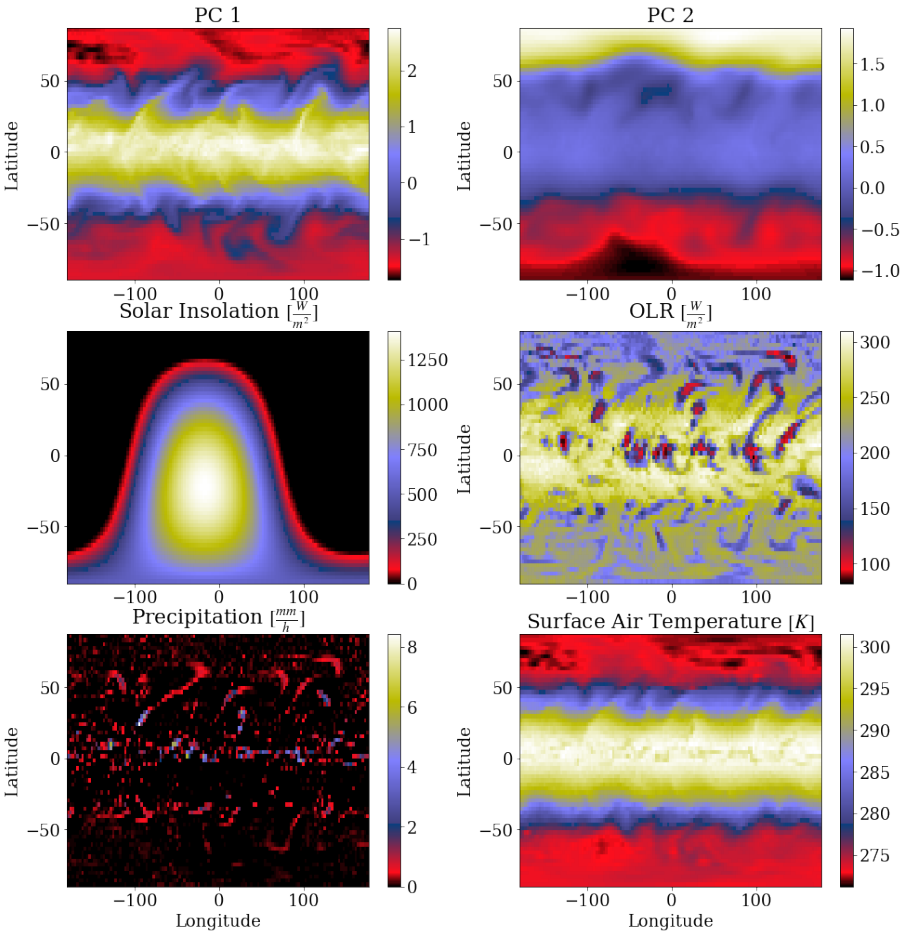

Figure A.9.: Latitude-Longitude plot of the first (upper left, PC1) and second (upper right, PC2) leading principal component of a PCA on the large-scale variables **X** and respective large-scale and subgrid-scale variables of the test data set for a particular time step. This Figure is directly reproduced from **Behrens** et al. 2022.

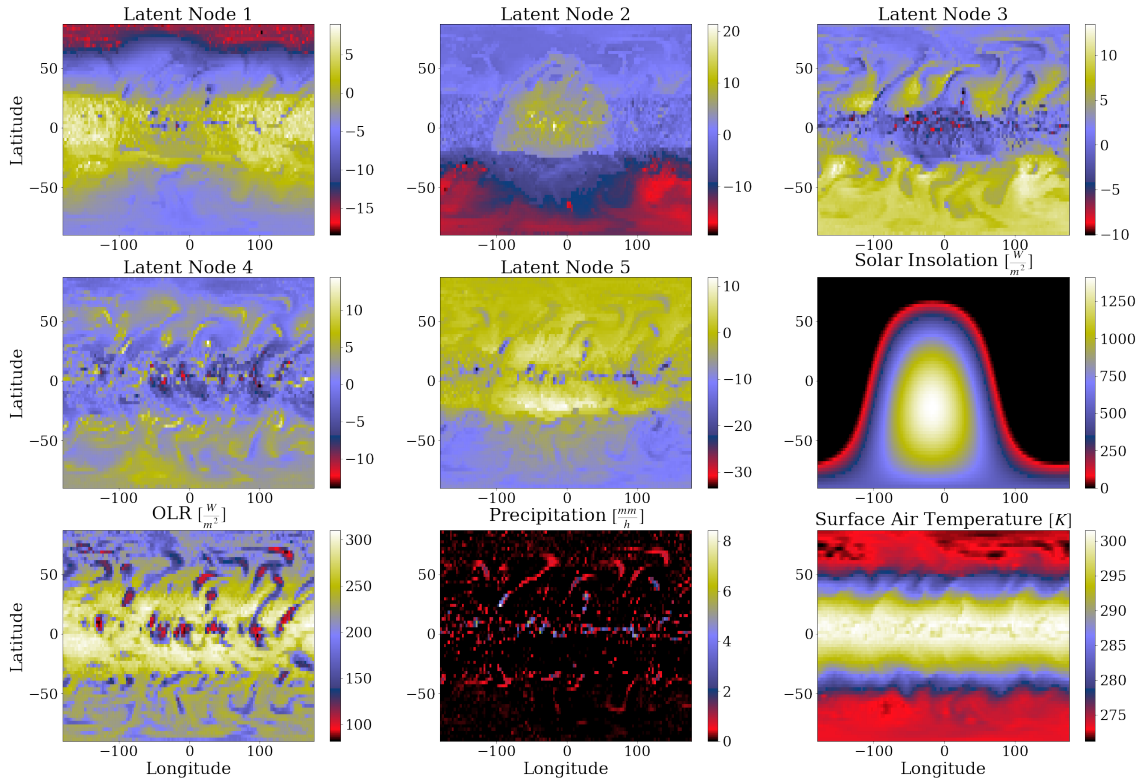

Figure A.10.: Latitude-Longitude plot of the latent variables of the VED (Latent Node 1 to 5) and respective large-scale and subgrid-scale variables of the test data set for the same time step as in Figure A.9. This Figure is directly reproduced from **Behrens** et al. 2022.

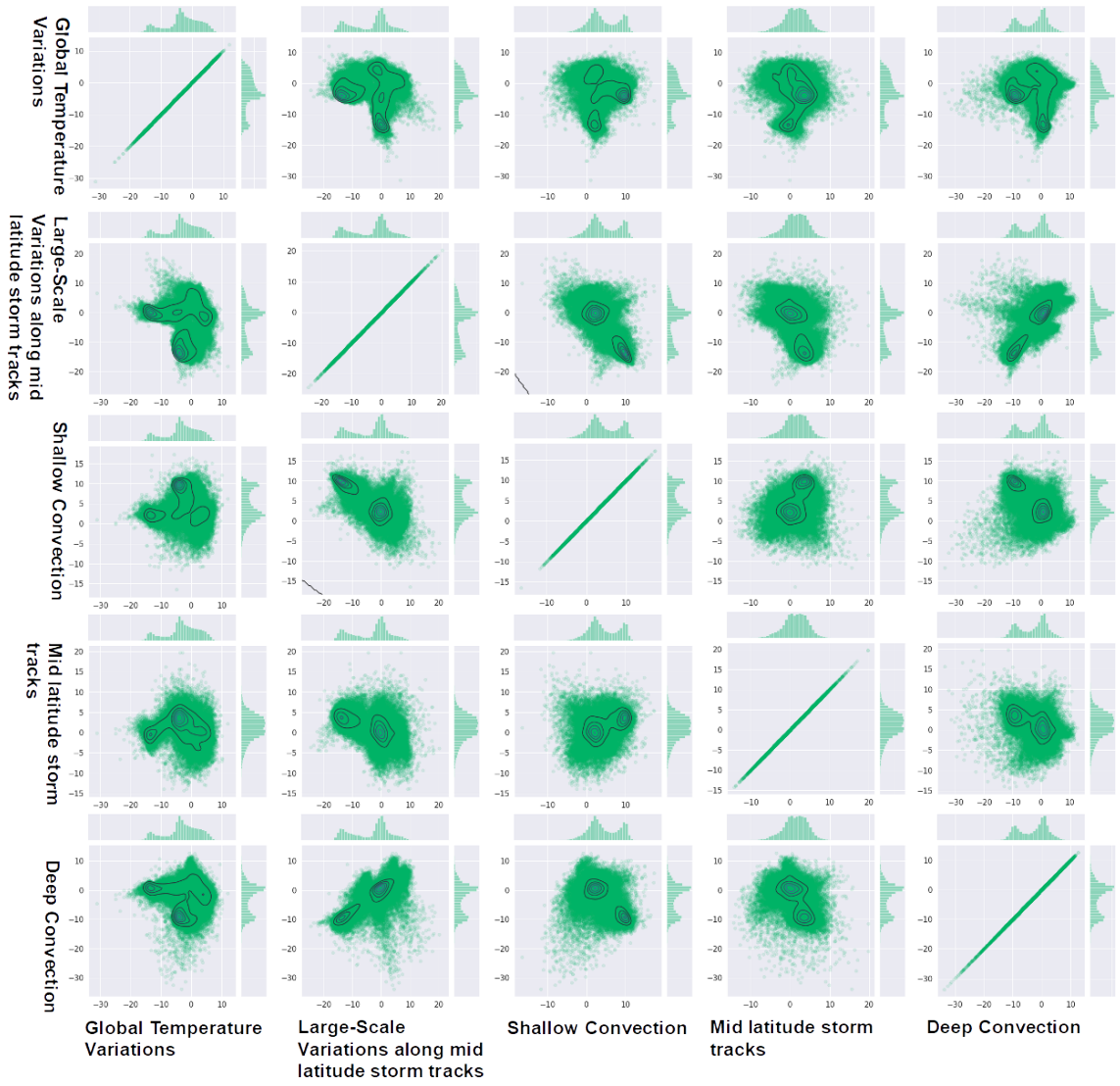

Figure A.11.: 2D Density (blue contours) and scatter (light green) plots of the projection of two latent variables of the five dimensional latent space of VED in combination with the marginal distributions of the two latent variables. The first row represents the projection of Node 1 (Global temperature variations) onto all other four latent variables and itself. The second row shows the plots for Node 2 (Large-Scale variations along mid latitude storm tracks). The third row represents Node 3 (Shallow Convection). The forth and fifth row shows the plots for Node 4 (Mid latitude storm track) and Node 5 (Deep Convection). All plots are based on 50000 randomly selected samples from the test data. This Figure is directly reproduced from **Behrens** et al. 2022.

averages of surface air temperatures of around 295K in this projection. These temperatures are around 5K colder than the maximum of the conditional averages seen for this particular projection, which is in agreement with the original hypothesis. Overall these results indicate the power of the VED with respect to the interpretability and meaningfulness of the latent space and stored physical concepts in the lower-order manifold.

A.4. Alternative VED and cVAE Structure

A.4.1. VED$_{X\rightarrow Y}$

VED$_{X\rightarrow Y}$ closely mirrors the original SP with similar output variables to those of the reference ANN. It uses a set of convection related CAM climate variables **X** as input to the network, except for meridional wind profiles which were additionally used in Rasp et al. 2018. For this variational network, we couple the encoder to a regular feed forward neural net with 3 hidden layers. The resulting variational network VED$_{X\rightarrow Y}$ (Figure 3.1) reproduces the convection-related SP output variables **Y** used in Rasp et al. 2018. The concatenated output vector **Y** has a length of 65 (65 output nodes). It contains the vertical profiles of temperature **dT(p)/dt** and specific humidity tendencies **dq(p)/dt**, the shortwave / longwave fluxes at the model top / surface $\mathbf{Q}_{sw/lw\ top/surf}$ and the precipitation rate **precip**. The coupled decoding feed forward neural net has three hidden layers with 353 nodes in each layer. The associated loss function is given in Equation A.1.

$$\text{VED loss}_{\text{X}\rightarrow\text{Y}} = \text{reconstruction loss}_{\text{X}\rightarrow\text{Y}} + \lambda\ \text{KL loss} \tag{A.1}$$

The reconstruction loss (Equation A.2) of VED$_{X\rightarrow Y}$ is defined as the MSE between the emulated $\mathbf{Y}^{emul}$ and **Y**.

$$\text{reconstruction loss}_{\text{X}\rightarrow\text{Y}} = \frac{1}{M}\times\frac{1}{N}\sum_{i=1}^{M=65}\sum_{j=1}^{N=\text{batch size}}(Y_{ij}-Y_{ij}^{emul})^2 \tag{A.2}$$

$$\text{KL loss} = \frac{1}{2}\times\frac{1}{N}\sum_{j=1}^{N=\text{batch size}}\sum_{k=1}^{K=\text{latent space width}}\left[-1-\ln\sigma_{jk}^2+\mu_{jk}^2+\sigma_{jk}^2\right] \tag{A.3}$$

$$\lambda\ \epsilon\ \mathbb{R}_+ \tag{A.4}$$

The hyperparameters used for VED$_{X\rightarrow Y}$ are displayed in Table A.5, and the model architecture is illustrated in Figure A.15.

VED$_{X\rightarrow Y}$ (test MSE = 0.157) reproduces the mean statistics with increased skill compared to VED (test MSE = 0.165) using the VED output normalization. The emulation skill of the spatio-temporal tropical variability is of the order of that of VED and slightly reduced with respect to reference ANN. However we see a decreased interpretability of the latent space of VED$_{X\rightarrow Y}$ in comparison to VED, which is a major disadvantage of the VED$_{X\rightarrow Y}$ network architecture. The 2D PCA compressed latent space of VED$_{X\rightarrow Y}$ generally shows a weak minimum to maximum

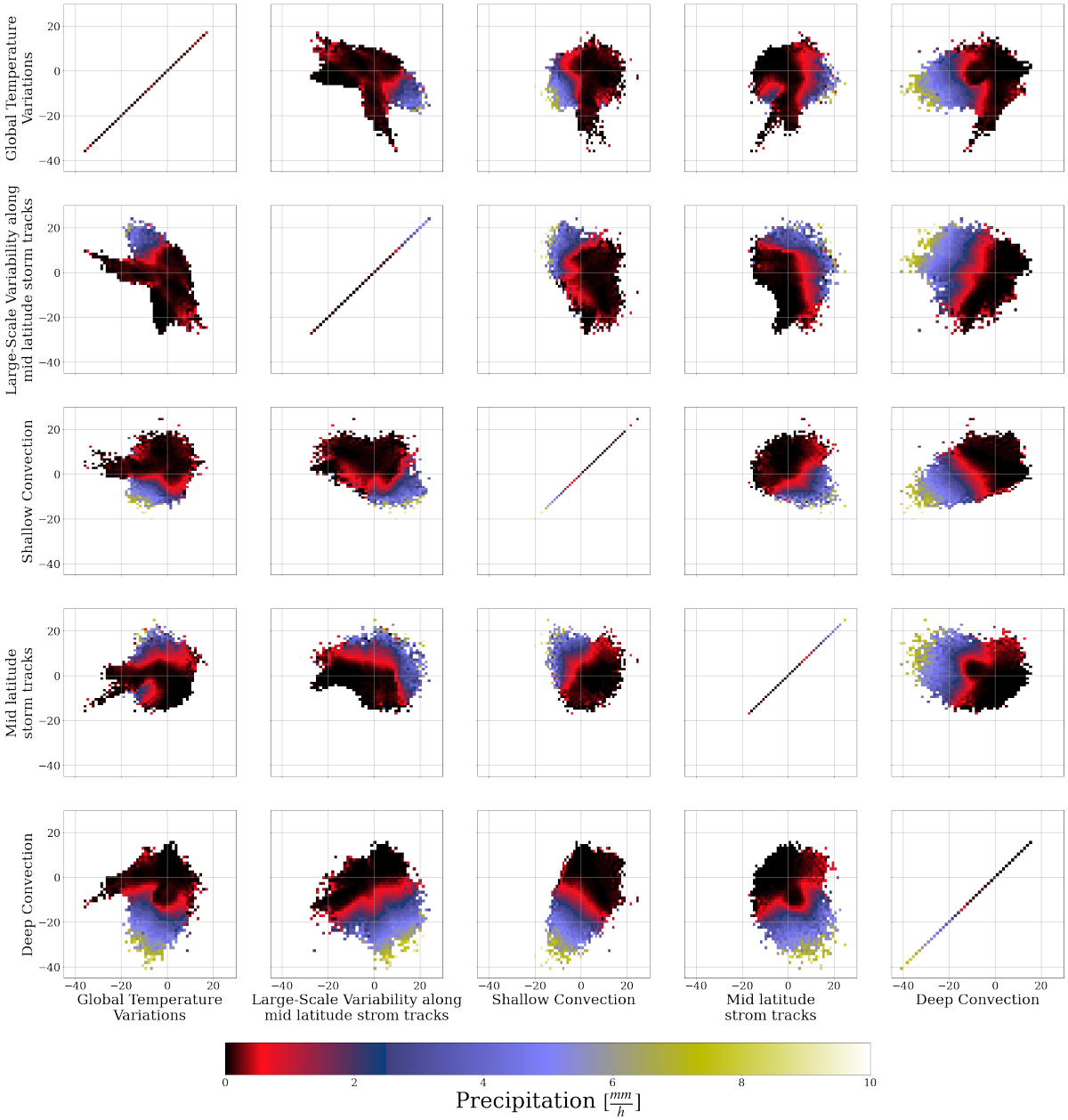

Figure A.12.: Similar to Figure A.11 a projection of one latent variable on all other latent variables and itself. The color coding reveals the conditional average of precipitation based on 1000000 randomly selected samples from the test data set. The first row shows the 2D projection of the first latent variable (Node 1, Global Temperature Variations) and all other latent variables. The second / third / fourth and fifth row depicts the projections of Node 2 (Large-Scale Variations along the mid latitude storm tracks) / Node 3 (Shallow Convection) / Node 4 (Mid latitude storm tracks) and Node 5 (Deep Convection). This Figure is directly reproduced from **Behrens** et al. 2022.

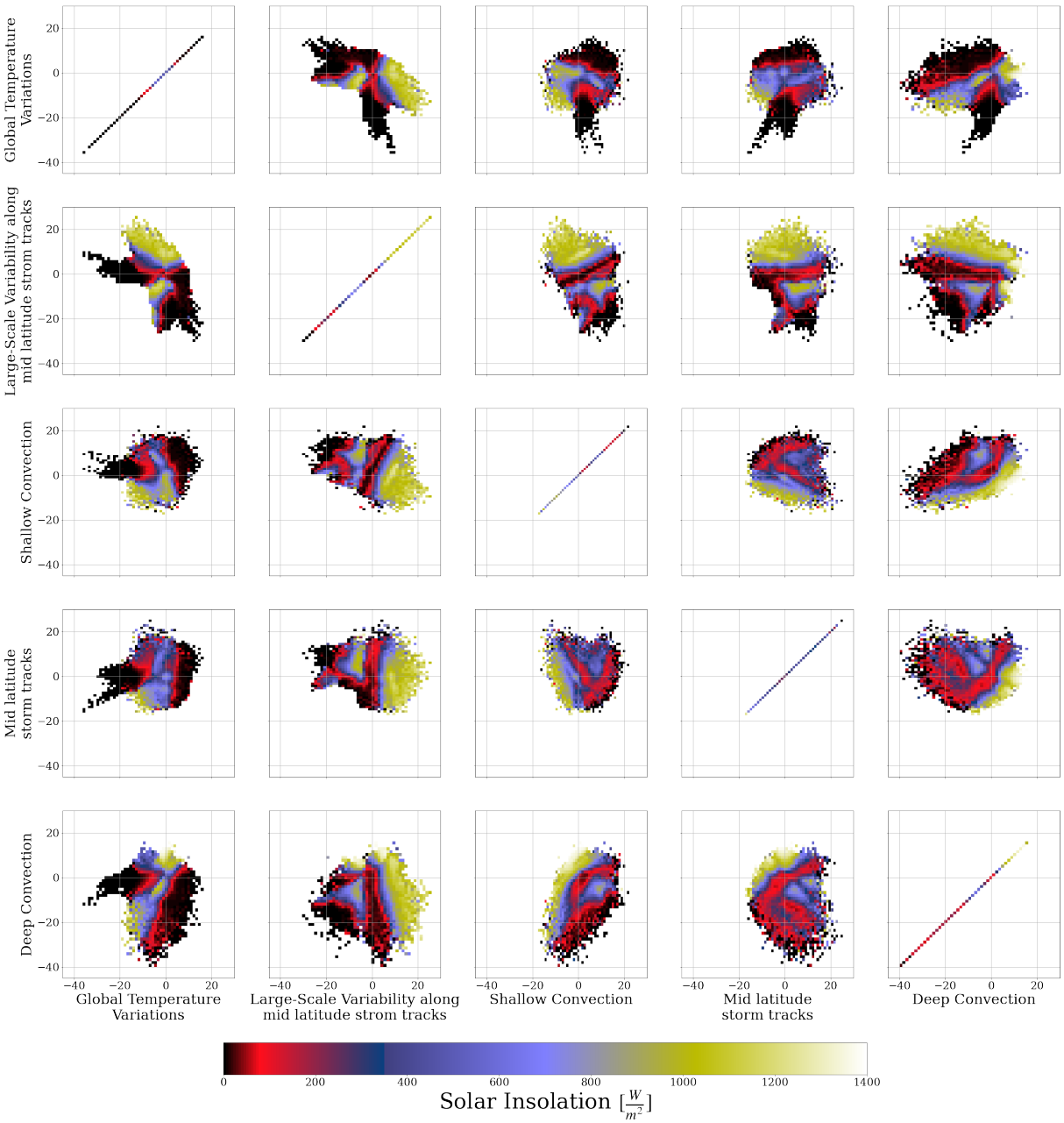

Figure A.13.: Similar to Figure A.12, but for the conditional averages of solar insolation in the projections. This Figure is directly reproduced from **Behrens** et al. 2022.

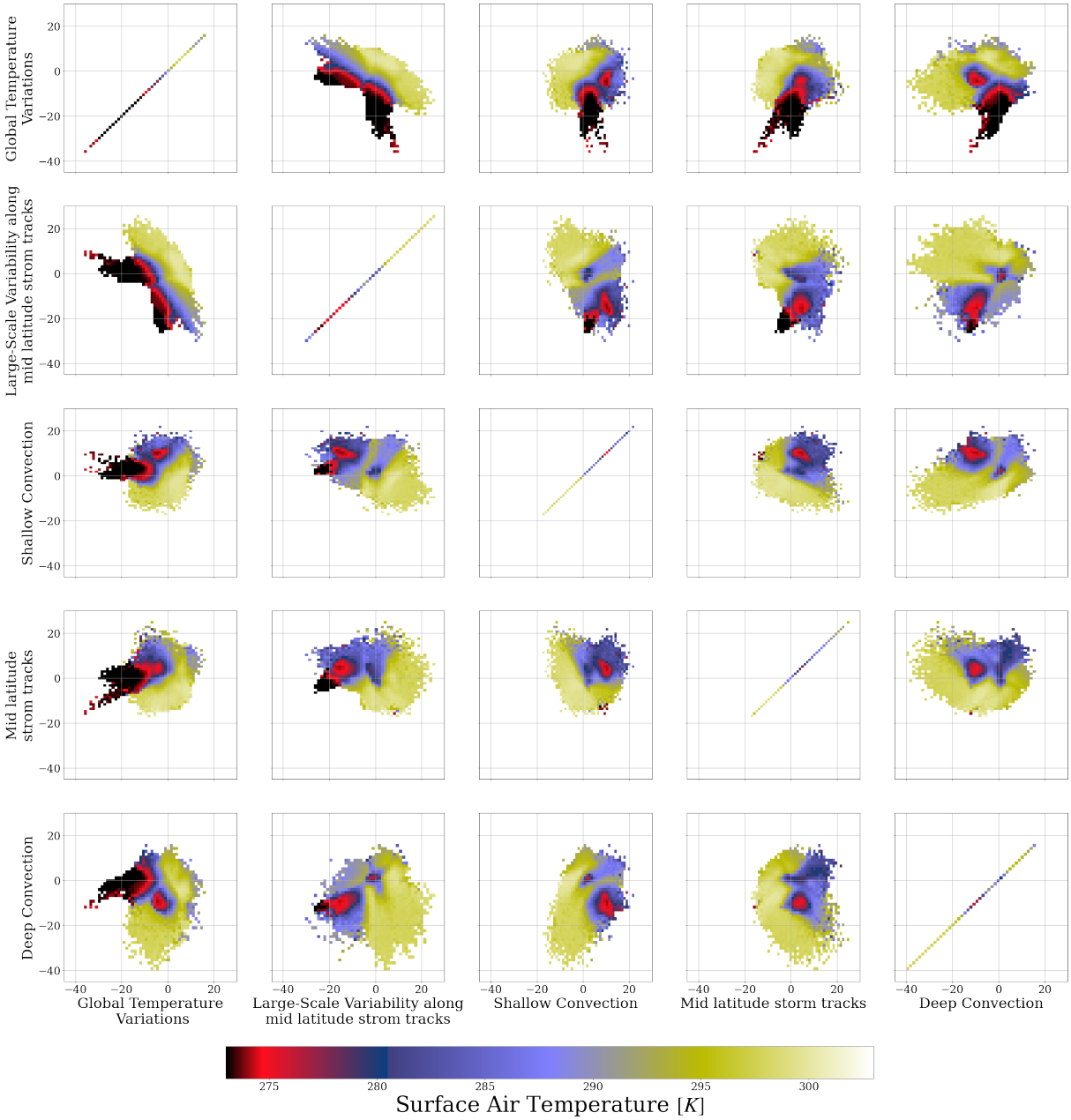

Figure A.14.: Similar to Figure A.12, but for the conditional averages of the surface air temperature in the projections. This Figure is directly reproduced from **Behrens** et al. 2022.

Hyperparameter $\mathbf{VED}_{X\rightarrow Y}$	**Values**
Learning Rate	0.00018238
Training / learning rate decrease	40 epochs, learning decrease every 7^{th} epoch by factor 5
Batch size	714
Latent Space Width	5 nodes
Node Size of Encoder	[64,463,463,232,116,58,29,5]
Node Size of Decoder [ANN]	[5,353,353,353,65]
Activation Encoder	[Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, Lambda]
Activation Decoder [ANN]	[Input, ReLU, ReLU, ReLU, ELU]
KL Annealing	Linear annealing from 2^{nd} to 7^{th} epoch

Table A.5.: Hyperparameters and architecture of the constructed $\text{VED}_{X\rightarrow Y}$ which uses large-scale CAM variables **X** to simulate SP variables **Y**. This Table is directly reproduced from **Behrens** et al. 2022.

distribution mostly focusing on the magnitude of convective processes (see Figure A.16) and faintly on geographic variability with respect to multiple subgrid-scale and large-scale climate variables (see Figure A.17, as an example for surface air temperatures). Samples from the two poles with anomalously cold surface air temperatures are not well separated in the 2D PCA compressed latent space of $\text{VED}_{X\rightarrow Y}$, in contrast to that seen for VED (see Figure A.17). We observe one surface air temperature minimum in the 2D PCA compressed latent space of $\text{VED}_{X\rightarrow Y}$. The minimum comprises samples from the austral high latitudes to the right and from boreal latitudes to the left. These low surface air temperatures are compressed within a very small fraction of the 2D PCA compressed latent space of $\text{VED}_{X\rightarrow Y}$ surrounded by mid-latitude temperatures in close distance. For VED we see a clearly improved adaption to these large-scale meridional temperature variations with well separated zones of austral and boreal polar samples. Likewise we see for VED that samples with increased precipitation are concentrated into two centers of action and the 2D PCA compressed latent space illustrates strong gradients with respect to conditional averages of precipitation, which is not the case for $\text{VED}_{X\rightarrow Y}$ (Figure A.16). This lack of interpretability of the latent space is a general limitation of $\text{VED}_{X\rightarrow Y}$ compared to VED or even ED for the identification of driving large-scale climate conditions and related convective processes globally.

A.4.2. cVAE

In general, a conditional Variational AutoEncoder Decoder (cVAE) predicts the distribution of a set of output variables conditioned on the input variables. The general model configuration of cVAE's enables the propagation of information about the state of output variables and also input variables through the latent space to the conditional decoder (Sohn et al. 2015). For the task to realistically reproduce **Y** and gain insights on the interpretability of the latent space,

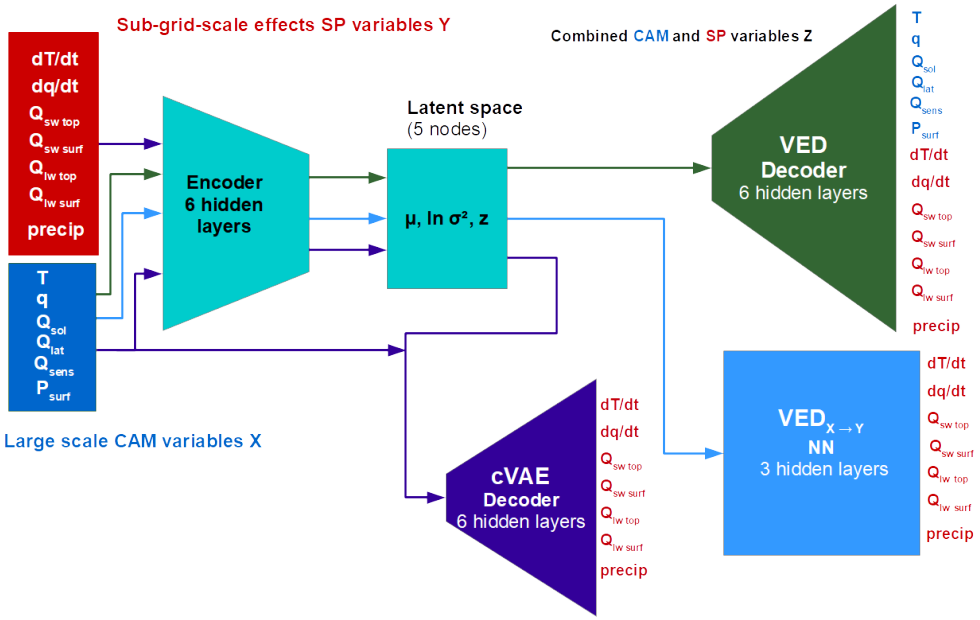

Figure A.15.: Combined schematic of the architecture of VED (green), $\text{VED}_{X\rightarrow Y}$ (light blue) and cVAE (purple arrows and network parts). The network structures in light blue are used for all variational networks with varying hyperparameters. This Figure is directly reproduced from **Behrens** et al. 2022.

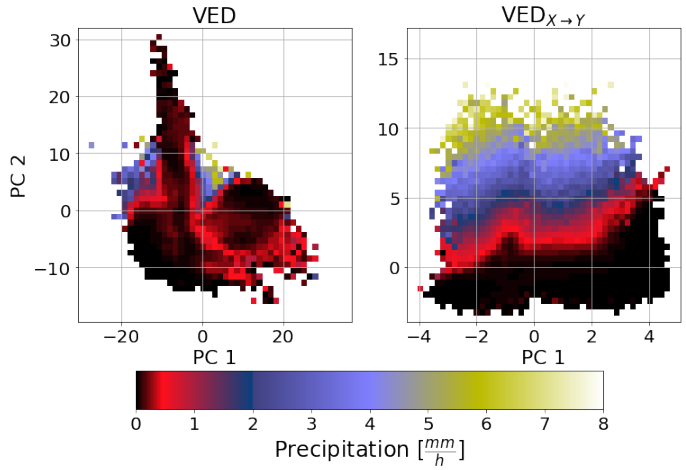

Figure A.16.: The 2D PCA compressed latent space of the VED (left) and $\text{VED}_{X\rightarrow Y}$ (right panel) and associated conditional average of precipitation of projected SP test data (similar to Figure 3.5). The x-axis / y-axis in all subplots indicates the 1^{st} / 2^{nd} leading PC of the 5D latent space in the respective panels. This Figure is directly reproduced from **Behrens** et al. 2022.

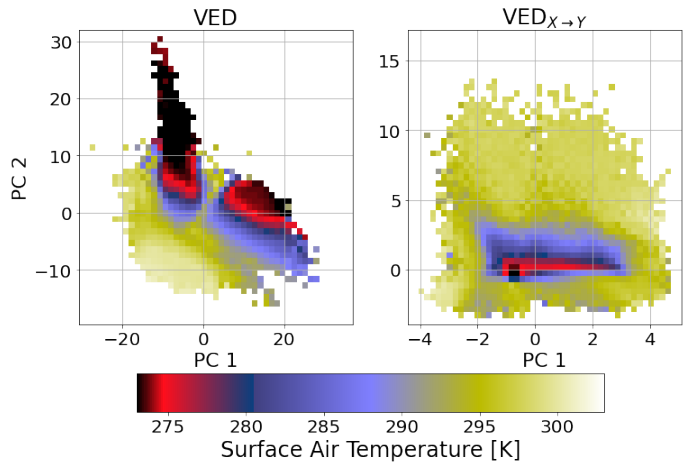

Figure A.17.: The 2D PCA compressed latent space of the VED (left) and VED$_{X\rightarrow Y}$ (right panel) and associated conditional average of surface air temperature of projected SP test data (similar to Figure 3.5). The x-axis / y-axis in all subplots indicates the 1^{st} / 2^{nd} leading PC of the 5D latent space in the respective panels. This Figure is directly reproduced from **Behrens** et al. 2022.

we construct one possible cVAE. The subgrid-scale variable vector **Y** is fed into the encoder together with large-scale CAM variables **X**, as can be seen in Figure A.15. **X** is an additional input to the decoding part of the network. As a result of that the latent space should illustrate a pronounced dependence on the subgrid-scale input features **Y** rather than on large-scale CAM variables **X**. The cVAE's loss function is defined as:

$$\text{cVAE loss} = \text{reconstruction loss}_{\text{cVAE}} + \lambda \text{ KL loss} \tag{A.5}$$

The associated reconstruction loss is defined as the MSE between $\mathbf{Y}^{emul}$ and **Y**, as can be seen in Equation 6.

$$\text{reconstruction loss}_{\text{cVAE}} = \frac{1}{M} \times \frac{1}{N} \sum_{i=1}^{M=65} \sum_{j=1}^{N=\text{batch size}} (Y_{ij} - Y_{ij}^{emul})^2 \tag{A.6}$$

$$\text{KL loss} = \frac{1}{2} \times \frac{1}{N} \sum_{j=1}^{N=\text{batch size}} \sum_{k=1}^{K=\text{latent space width}} [-1 - \ln \sigma_{jk}^2 + \mu_{jk}^2 + \sigma_{jk}^2] \tag{A.7}$$

$$\lambda \, \epsilon \, \mathbb{R}_+ \tag{A.8}$$

The used hyperparameters are displayed in Table A.6 and the model architecture can be seen in Figure 3.1.

Due to its deviating model architecture in comparison to the constructed VEDs (VED and VED$_{X\rightarrow Y}$), which are not trained with SP subgrid-scale variables **Y** as input data, the cVAE has an advantage against all evaluated models in training mode (during the model optimization). This advantage in training mode reflects in a strongly improved emulation skill of this network compared to the reference ANN. The MSE of cVAE in training mode with respect to SP training, validation or test data (0.049 / 0.050 / 0.050) is more than half as small as the one

Hyperparameter cVAE	Values
Learning Rate	0.00096133
Training / learning rate decrease	40 epochs, learning decrease every 7^{th} epoch by factor 5
Batch size	666
Latent Space Width	5 nodes
Node Size of Encoder	[[65,64],457,457,228,114,57,29,5]
Node Size of Decoder	[5,29,57,114,228,457,457,65]
Activation Encoder	[Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, Lambda]
Activation Decoder	[Input, ReLU, ReLU, ReLU, ReLU, ReLU, ReLU, ELU]
KL Annealing	Linear annealing from 2^{nd} to 7^{th} epoch

Table A.6.: Hyperparameters and architecture of the constructed conditional Variational AutoEncoder Decoder (cVAE) which uses subgrid-scale SP variables **Y** and large-scale CAM variables **X** to simulate subgrid-scale SP variables **Y**. This Table is directly reproduced from **Behrens** et al. 2022.

of reference ANN (0.133 / 0.135 / 0.135) using the VED output normalization. We observe similar emulation capabilities for the related coefficients of determination R^2 of the lower tropospheric specific humidity and temperature tendencies. More than 96% of the horizontal grid points have a R^2 value larger than 0.7 for 700 hPa temperature tendencies in the case of the cVAE in training mode. For cVAE, only 38% of the grid points exceed a coefficient of determination of 0.7. Nevertheless the emulation capabilities in test mode, where only the CAM climate variables **X** are fed into cVAE, are remarkably weaker than for all other evaluated networks. This is one clear disadvantage of the "brute-force training strategy" of the cVAE with our architecture, where we train the encoder and decoder together. The strong decrease in emulation skill between training and test mode suggests that the largest portion of optimization goes into the emulation of the subgrid-scale variables **Y**. Another discouraging point is the overall poorly developed interpretability of the latent space of cVAE with respect to essential subgrid-scale and climate variables like outgoing longwave radiation, solar insolation or surface air temperature. cVAE is not capable to distinguish between day and night-time conditions in its latent space. This is a crucial benchmark of all other evaluated models. Overall, cVAE focuses in its latent space exclusively on variations in convective moistening and heating tendencies or the related formation of precipitation. This clearly limits the interpretability of drivers of convective predictability in the latent space of cVAE. Furthermore it suggests that key information about the background climate state of convective processes are dominantly propagated trough the additional link of **X** to the decoder of cVAE (see Figure A.15). This leads to the fact that the encoding of large-scale information in the latent space of cVAE in training mode is clearly outperformed by a traditional PCA on the climate variables **X**. Despite these discouraging results, we think cVAE could be upgraded towards a generative and stochastic parameterization of SP. Pan et al. 2022 described that their initial cVAE structure exhibited large differences in performance between the training and test

Latent Node 1

Global Temperature variations	10^{th} perc	25^{th} perc	50^{th} perc	75^{th} perc	90^{th} perc
$\mathbf{Q}_{sw\ top}$ $[\frac{W}{m^2}]$	4	115	451	36	6
$\mathbf{Q}_{sw\ surf}$ $[\frac{W}{m^2}]$	-1	49	284	24	1
$\mathbf{Q}_{lw\ top}$ $[\frac{W}{m^2}]$	181	221	241	260	275
$\mathbf{Q}_{lw\ surf}$ $[\frac{W}{m^2}]$	55	12	28	60	44
precip $[\frac{mm}{h}]$	0.11	0.01	0.03	0.12	0.07
$\mathbf{P}_{surf}$ [hPa]	933	982	995	995	989
$\mathbf{Q}_{sol}$ $[\frac{W}{m^2}]$	15	263	748	45	8
$\mathbf{Q}_{sens}$ $[\frac{W}{m^2}]$	25	9	3	9	12
$\mathbf{Q}_{lat}$ $[\frac{W}{m^2}]$	52	19	39	85	163

Table A.7.: Generated shortwave and longwave heat flux at the model top and surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of $\mathbf{z}_{median}$ (4^{th} column, 50^{th} perc) and $\mathbf{z}_{translation}$ of the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile of latent node 1 (Global Temperature variations). This Table is directly reproduced from **Behrens** et al. 2022.

mode too. Therefore they developed a step-wise concept, where first the decoder is trained on **X**, then the encoder on **Y** and later the entire network on the complete variable list **O**. With this concept of step-wise training they were able to drastically improve the emulation abilities of their cVAE (Pan et al. 2022). In our case this upgraded training strategy might result in an enhanced interpretability of the latent space of cVAE with respect to large-scale drivers of convective predictability similar to results shown in this study for VED.

A.5. Generated SP/CAM Variables with $\mathbf{z}_{translation}$ / $\mathbf{z}_{median}$ and Squared Pearson Correlation $\mathbf{R}^2$ Plots between Latent Nodes and Vertical Profiles

This section comprises the Tables A.7-A.11 of generated 2D variables in **X** and **Y** for each latent node with our generative modeling approach. Additionally the squared Pearson correlation R^2 between the Nodes 1 to 5 and vertical profiles of **dq/dt**, **dT/dt**, **q** and **T** are displayed for space-time series (Figure A.18) or time series (Figure A.19) respectively. For Figure A.18 the Pearson correlation is computed based on the concatenated space-time series (with the shape [horizontal grid-cells **H** × time steps **P**, latent space width **K** or output variable size **M**]) of the latent nodes and profiles in **O**, which means that these arrays include information about the large-scale geographic variability, e.g. the large meridional temperature and specific humidity contrasts between the tropics and poles. For Figure A.19 the Pearson correlation is calculated in each horizontal grid-cell between the time series (with the shape [**P**, **K** or **M**] of the latent nodes and output profiles in **O**. As a second step the median of the Pearson correlation coefficients is calculated across all horizontal grid-cells **H**.

Latent Node 2

Large-scale variability along the mid latitude storm tracks	10^{th} perc	25^{th} perc	50^{th} perc	75^{th} perc	90^{th} perc
$\mathbf{Q}_{sw\ top}$ $[\frac{W}{m^2}]$	987	1092	451	57	158
$\mathbf{Q}_{sw\ surf}$ $[\frac{W}{m^2}]$	773	845	284	30	49
$\mathbf{Q}_{lw\ top}$ $[\frac{W}{m^2}]$	252	249	241	205	173
$\mathbf{Q}_{lw\ surf}$ $[\frac{W}{m^2}]$	85	73	28	44	13
precip $[\frac{mm}{h}]$	-0.01	0.00	0.03	0.12	0.15
$\mathbf{P}_{surf}$ [hPa]	983	989	995	993	992
$\mathbf{Q}_{sol}$ $[\frac{W}{m^2}]$	1214	1347	748	125	443
$\mathbf{Q}_{sens}$ $[\frac{W}{m^2}]$	19	9	3	6	23
$\mathbf{Q}_{lat}$ $[\frac{W}{m^2}]$	83	55	39	86	101

Table A.8.: Generated shortwave and longwave heat flux at the model top and surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of $\mathbf{z}_{median}$ (4^{th} column, 50^{th} perc) and $\mathbf{z}_{translation}$ of the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile of latent node 2 (Large-scale variability along mid latitude storm tracks). This Table is directly reproduced from **Behrens** et al. 2022.

Latent Node 3

Shallow Convection	10^{th} perc	25^{th} perc	50^{th} perc	75^{th} perc	90^{th} perc
$\mathbf{Q}_{sw\ top}$ $[\frac{W}{m^2}]$	5	134	451	1200	1112
$\mathbf{Q}_{sw\ surf}$ $[\frac{W}{m^2}]$	2	49	284	925	838
$\mathbf{Q}_{lw\ top}$ $[\frac{W}{m^2}]$	178	220	241	251	255
$\mathbf{Q}_{lw\ surf}$ $[\frac{W}{m^2}]$	3	8	28	73	72
precip $[\frac{mm}{h}]$	0.08	0.05	0.03	0.04	0.05
$\mathbf{P}_{surf}$ [hPa]	985	999	995	991	991
$\mathbf{Q}_{sol}$ $[\frac{W}{m^2}]$	15	329	748	1488	1468
$\mathbf{Q}_{sens}$ $[\frac{W}{m^2}]$	-15	-11	3	30	51
$\mathbf{Q}_{lat}$ $[\frac{W}{m^2}]$	-33	-13	39	207	242

Table A.9.: Generated shortwave and longwave heat flux at the model top and surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of $\mathbf{z}_{median}$ (4^{th} column, 50^{th} perc) and $\mathbf{z}_{translation}$ of the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile of latent node 3 (Shallow Convection). This Table is directly reproduced from **Behrens** et al. 2022.

Latent Node 4

Mid latitude frontal systems	10^{th} perc	25^{th} perc	50^{th} perc	75^{th} perc	90^{th} perc
$\mathbf{Q}_{sw\ top}$ $[\frac{W}{m^2}]$	440	456	451	435	442
$\mathbf{Q}_{sw\ surf}$ $[\frac{W}{m^2}]$	317	317	284	266	270
$\mathbf{Q}_{lw\ top}$ $[\frac{W}{m^2}]$	175	202	241	224	215
$\mathbf{Q}_{lw\ surf}$ $[\frac{W}{m^2}]$	59	53	28	39	43
precip $[\frac{mm}{h}]$	0.00	0.00	0.03	0.15	0.25
$\mathbf{P}_{surf}$ [hPa]	1001	1000	995	990	986
$\mathbf{Q}_{sol}$ $[\frac{W}{m^2}]$	625	678	748	741	746
$\mathbf{Q}_{sens}$ $[\frac{W}{m^2}]$	-5	-4	3	8	15
$\mathbf{Q}_{lat}$ $[\frac{W}{m^2}]$	29	27	39	75	97

Table A.10.: Generated shortwave and longwave heat flux at the model top and surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of $\mathbf{z}_{median}$ (4^{th} column, 50^{th} perc) and $\mathbf{z}_{translation}$ of the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile of latent node 4 (Mid latitude frontal systems). This Table is directly reproduced from **Behrens** et al. 2022.

Latent Node 5

Deep Convection	10^{th} perc	25^{th} perc	50^{th} perc	75^{th} perc	90^{th} perc
$\mathbf{Q}_{sw\ top}$ $[\frac{W}{m^2}]$	206	577	451	172	7
$\mathbf{Q}_{sw\ surf}$ $[\frac{W}{m^2}]$	80	327	284	109	0
$\mathbf{Q}_{lw\ top}$ $[\frac{W}{m^2}]$	188	208	241	254	266
$\mathbf{Q}_{lw\ surf}$ $[\frac{W}{m^2}]$	24	26	28	93	113
precip $[\frac{mm}{h}]$	0.60	0.24	0.03	0.01	-0.01
$\mathbf{P}_{surf}$ [hPa]	989	989	995	999	998
$\mathbf{Q}_{sol}$ $[\frac{W}{m^2}]$	489	1036	748	264	4
$\mathbf{Q}_{sens}$ $[\frac{W}{m^2}]$	4	2	3	3	6
$\mathbf{Q}_{lat}$ $[\frac{W}{m^2}]$	57	51	39	64	80

Table A.11.: Generated shortwave and longwave heat flux at the model top and surface, precipitation, surface pressure, solar insolation, sensible and latent heat flux of $\mathbf{z}_{median}$ (4^{th} column, 50^{th} perc) and $\mathbf{z}_{translation}$ of the 10^{th}, 25^{th}, 75^{th} and 90^{th} percentile of latent node 5 (Deep Convection). This Table is directly reproduced from **Behrens** et al. 2022.

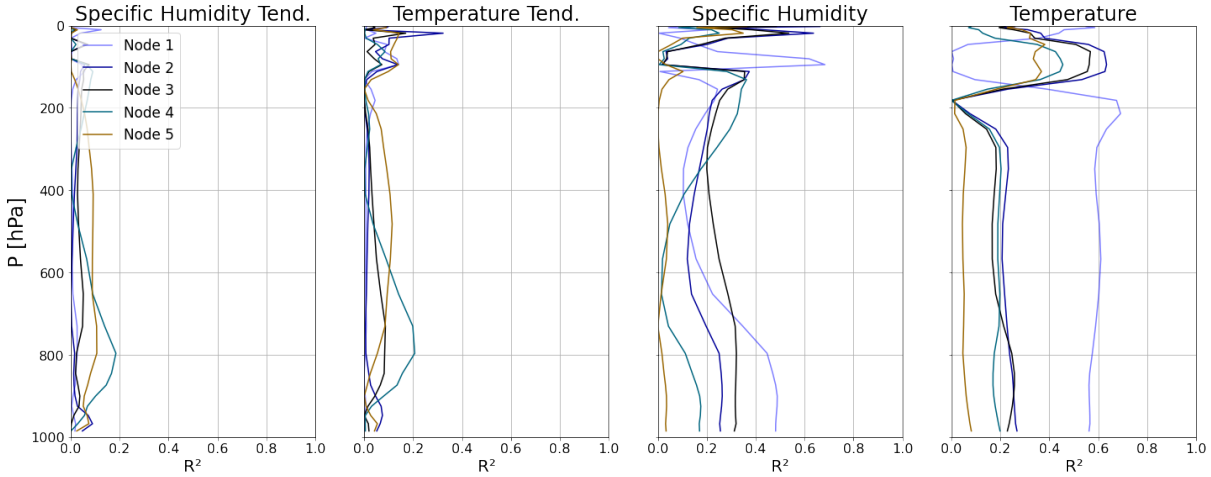

Figure A.18.: Squared Pearson correlation coefficient (linear explained variance) R^2 between the latent nodes of VED and predicted vertical profiles of specific humidity tendency (**dq/dt**), temperature tendency (**dT/dt**), specific humidity (**q**) and temperature (**T**) in space-time (which features the large meridional gradients of **q** and **T**). The light blue line resembles the R^2 value for latent node 1 / Global Temperature variations. The dark blue / black / dark cyan / bronze curve denotes the explained variance of latent node 2 (Large-scale variability along storm tracks) / 3 (Shallow Convection) / 4 (Mid latitude frontal system) / 5 (Deep Convection). This Figure is directly reproduced from **Behrens** et al. 2022.

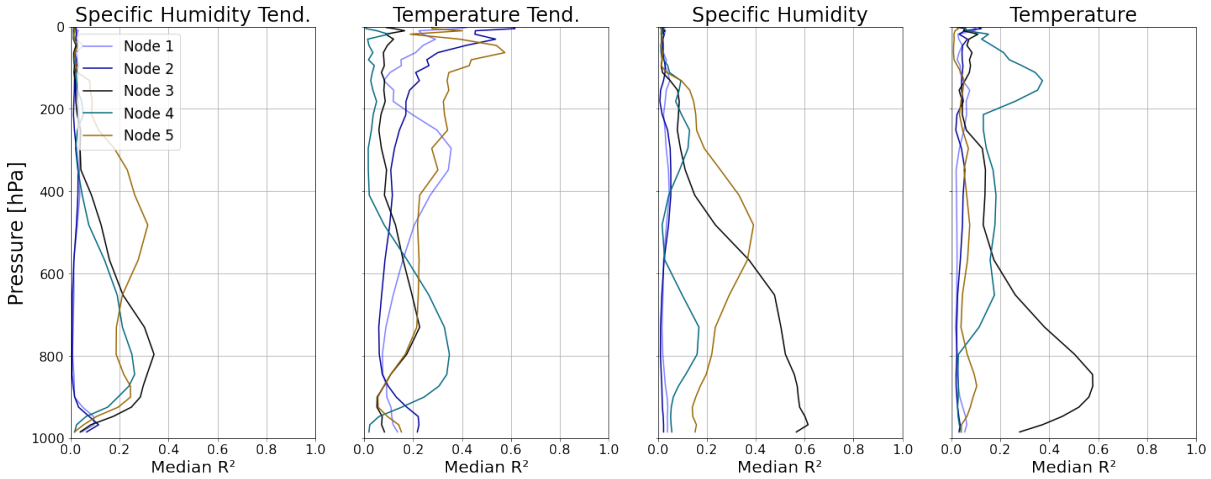

Figure A.19.: Median Squared Pearson correlation coefficient (linear explained variance) R^2 between the latent nodes of VED and predicted vertical profiles of specific humidity tendency (**dq/dt**), temperature tendency (**dT/dt**), specific humidity (**q**) and temperature (**T**) in time (without large meridional gradients of **q** and **T**). The light blue line resembles the median R^2 value for latent node 1 / Global Temperature variations. The dark blue / black / dark cyan / bronze curve denotes the median explained variance of latent node 2 (Large-scale variability along storm tracks) / 3 (Shallow Convection) / 4 (Mid latitude frontal systems) / 5 (Deep Convection). This Figure is directly reproduced from **Behrens** et al. 2022.

B. Supporting materials for Chapter 4 and 5: Improving Atmospheric Processes in Earth System Models with Deep Learning Ensembles and Stochastic Parameterizations

This section is reproduced from the supporting material of my paper that is currently in review (**Behrens** et al. 2024) with minor modifications of the naming of different deep learning models and ensembles. All Figures and Tables were produced from me as author of the thesis. Moreover I led the writing of the text for the supporting material that is shown in this appendix.

B.1. Introduction

Section B.2 describes the overall network configurations and Normalization. This section explains the hyperparameter tuning (subsection B.2.1), the Input and Output normalization (subsection B.2.2) and shows the hyperparameter of the best-performing ANNs (subsection B.2.3) and VEDs (subsection B.2.4). Section B.3 contains all supporting Figures with respect to the deterministic metrics. Section B.4 includes the additional Figures with respect to the ensemble or uncertainty metrics. Section B.5 describes our approach to find a suitable value for the applied latent space perturbation α with a static magnitude or a varying magnitude that is varying across the latent dimensions. We use VED 1 here as a baseline model. Section B.6 contains all supporting Figures related to the online runs.

B.2. Network Configurations and applied Normalizations

B.2.1. Hyperparameter Tuning

We conducted hyperparameter tuning experiments for two model types Artificial Neural Networks (ANN) and Variational Encoder Decoder structures (VEDs). For the ANNs we tested in total 116 suitable configurations. We run the ANNs over 15 epochs with two learning rate steps after the 5^{th} and 10^{th} epoch by dividing the initial learning rate by factor 5 and 25. We use Adam (Kingma and Ba 2014) as optimizer during the training. We use the same training and validation sets as in the main text, the first 7 consecutive days from each month ofthe year 2013 (training) / 2014 (validation). We selected the validation mean-square error of the subgrid SP variables $\boldsymbol{Y}$ as our hyperparameter optimization objective. We further saved

Hyperparameter range of ANNs	Values
Initial learning rate	10^{-4} to 5×10^{-3}
Batch size	200 to 13824
Activation function of hidden layers	ReLU, ELU, leaky ReLU, Tanh
Node Size of hidden layers	200 to 500
Depth of ANNs in hidden layers	4 to 8 hidden layers

Table B.1.: Hyperparameter range of the search for skilful ANNs, which reproduce SP subgrid variables $\boldsymbol{Y}$ with large-scale CAM variables and CAM precip $\boldsymbol{X}$ as input data set. The hyperparameter search was conducted over 116 trials and 15 epochs with a learning rate decrease after every 5^{th} epoch by a factor of 5. This Table is reproduced with minor modifications from **Behrens** et al. 2024.

the accuracy and mean absolute error as additional evaluation metrics for the validation and training data. Table B.1 details the hyperparameter and the associated range / options we tested.

We observed that the most sensitive hyperparameters are the initial learning rate and the activation function of the hidden layers, where ELU overall had the best performance.

For the Variational Encoder Decoder structures we conducted a similar hyperparameter tuning experiment. One major difference to the ANN is the presence of a latent space (lower-dimensional) space between the encoding and decoding part of the network. The latent space width is one of the main tuning parameter of these networks, like it was shown in **Behrens** et al. 2022. For the VEDs we prescribed the dimensionality reduction or expansion in the Encoder and Decoder. The fourth or third last hidden layer of the Encoder in front of the latent space has the half of the original node size of the first hidden layer, if the Encoder has more than 4 hidden layers or 4 hidden layer. The third or second last hidden layer of the Encoder in front of the latent space has a quarter of the original node size if the Encoder has more than 4 hidden layer or 4 hidden layers. While the second last or last hidden layer of the Encoder in front of the latent space has an eight of the original node size, if the Encoder has more than 4 hidden layers or 4 hidden layers. Additionally if the Encoder has more than 4 hidden layers, the last hidden layer consists of $\frac{1}{16}$ of Nodes of the first hidden layer.

The Decoder is mirroring the Encoder with an increase from the first to the third or fourth hidden layer from an eight or a $\frac{1}{16}^{th}$ to a half of the node size of the last hidden layer of the Decoder.

VEDs have an additional KL loss term in their loss function. We chose a static KL regularization term to use it as an additional hyperparameter for the network configuration which gives us an active tuning knob to score a suitable balance between reconstruction (here a mean square error (mse) loss is used) and the KL loss term.

As an objective for the hyperparameter tuning of the VEDs we set the validation loss (sum of reconstruction loss and annealed kl loss, see Equation 4.5). The learning rate schedule, Adam (Kingma and Ba 2014) as optimizer and the training over 15 epochs, the training and validation set is the same as before. In total we conduct 60 trial with varying hyperparamters. Table B.2 shows the evaluated hyperparameters for the VEDs and the associated ranges.

Hyperparameter range of VEDs	Values
Initial learning rate	10^{-4} to 5×10^{-3}
λ kl regularization coefficient	5×10^{-5} to 10^{-3}
Batch size	200 to 13824
Latent Space Width	2 to 15 latent nodes
Activation function of hidden layers	ReLU, ELU, leaky ReLU, Tanh
Node Size of first / last hidden layers of Encoder / Decoder	200 to 500
Depth of VEDs in hidden layers	5 to 6 hidden layers

Table B.2.: Hyperparameter range of search for skilful Variational Encoder Decoder (VED) structures, which reproduce SP subgrid variables $\boldsymbol{Y}$ with large-scale CAM variables and CAM precip $\boldsymbol{X}$ as input data set. The hyperparameter search was conducted over 60 trials and 15 epochs with a learning rate decrease after every 5^{th} epoch by a factor of 5. This Table is directly reproduced from **Behrens** et al. 2024.

We observed that the initial learning rates, the latent space width in combination with the kl regularization factor λ are the most sensitive hyperparameters. A larger latent space width in combination with a smaller λ is beneficial for the overall network performance with our approach.

B.2.2. Input, Output normalization and computation of tendency terms before coupling

Regarding the used Input and Output normalization, we built up on existing knowledge and experience from previous papers (**Behrens** et al. 2022; Rasp et al. 2018) when it comes to the normalization of large-scale CAM variables and CAM precipitation $\boldsymbol{X}$ (input) and subgrid SP variables $\boldsymbol{Y}$ (output normalization). Regarding the input normalization we used the same strategy as is presented in Rasp et al. 2018 or **Behrens** et al. 2022. We computed a longterm mean (84 days = period of training data set) for all variables and all levels. We subtract the mean array from each input data sample and divide the residuals by the range between longterm minimum and maximum anomaly. With this input normalization we constrain the normalized inputs $\boldsymbol{X}$ into the range of [-1,1].

For the output normalization we use a similar strategy as is presented in **Behrens** et al. 2022. We normalize the $\dot{q}$(p) profile by the longterm maximum standard deviation along the vertical axis (over two 2 months, June and July of 2013) of all levels, which comes from the surface layer. For $\dot{T}$(p) we found the maximum standard deviation also in the surface layer and used this value for the output normalization. For $\dot{q}_{cl}$(p) we used the standard deviation from 831 hPa (level 22) for the output normalization, while for $\dot{q}_{ci}$(p) from 244 hPa (level 14). The remaining 8 2D SP variables in $\boldsymbol{Y}$ were standardised accordingly.

Equation B.1 illustrates the general computation of the tendency terms before coupling for the example of $\dot{q}$(p). Herein QBC(p) is the vertical profile of specific humidity with the updates from SP but before the radiative adjustment and coupling to CLM5, QBP(p) is the

vertical profile of the specific humidity before calling SP and dt is the native CESM time step of 1800s.

$$\dot{q}(p) = \frac{\text{QBC(p)} - \text{QBP(p)}}{\text{dt}} \tag{B.1}$$

B.2.3. ANN ensemble: Hyperparameter of all ANNs

We evaluated the ANN hyperparameter tuning experiment and selected the 7 best - performing ANNs to form the base for our deterministic and stochastic ANN SP parameterizations ($\overline{\text{ANN}}$ and ANN-ensemble). Table B.3 shows the hyperparameter configuration of the 7 ANNs. We train all ANNs over 40 epochs with a learning rate decrease after every 7^{th} epoch by a factor of 5 using ADAM (Kingma and Ba 2014). The hyperparameter setting of ANN 1 are moreover used fro ANN-dropout

	ANN 1	**ANN 2**	**ANN 3**	**ANN 4**	**ANN 5**	**ANN 6**	**ANN 7**
Initial learning rate	6.16 $\times 10^{-4}$	3.36 $\times 10^{-4}$	4.82 $\times 10^{-4}$	4.72 $\times 10^{-4}$	12.62 $\times 10^{-4}$	13.73 $\times 10^{-4}$	4.74 $\times 10^{-4}$
Batch size	3551	9402	8833	9802	10740	11162	7800
Activation function	ELU	ELU	ELU	ELU	ELU	ELU	ELU
Activation function Output layer	Linear	Linear	Linear	Linear	Linear	Linear	Linear
Node Size	405	455	422	350	323	433	279
Depth ANNs [hid. lay.]	4	6	8	8	4	5	8

Table B.3.: Hyperparameters of the best-performing ANNs that form the base for the stochastic and deterministic ANN ensemble. This Table is reproduced with minor modifications from **Behrens** et al. 2024.

B.2.4. VED ensemble: Hyperparameter of all VEDs

Table B.4 shows the hyperparameter of the 7 best-performing VEDs, all except VED 6 (due to unstable behaviour especially on the test data set) form the "quasi-deterministic" VED ensemble ($\overline{\text{VED}}$). Additionally VED 1 is used as the example model, on which we apply our latent space perturbation approach (VED-varying, VED-static).

	VED 1	**VED 2**	**VED 3**	**VED 4**	**VED 5**	**VED 6***	**VED 7**
Initial learning rate	16.12 $\times 10^{-4}$	4.41 $\times 10^{-4}$	6.52 $\times 10^{-4}$	14.57 $\times 10^{-4}$	10.13 $\times 10^{-4}$	7.18 $\times 10^{-4}$	6.17 $\times 10^{-4}$
Batch size	9123	9047	8627	2313	4624	2770	8821
Activation function	ELU	ELU	ELU	leaky ReLU	ELU	ELU	leaky ReLU
kl weight λ	6.8 $\times 10^{-5}$	5.3 $\times 10^{-5}$	11.2 $\times 10^{-5}$	5.0 $\times 10^{-5}$	6.8 $\times 10^{-5}$	17.2 $\times 10^{-5}$	7.2 $\times 10^{-5}$
Encoder Node Size	[109,307, 307,154, 77,39, 20,13]	[109,411, 411,206, 103,52, 26,10]	[109,426, 426,213, 107,54, 27,9]	[109,359, 359,180, 90,45, 23,12]	[109,337, 337,169, 85,43, 22,13]	[109,411, 411,206, 103,52, 21,13]	[109,492, 246,123, 62,31,6]
Decoder Node Size	[13,20, 39,77, 154,307, 307,112]	[10,26, 52,103, 206,411, 411,112]	[9,27, 54,107, 213,426, 426,112]	[12,23, 45,90, 180,359, 359,112]	[13,22, 43,85, 169,337, 337,112]	[13,21, 52,103, 206,411, 411,112]	[6,31,62, 123,246, 492,112]
Depth Encoder / Decoder [hid. lay.]	6	6	6	6	6	6	5

Table B.4.: Hyperparameters of the 7 best-performing VEDs. The * denotes VED 6, which shows unstable behaviour on the validation and test data set if the model is trained over 40 epochs. Therefore we exclude this VED from the "quasi-deterministic" VED ensemble presented in the paper. This Table is directly reproduced from **Behrens** et al. 2024

B.3. Reproduction of subgrid convective processes with ensembles

Figure B.1 shows the median Coefficient of Determination (R^2) of the vertical profiles of $\dot{q}_{cl}$ and $\dot{q}_{ci}$ for all developed staochastic and

Figure B.2 depicts the latitude longitude plots of the coefficient of determination R^2 of $\dot{q}$ on 956 hPa for $\overline{\text{ANN}}$ (subplot a) and ANN-ensemble (b), ANN 1 (c) as an example of a single ANN realisation and ANN-dropout (d).

Figure B.3 shows the median coefficients of determination for the 2D SP precipitation and radiative fluxes for all approaches.

Figure B.4 shows the median MAEs of the vertical profiles of $\dot{q}$, $\dot{T}$, $\dot{q}_{cl}$, $\dot{q}_{ci}$ for ensemble and stochastic parameterizations.

Figure B.5 shows the median MAEs for the remaining 8 SP variables. Note that we used the original output normalized predictions and test data to compile this plot. The associated y-axis reflects therefore the median MAE with respect to the used standard deviations.

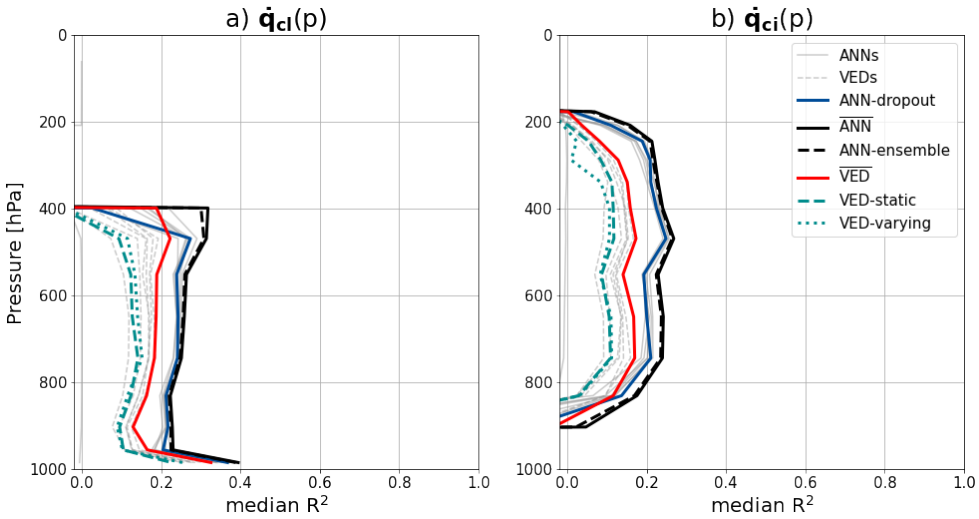

Figure B.1.: Vertical profiles of median Coefficient of Determination (R^2) for a) cloud liquid water tendency $\dot{q}_{cl}$, b) cloud ice water tendency $\dot{q}_{ci}$ of different individual ANNs and VEDs and in the background (grey solid and dashed lines), ANN-dropout (solid navy blue); $\overline{ANN}$ and ANN-ensemble (solid and dashed black), $\overline{VED}$ (solid red); VED-static (dashed cyan) and VED-varying (dotted cyan line). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

B.4. Uncertainty Estimates of subgrid convective processes with stochastic and deterministic ensembles

Figure B.6 shows the aggregated continuous rank probability score (CRPS) for all approaches over all SP variables $\boldsymbol{Y}$ with respect to the output loss dictionary. Figure B.7 shows the CRPS of all approaches with respect to $\dot{q}_{cl}$ on 831 hPa. Figures B.8 and B.9 illustrate the CRPS of surface $\dot{q}$ and $\dot{T}$.

Figure B.10 - B.12 depicts the probability integral transform histograms of upper tropospheric $\dot{q}_{ci}$, $\dot{q}_{cl}$ in the upper part of the planetary boundary layer, surface $\dot{q}$ and surface $\dot{T}$.

B.5. Hyperparameter tuning of the latent space perturbation α_i

To score a balance between reproduction skill and calibration of the ensemble spread based on a single VED (we select VED 1) with perturbation of the latent space (Approach 3), we conduct a further hyperparameter optimization. We compute the PIT distance (Equation B.2, following Haynes et al. 2023), where B is the number of bins in the PIT histogram, E_b is the number of samples within a distinct bin, E is the total number of evaluated samples and b is the ID of a distinct bin. We use the median of PIT distances of all SP variables as a first metric for the intra-ensemble spread.

$$\text{PIT distance} = [\frac{1}{B}\sum_{b=1}^{B}(\frac{E_b}{E} - \frac{1}{B})]^{\frac{1}{2}} \tag{B.2}$$

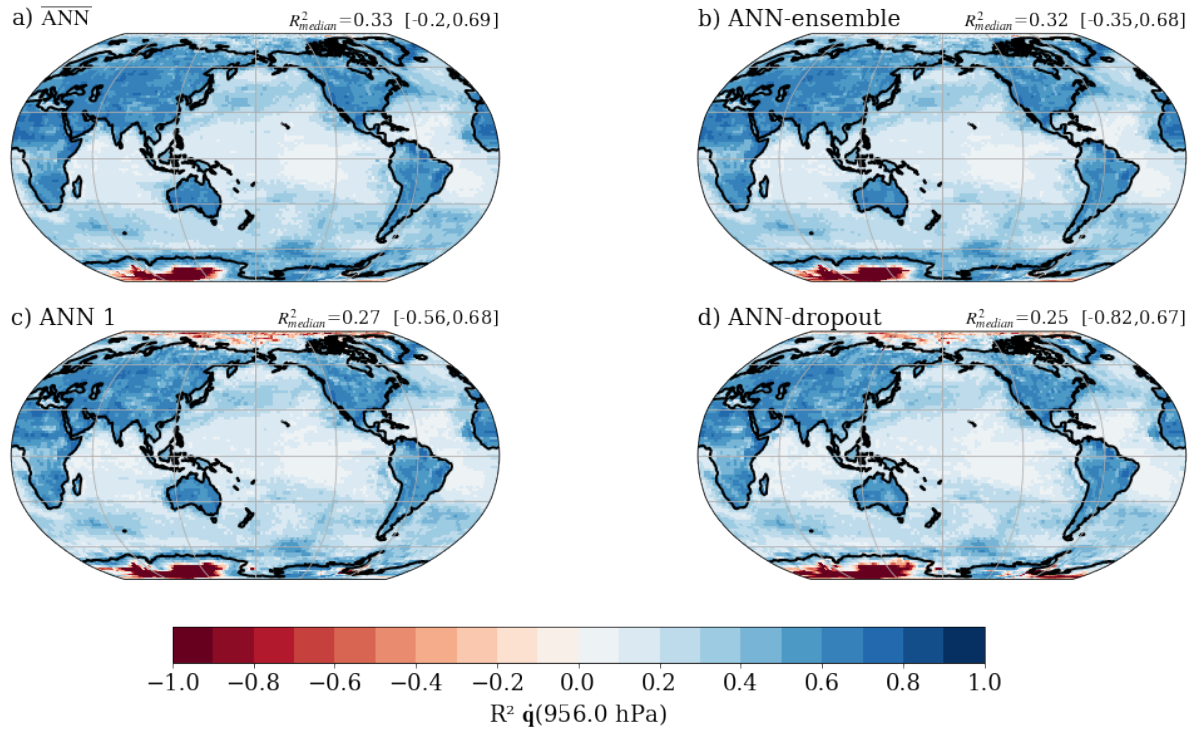

Figure B.2.: Coefficient of determination R^2 of the specific humidity tendency $\dot{\boldsymbol{q}}$ on 956 hPa of a) $\overline{\text{ANN}}$, b) ANN-ensemble, c) ANN 1 as an example of an individual member and d) ANN-dropout. The median, the 2.5^{th} and 97.5^{th} percentiles of the horizontal R^2 field are shown above each panel to the right. This Figure is reproduced with minor modifications from **Behrens** et al. 2024

The second metric is the median of all coefficients of determination R^2, computed this time over the concatenated space-time axis, of the SP variables $\boldsymbol{Y}$, which measures the reproduction skill.

These two metrics are complemented by the median and mean CRPS across all SP variables $\boldsymbol{Y}$ as third metric, which focuses both on the reproduction skill and the calibration of the ensemble spread.

These three metrics give us a robust toolbox to find a good magnitude of either a static latent perturbation α_i or varying α_i along all latent dimensions . We picked the VED 1 as an example to find both a suitable static α and varying α along its 13 latent dimensions. Therefore we selected 100 time steps (~ 1.4 million samples) and generated a 7 member ensemble, which is fed then into the decoder.

This step is then repeated a few times in an algorithm and all metrics are tracked for the respective static α_i or varying α_i arrays. As a first step we conduct a search for the static α_i between 0 and 1 using a step size of 0.1. For mean / median CRPS we found a global minimum between 0.1 and 0.3. The same is also true if we focus on the sum of 1 minus the median R^2 and median PIT difference, where we see a decrease until 0.15 to 0.4 and an increase afterwards, which is in line with the decay of reproduction skill with increasing degree of latent space perturbation.

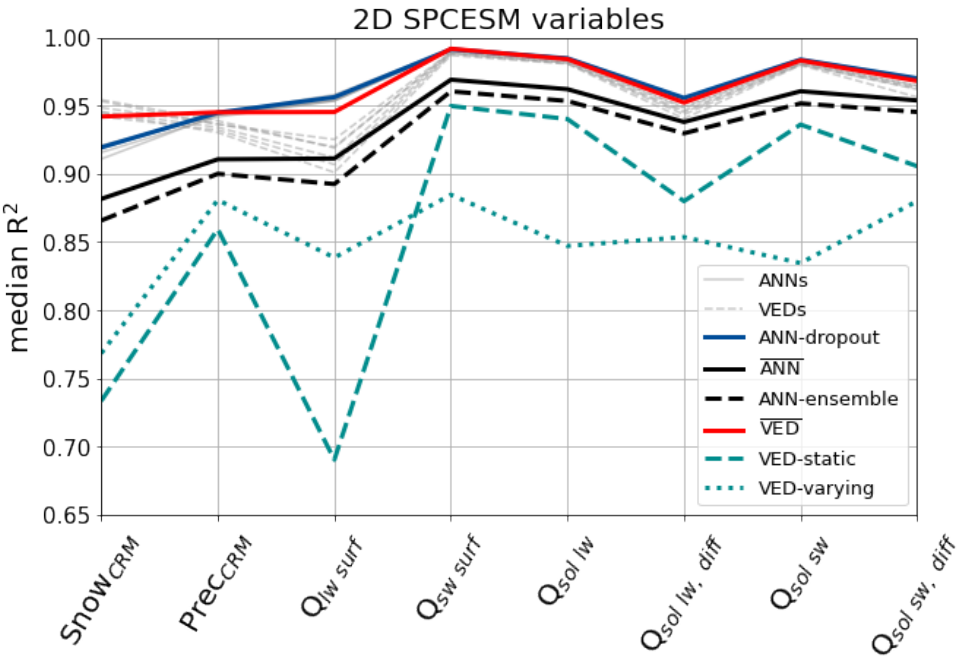

Figure B.3.: Median coefficient of determination R^2 for the remaining 8 2D output variables of different individual ANNs and VEDs in the background (solid and dashed grey lines), ANN-dropout (navy blue); $\overline{\text{ANN}}$ and ANN-ensemble (solid and dashed black line), $\overline{\text{VED}}$ (solid red line); VED-static (dashed cyan line) and VED-varying (dotted cyan line). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

As a second step we 'fine-search" the α-range between 0.1 and 0.4 with a stepping of 0.01. In this case the sum between 1 minus median R^2 and the median PIT distance has a minimum at α_i=0.40. For the median CRPS of all $\boldsymbol{Y}$ we find the minimum around α_i=0.36, while for the mean CRPS the minimum is located at 0.19. We test this approach also for higher and lower percentiles for CRPS and the sum term and do not find strong shifts of a suitable static α_i that provides a good balance between reproduction skill and the calibration of the spread. In general we find that a static α_i of around 0.2 to 0.4 provides an improved CRPS and PIT distances, while not dramatically reducing the prediction skill of VED 1.

For the varying α_i along the latent dimensions of VED 1 we conduct in total 2800 trials based on 50 randomly drawn time steps. Here we use first a range from 0 to 2.5 to randomly draw values for each α_i, where i is a distinct latent dimension. Later we reduce the range from 0 to 1, which results in an increase of CRPS, R^2 and PIT distance values. To evaluate the skill and to get the best performance, we search for those α_i arrays that have a median CRPS smaller than the 2.5^{th} percentile of all median CRPS values, and a median loss term based on PIT distance and R^2 also smaller than the overall 2.5^{th} percentile. We select two favourable α arrays out of the entire set, see Table B.5. We use α array 1, which was drawn in a pre-hyperparameter search where we only focused on improving the PIT distance, due to its improved CRPS and PIT compared to α array 2. Compared to the static α approach the varying α arrays have a smaller median CRPS with a comparable median loss term (1 - R^2

	α **array 1**	α **array 2**
alpha array	[0.09 0.52 0.07 0.73 0.4 0.33 0.77 0.29 0.95 0.61 0.73 0.84 0.35]	[0.25, 0.05, 0.25, 0.68, 0.77, 0.09, 0.61, 0.92, 0.02, 0.44, 0. , 0.15, 0.93]
median CRPS	0.0203	0.0201
mean CRPS	0.0453	0.0448
median R^2	0.266	0.320
median PIT distance	0.00144	0.00165
1 - median R^2 + median PIT distance	0.735	0.681

Table B.5.: Suitable α arrays for the perturbation of the latent space of VED 1. Illustrated are the α arrays and key performance metrics to put them into context with the static α approach. The metrics are computed over 100 randomly drawn time steps similar to Figure B.13 and B.15. This Table is directly reproduced from **Behrens** et al. 2024.

- PIT distance). This indicates an improved calibration of the intra-ensemble spread, which does not lead to a decay in reproduction skill.

B.6. Online results: Evaluation of developed stochastic and deterministic ensemble parameterizations and related benchmark parameterizations

Figure B.16 and Figure B.17 shows the time series of the mean RMSE of specific humidity respectively temperature below 200 hPa simulated with the developed deterministic, stochastic ensemble parameterizations and ANNs with respect to an independent run with a superparameterization in CESM2.

Figure B.18 shows the zonal averages of the temperature field for the period February - June 2013 with a superparameterization coupled to CESM, related differences between SP-CESM and our developed ensemble parameterizations and also the differences between SP-CESM and the CESM2 run with the Zhang-McFarlane scheme.

Figure B.19 shows the zonal averages of the specific humidity field for the period February - June 2013 with a superparameterization coupled to CESM, related differences between SP-CESM and our developed ensemble parameterizations and also the differences between SP-CESM and the CESM2 run with the Zhang-McFarlane scheme.

Figure B.20 shows the precipitation histograms of the developed deterministic and stochastic ensemble parameterizations in comparison to the superparameterization and the Zhang-McFarlane Scheme Zhang and McFarlane 1995 based on 10 million randomly drawn samples from the period February to June 2013.

Figure B.21 depicts the global maps of median precipitation of the CESM runs with the different parameterizations for the period February to June 2013.

Figure B.22 shows the regions on the globe that we select for the evaluation of the represented diurnal cycle of all parameterizations.

Figure B.23 shows the diurnal cycles of precipitation simulated with the superparameterization, the developed deterministic and stochastic ensemble parameterizations and the Zhang-McFarlane scheme over the regions illustrated in Figure B.22.

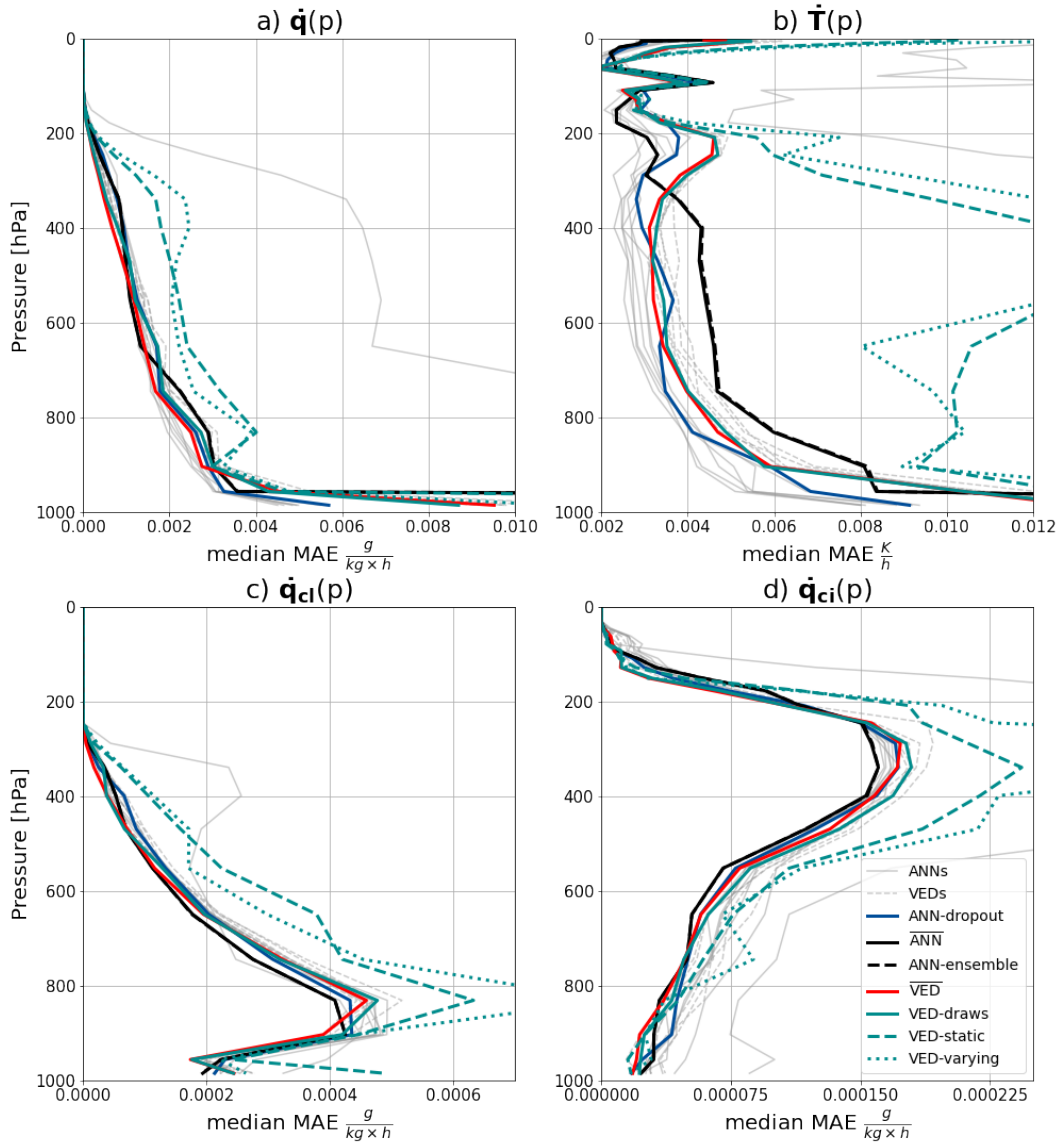

Figure B.4.: Vertical profiles of the median mean absolute error (MAE) for specific humidity tendency (a, $\dot{q}$), temperature tendency (b, $\dot{T}$), cloud liquid tendency (c, $\dot{q}_{cl}$) and cloud ice tendency (d, $\dot{q}_{ci}$) of the individual ANNs and VEDs in the background (grey), ANN-dropout (solid navy blue); $\overline{\text{ANN}}$ and ANN-ensemble (solid and dashed black), $\overline{\text{VED}}$ (solid red); VED-draws (solid cyan), VED-static (dashed cyan) and VED-varying (dotted cyan line). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

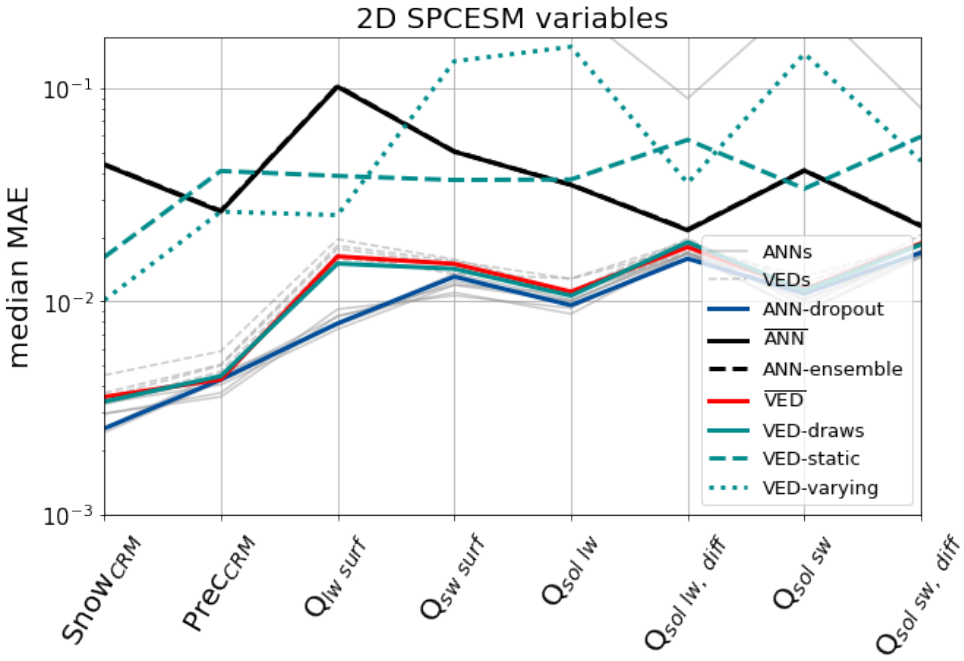

Figure B.5.: Median mean absolute error (MAE) of the 2D output variables of different individual ANNs and VEDs in the background (solid and dashed grey lines), ANN-dropout (navy blue) ; $\overline{\text{ANN}}$ and ANN-ensemble (solid and dashed black line), $\overline{\text{VED}}$ (solid red line); VED-draws (solid cyan line), VED-static (dashed cyan line) and VED-varying (dotted cyan line). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

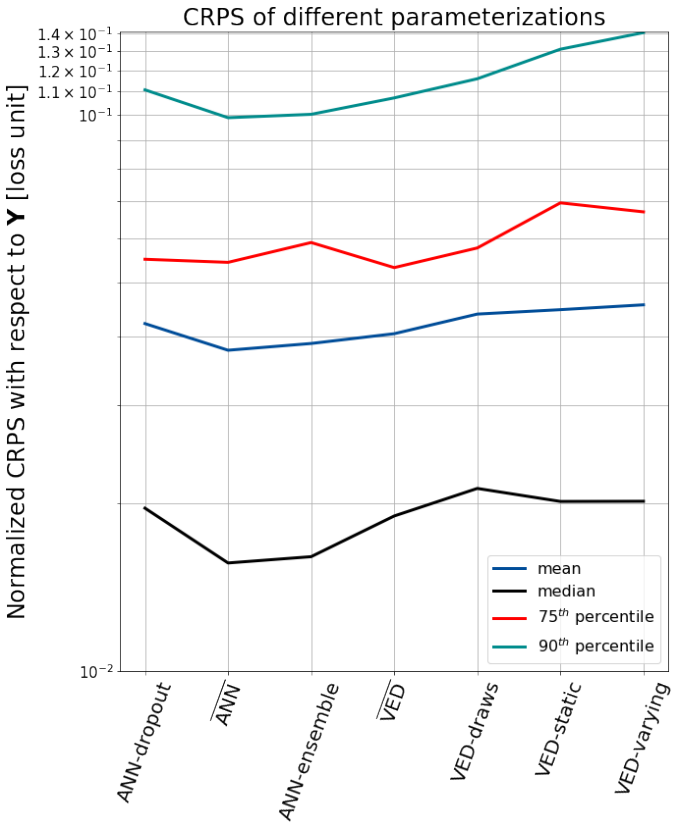

Figure B.6.: Aggregated Continuous Rank Probability Score (CRPS) for different stochastic and deterministic parameterizations. The blue line indicates the mean, the black line illustrates the median, the red line the 70^{th}, the cyan line the 90^{th} percentile computed over all SPCESM variables Y based on 500 randomly drawn time steps from test data. The y-axis illustrates the normalized CRPS loss and the evaluated parameterizations are shown along the x-axis with the respective name as tick label. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

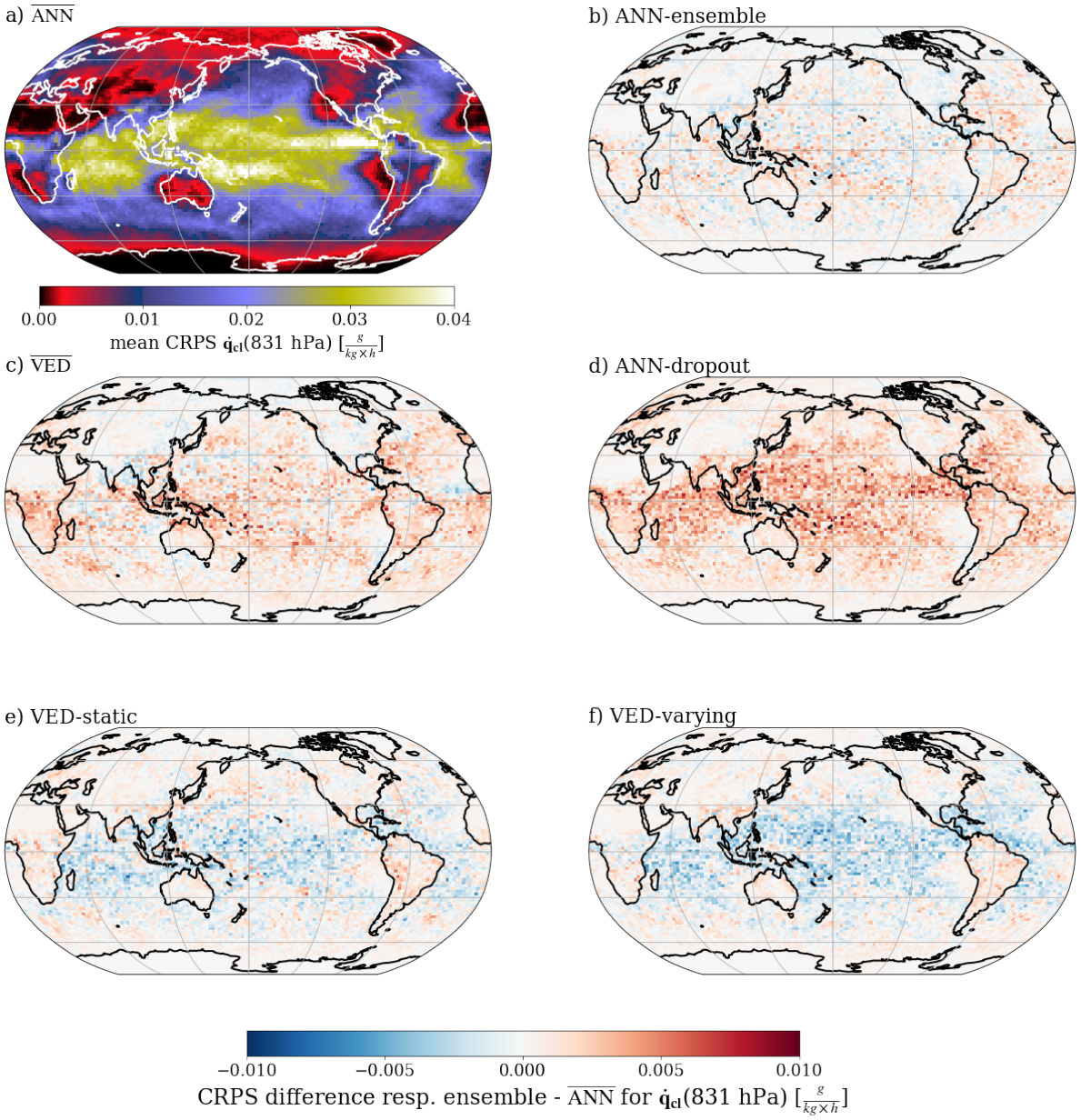

Figure B.7.: CRPS of $\dot{q}_{cl}$ in the upper planetary boundary layer on 831 hPa. The order of the shown parameterizations is identical to Figure 5.4. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

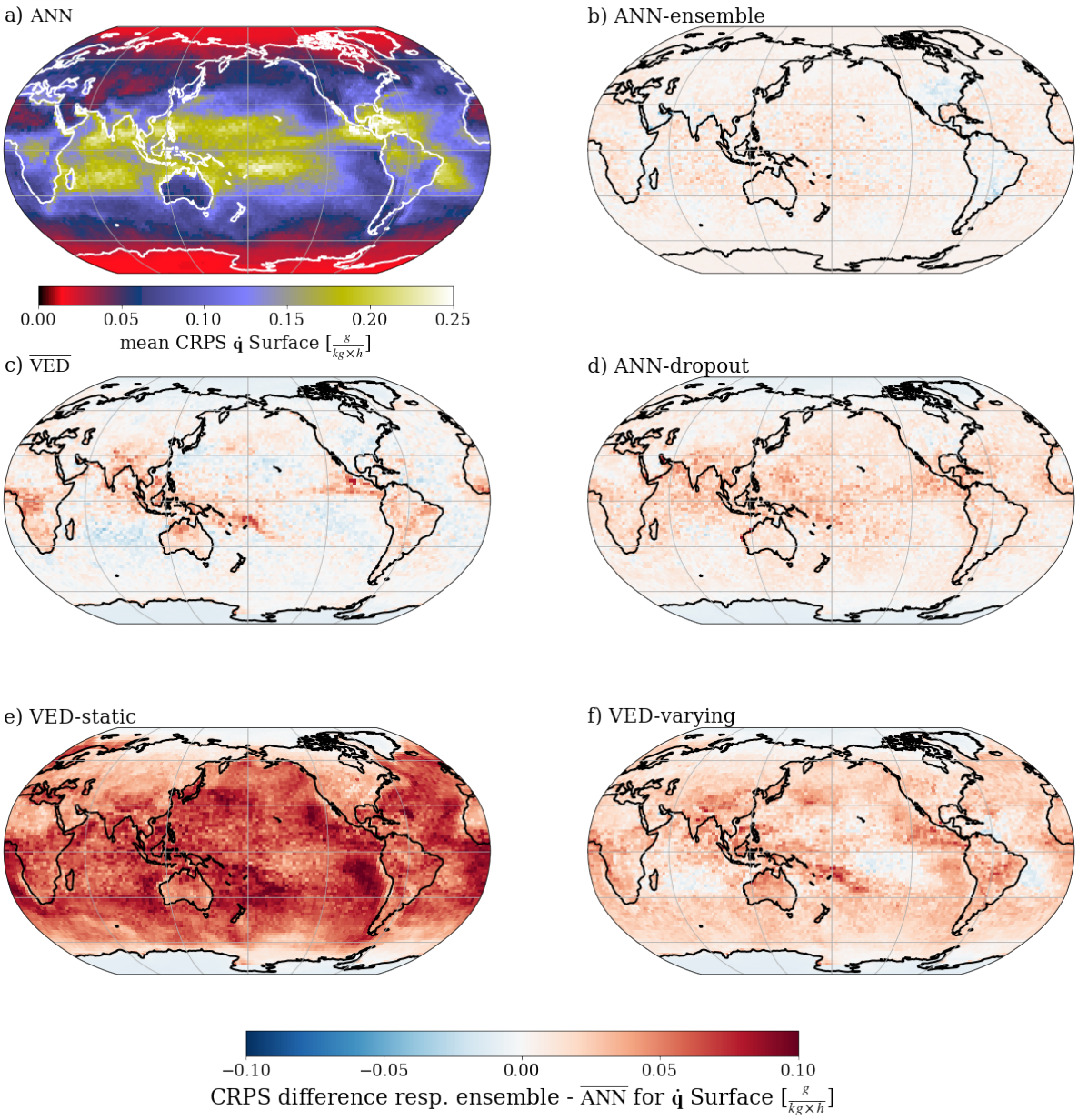

Figure B.8.: CRPS of surface $\dot{q}$. The order of the shown parameterizations is identical to Figure 5.4. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

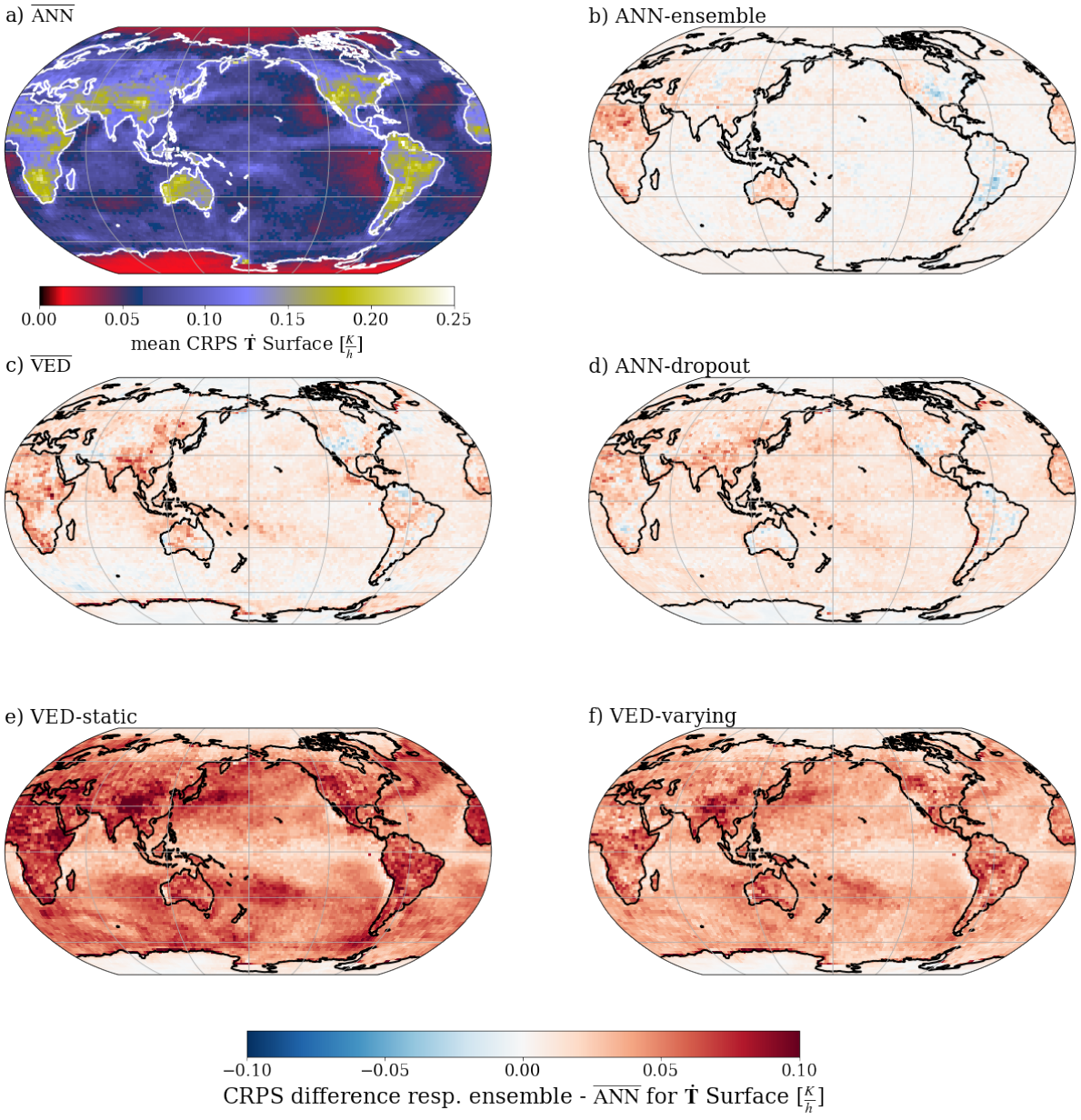

Figure B.9.: CRPS of surface $\dot{T}$. The order of the shown parameterizations is identical to Figure 5.4. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

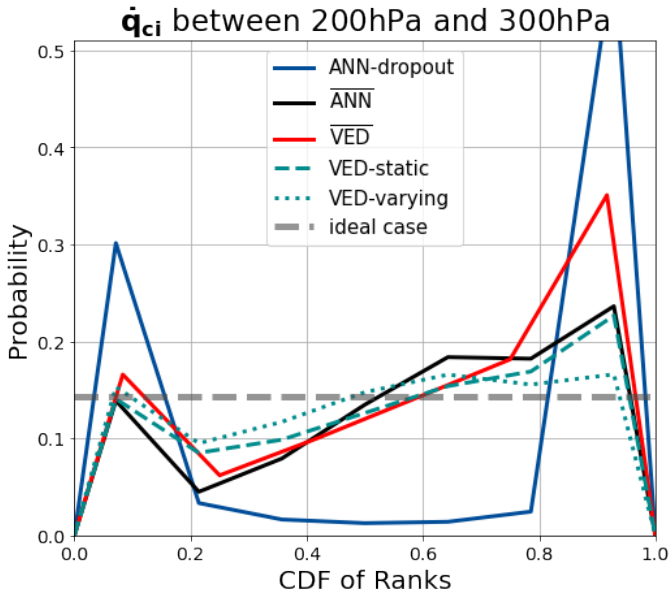

Figure B.10.: Probability Integral Transform (PIT) histogram of $\dot{q}_{ci}$ in the upper troposphere between 200 and 300 hPa. The x-axis represent the CDF of the ranks with respect to the number of ensemble members. The y-axis depicts the probability associated with each rank. The PIT histograms are based on 400 randomly drawn time steps from the test data set. The thick dashed gray line in the subplot in horizontal direction symbolises the perfect PIT histogram. The PIT curve of ANN-dropout is shown in blue and the PIT curves of $\overline{\text{ANN}}$ and ANN-ensemble in solid and dashed black. The PIT curve of $\overline{\text{VED}}$ is depicted in red. Additionally the PIT curves of VED-static and VED-varying are shown in dashed and dotted cyan. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

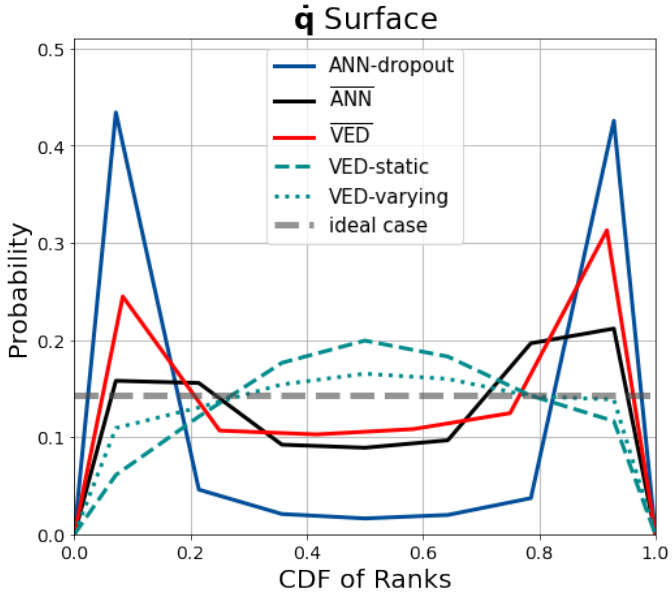

Figure B.11.: The PIT histograms for QBCTEND at the surface. The PIT histograms are again based on 400 randomly drawn time steps from the test data set. The color coding for the evaluated ensemble methods is identical to Figure B.10. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

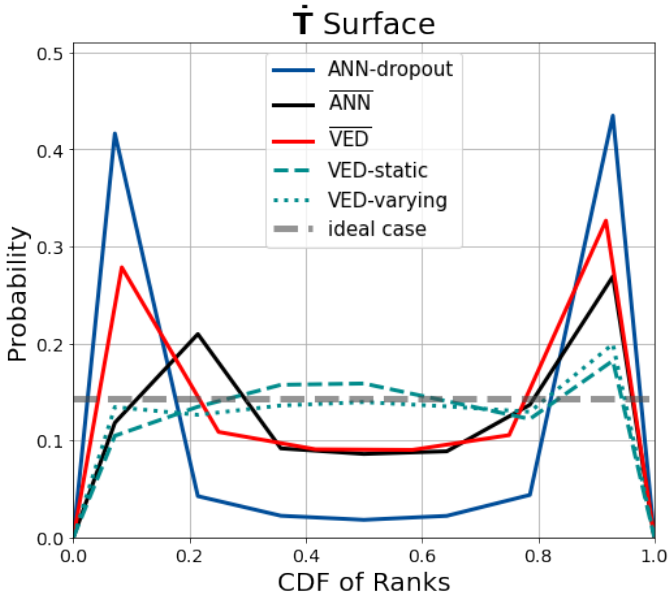

Figure B.12.: The PIT histograms for $\dot{T}$ at the surface. The PIT histograms are again based on 400 randomly drawn time steps from the test data set. The color coding for the evaluated ensemble methods is identical to Figure B.10. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

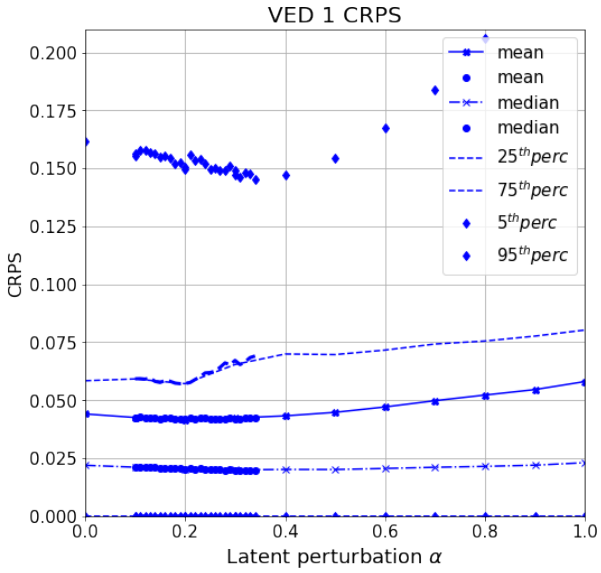

Figure B.13.: Aggregated CRPS over all SP variables $\boldsymbol{Y}$ as a function of static latent space perturbation α. Shown are the median, mean, the 5^{th} ,25^{th}, 75^{th}, 95^{th} percentile for both the coarse (in the range $\alpha = [0, 1]$) and fine ($\alpha = [0.1, 0.4]$) hyperparameter search. This Figure is directly reproduced from **Behrens** et al. 2024.

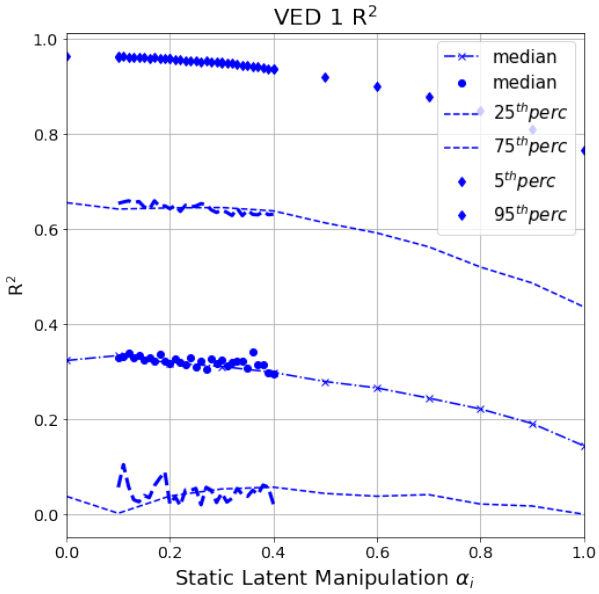

Figure B.14.: Aggregated coefficient of determination R^2 over all SP variables $\boldsymbol{Y}$ as a function of static latent space perturbation α. Shown are the median, mean, the 5^{th}, 25^{th}, 75^{th}, 95^{th} percentile for both the coarse (in the range $\alpha = [0, 1]$) and fine ($\alpha = [0.1, 0.4]$) hyperparameter search. This Figure is directly reproduced from **Behrens** et al. 2024.

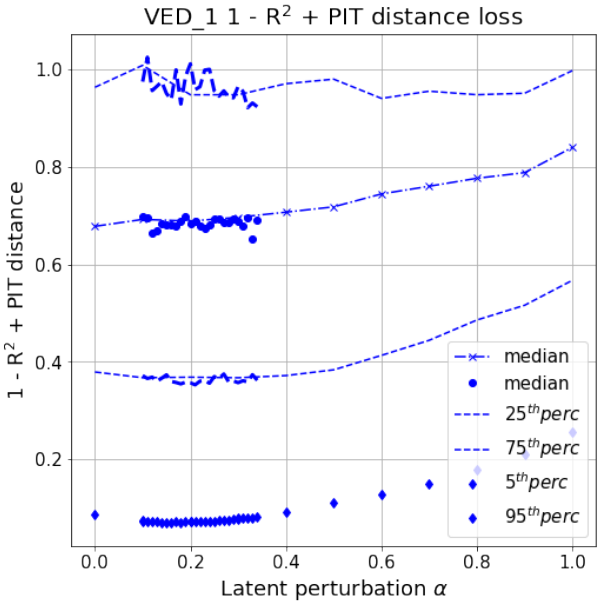

Figure B.15.: Aggregated loss function (1-R^2+PIT distance) over all SP variables $\boldsymbol{Y}$ as a function of static latent space perturbation α. Shown are the median, the 5^{th}, 25^{th}, 75^{th}, 95^{th} percentile for both the coarse (in the range $\alpha = [0, 1]$) and fine ($\alpha = [0.1, 0.4]$) hyperparameter search. This Figure is directly reproduced from **Behrens** et al. 2024.

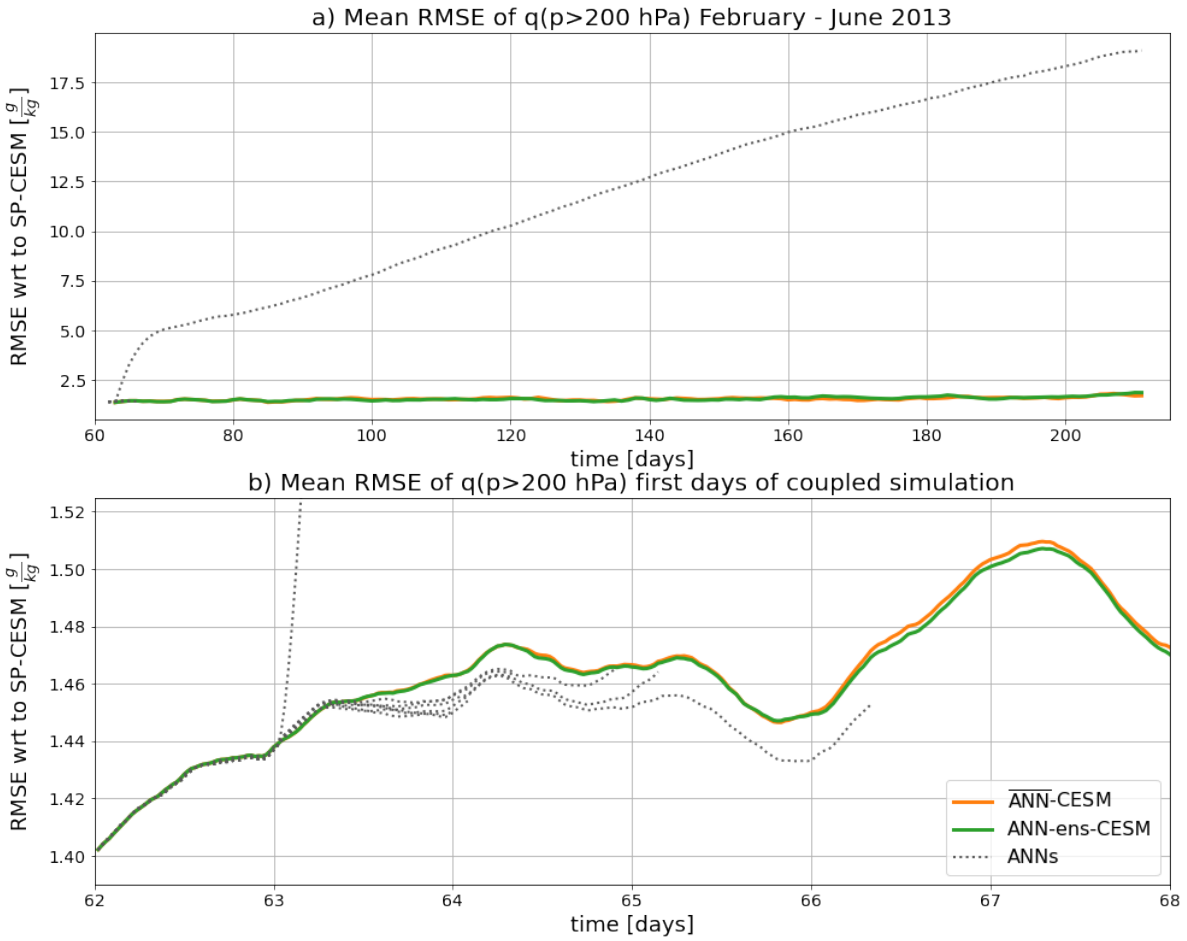

Figure B.16.: Mean Root Mean Squared Error (RMSE) of q below 200 hPa of the coupled runs with the deterministic ANN ensemble parameterization ($\overline{\text{ANN}}$-CESM, orange), the stochastic ANN ensemble parameterization (ANN-ens-CESM, green) and individual ANNs that for the ensembles (dotted grey lines) with respect to the independent run with the superparamterization (SP-CESM). Subplot a) depicts the mean RMSE timeseries from beginning of February to the end of June 2013. Subplot b) shows the timeseries zoomed in on the first six days of the simulations. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

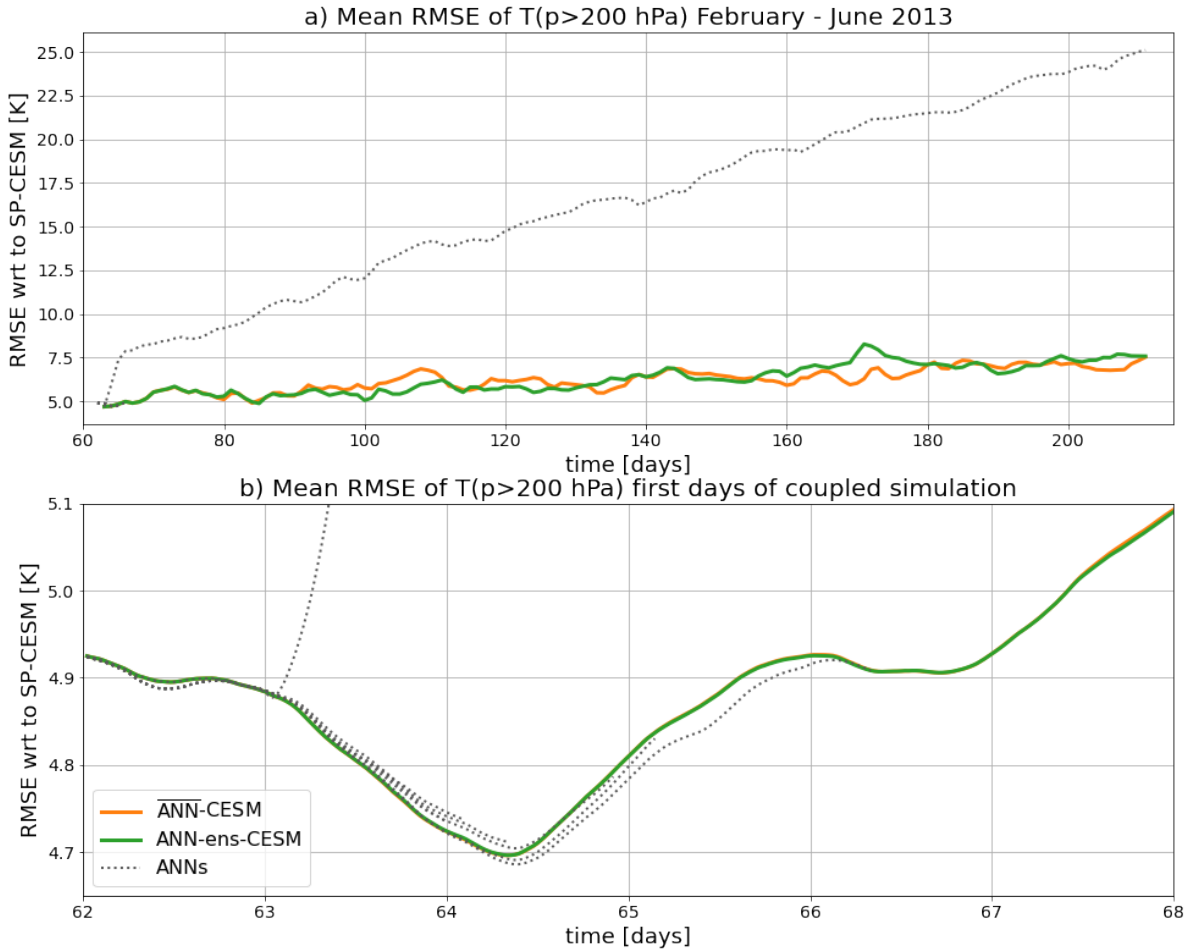

Figure B.17.: Mean Root Mean Squared Error (RMSE) of $\boldsymbol{T}$ below 200 hPa of the coupled runs with the deterministic ANN ensemble parameterization ($\overline{\text{ANN}}$-CESM, orange), the stochastic ANN ensemble parameterization (ANN-ens-CESM, green) and individual ANNs that for the ensembles (dotted grey lines) with respect to the independent run with the superparamterization (SP-CESM). Subplot a) depicts the mean RMSE timeseries from beginning of February to the end of June 2013. Subplot b) shows the timeseries zoomed in on the first six days of the simulations. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

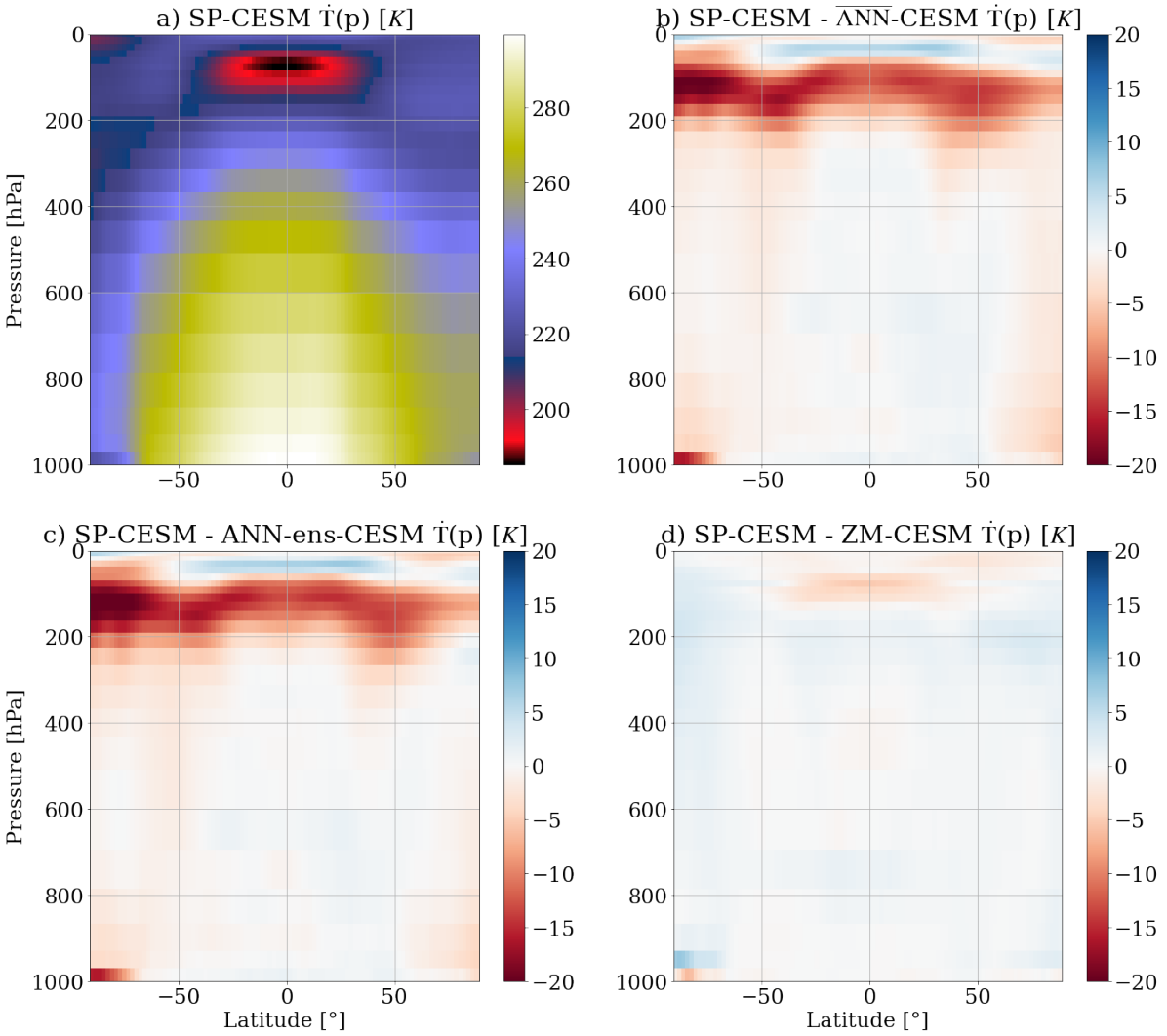

Figure B.18.: Zonal averages of the temperature field of SP-CESM over the period February to June 2013 (panel a), the difference in zonal averages between SP-CESM and CESM2 run with the deterministic ensemble parameterization ($\overline{\text{ANN}}$-CESM, panel b), between SP-CESM and CESM2 with the stochastic ensemble parameterization (ANN-ens-CESM, panel c) and between SP-CESM and with the Zhang-McFarlane scheme (ZM-CESM, panel d). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

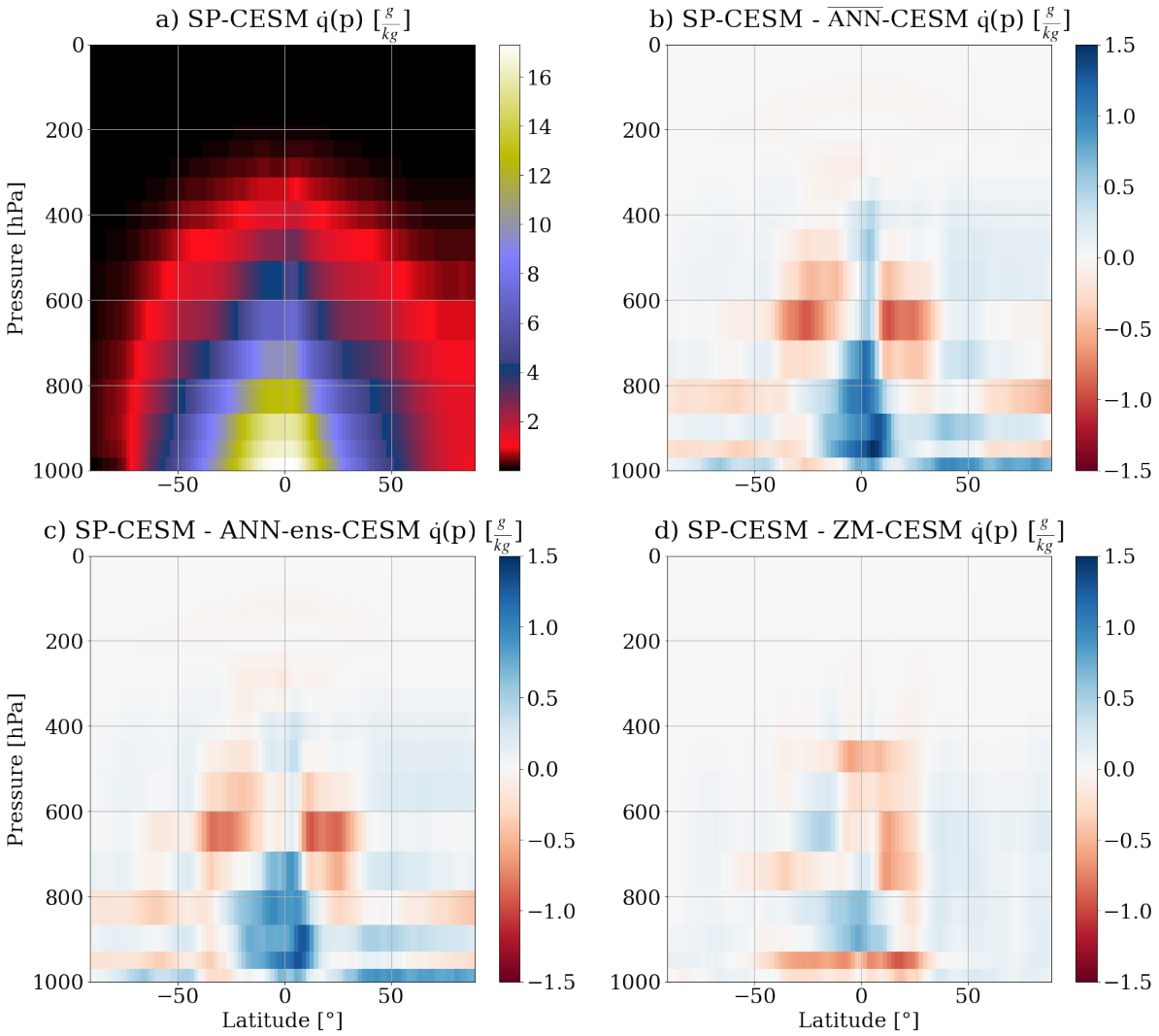

Figure B.19.: Zonal averages of the specific humidity field of SP-CESM over the period February to June 2013 (panel a), the difference in zonal averages between SP-CESM and CESM2 run with the deterministic ensemble parameterization ($\overline{\text{ANN}}$-CESM, panel b), between SP-CESM and CESM2 with the stochastic ensemble parameterization (ANN-ens-CESM, panel c) and between SP-CESM and with the Zhang-McFarlane scheme (ZM-CESM, panel d). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

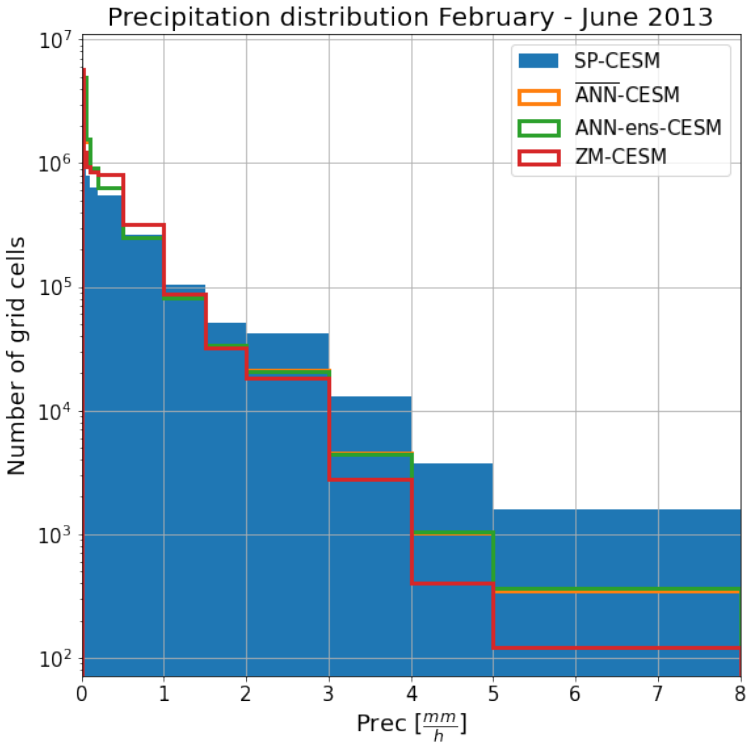

Figure B.20.: Precipitation histograms based on 10 million randomly drawn samples from the CESM2 runs with the superparameterization (SP-CESM, blue filled histogram), the deterministic ANN ensemble parameterization ($\overline{\text{ANN}}$-CESM, orange), the stochastic ANN ensemble parameterization (ANN-ens-CESM, green) and the Zhang-McFarlane scheme (ZM-CESM, red histogram) for the period February to June 2013. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

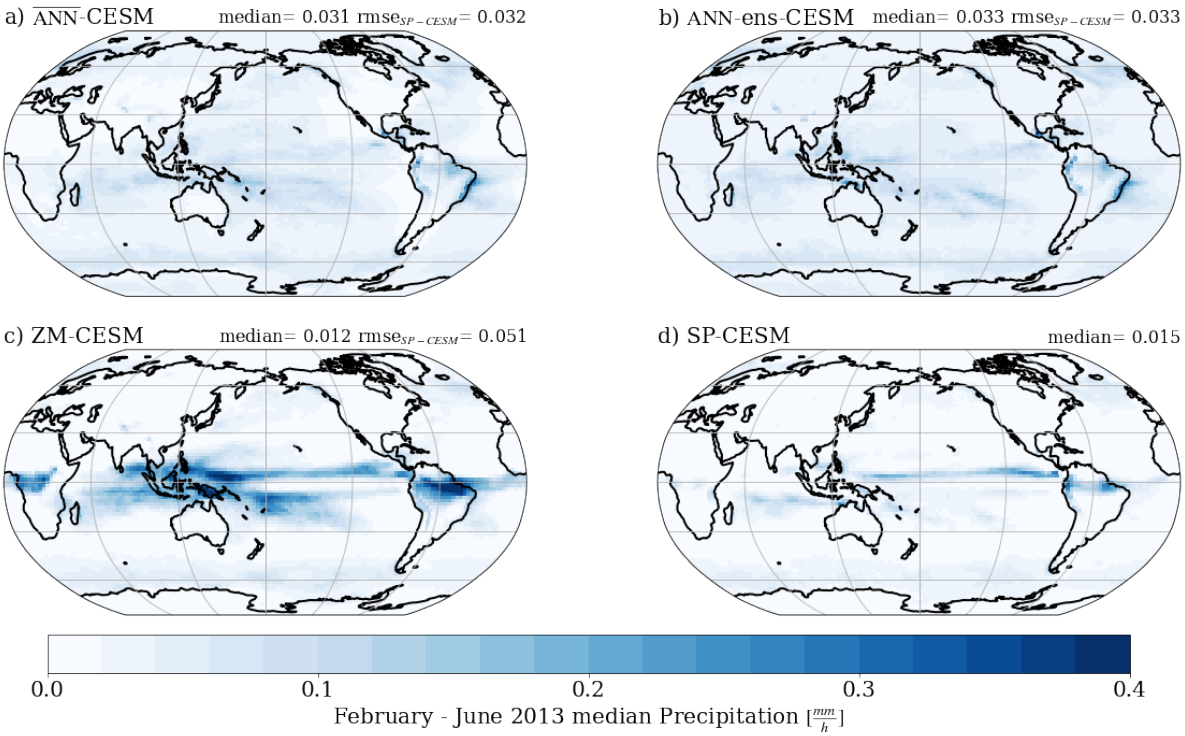

Figure B.21.: Global maps of the simulated median precipitation in CESM2 runs with the deterministic ANN ensemble parameterization ($\overline{\text{ANN}}$-CESM, panel a), the stochastic ANN ensemble parameterization (ANN-ens-CESM, b), the superparameterization (SP-CESM, c) and the Zhang-McFarlane scheme (ZM-CESM, d) for the period February to June 2013. The median value of the global map and the RMSE with respect to SP-CESM is shown above each respective panel. This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

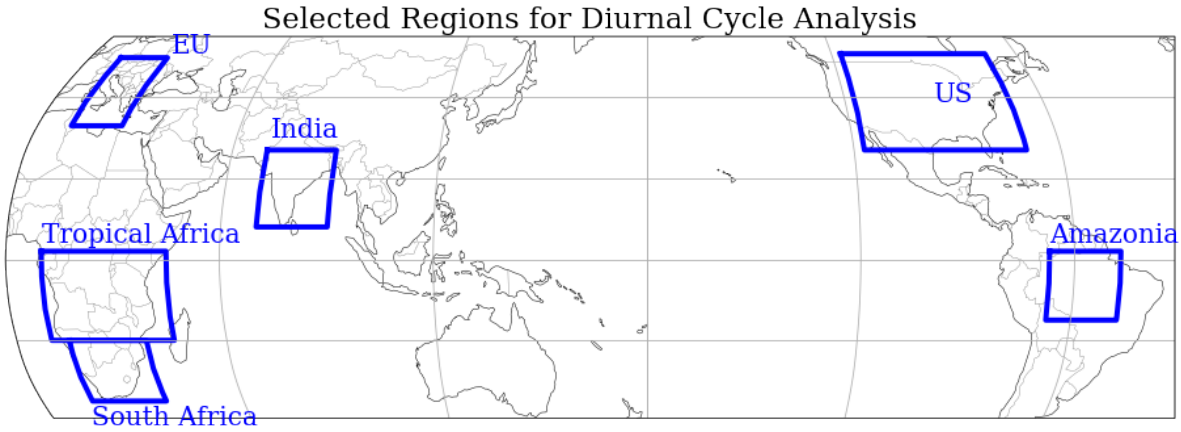

Figure B.22.: Regions that are used for the evaluation of the represented diurnal cycle in Figure B.23. This Figure is directly reproduced from **Behrens** et al. 2024.

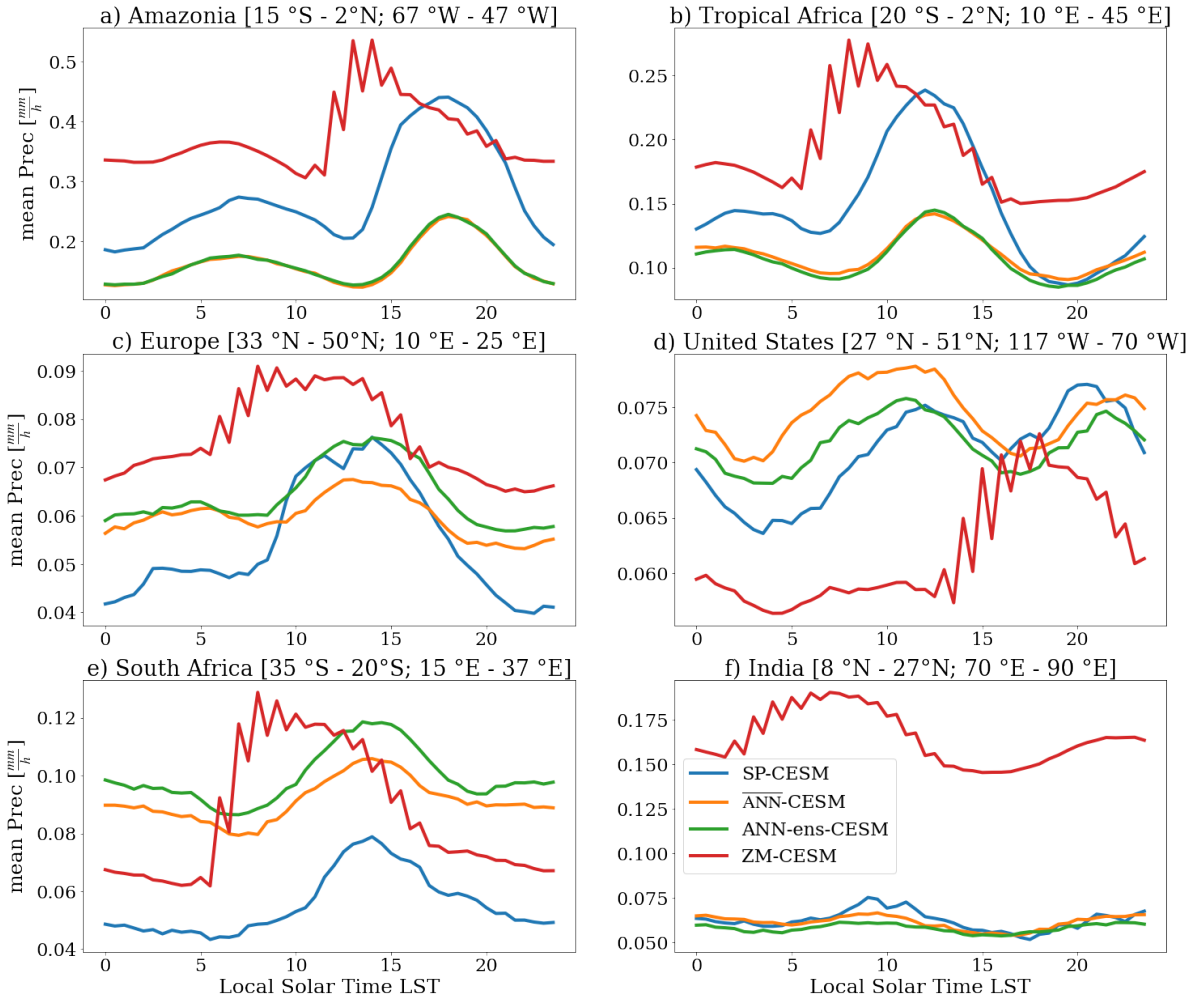

Figure B.23.: The simulated diurnal cycle of precipitation represented by the different parameterizations for the period February to June 2013 over Amazonia (panel a), tropical Africa (panel b), Europe (panel c), the United States (panel d), South Africa (panel e) and India (panel f). The diurnal cycle with the superparameterization is displayed by the blue line in each panel (SP-CESM), the deterministic ensemble parameterization by the orange line ($\overline{\text{ANN}}$-CESM), the stochastic ensemble by the green line (ANN-ens-CESM) and the Zhang-McFarlane scheme by the red line (ZM-CESM). This Figure is reproduced with minor modifications from **Behrens** et al. 2024.

List of Abbreviations

AED	AutoEncoder Decoder	23
ANN	Artificial Neural Network	23
CALIPSO	Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observation	9
CAM	Community Atmosphere Model	16
CAPE	Convective Available Potential Energy	9
CDF	Cumulative Distribution Function	82
CESM	Community Earth System Model	16
CCL	Convective Condensation Level	9
CGAN	Conditional Generative Adversarial Network	32
CIN	Convective Inhibition	9
CLM5	Community Land Model version 5	62
CMIP	Coupled Model Intercomparison Project	1
CO_2	carbon dioxide	2
CRPS	Continuous Rank Probability Score	77
cVAE	conditional Variational AutoEncoder Decoder	103
DL	deep learning	v
DYAMOND	Dynamics of the Atmospheric General Circulation Modeled On Non-Hydrostatic Domains	21
ECS	Equilibrium Climate Sensitivity	2
ED	Encoder Decoder	27
ERA5	European Centre for Medium-Range Weather Forecasts fifth-generation reanalysis	6
ESACCI-CLOUD	European Space Agency's Climate Change Initiative Cloud	6
ESM	Earth System Model	1
FKB	Fortran-Keras-Bridge	78
GPCP-SG	Global Precipitation Climatology Project - Satellite-Gauge	6
IPCC	Intergovernmental Panel on Climate Change	1

ITCZ Inter-Tropical Convergence Zone . . . 7

LCL Lifting Condensation Level . . . 9

LST Local Solar Time . . . 90

MAE Mean Absolute Error . . . 71

MCD Monte Carlo Dropout . . . 68

MH-NOAA/OI-SST-SIC Merged Hadley - National Oceanic and Atmospheric Administration / Optimum Interpolation Sea Surface Temperature and Sea Ice Concentration data set . . . 62

ML Machine Learning . . . v

MSE Mean Squared Error . . . 25

MJO Madden Julian Oscillation . . . 18

MMF multiscale modeling framework . . . 62

nextGEMS Next Generation Earth System Models . . . 22

NICAM Nonhydrostatic Icosahedral Atmospheric Model . . . 21

PC Principal Component . . . 43

PCA Principal Component Analysis . . . 42

PIT Probability Integral Transform . . . 82

R^2 Coefficient of Determination . . . 40

RMSE root mean square error . . . 78

SAM System for Atmospheric Modeling . . . 19

SP Superparameterization . . . 17

SPCAM Super Parameterized Community Atmosphere Model . . . 18

SPCESM Super Parameterized Earth System Model . . . 18

SPPT Stochastic Perturbed Parameter Tendencies scheme . . . 31

SRM Storm Resolving Model . . . 21

SST Sea Surface Temperature . . . 43

swcre shortwave cloud radiative effect . . . 13

VAE Variational Auto Encoder . . . 23

VED Variational Encoder Decoder . . . 29

WMO World Meteorological Organisation . . . 1

List of Figures

List of Tables

References

Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. A., & Murphy, K. (2018). Fixing a broken ELBO. *International Conference on Machine Learning*, 159–168. https://doi.org//10.48550/arXiv.1711.00464

Andersen, J. A., & Kuang, Z. (2012). Moist static energy budget of MJO-like disturbances in the atmosphere of a zonally symmetric aquaplanet. *Journal of Climate*, *25*(8), 2782–2804. https://doi.org/10.1175/JCLI-D-11-00168.1

Arakawa, A., & Schubert, W. H. (1974). Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I. *Journal of Atmospheric Sciences*, *31*(3), 674–701. https://doi.org/10.1175/1520-0469(1974)031<0674:IOACCE>2.0.CO;2

Behrens, **G.**, Beucler, T., Gentine, P., Iglesias-Suarez, F., Pritchard, M., & Eyring, V. (2022). Non-Linear Dimensionality Reduction With a Variational Encoder Decoder to Understand Convective Processes in Climate Models [e2022MS003130 2022MS003130]. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS003130. https://doi.org/10.1029/2022MS003130

Behrens, **G.**, Beucler, T., Iglesias-Suarez, F., Yu, S., Gentine, P., Pritchard, M., Schwabe, M., & Eyring, V. (2024). Improving Atmospheric Processes in Earth System Models with Deep Learning Ensembles and Stochastic Parameterizations. *In Review to Journal of Advances in Modeling Earth Systems*. https://doi.org/10.48550/arXiv.2402.03079

Berner, J., Achatz, U., Batté, L., Bengtsson, L., de la Cámara, A., Christensen, H. M., Colangeli, M., Coleman, D. R. B., Crommelin, D., Dolaptchiev, S. I., Franzke, C. L. E., Friederichs, P., Imkeller, P., Järvinen, H., Juricke, S., Kitsios, V., Lott, F., Lucarini, V., Mahajan, S., . . . Yano, J.-I. (2017). Stochastic Parameterization: Toward a New View of Weather and Climate Models. *Bulletin of the American Meteorological Society*, *98*(3), 565–588. https://doi.org/10.1175/BAMS-D-15-00268.1

Bessac, J., Christensen, H. M., Endo, K., Monahan, A. H., & Weitzel, N. (2021). Scale-Aware Space-Time Stochastic Parameterization of Subgrid-Scale Velocity Enhancement of Sea Surface Fluxes [e2020MS002367 2020MS002367]. *Journal of Advances in Modeling Earth Systems*, *13*(4), e2020MS002367. https://doi.org/10.1029/2020MS002367

Beucler, T., Rasp, S., Pritchard, M., & Gentine, P. (2019). Achieving conservation of energy in neural network emulators for climate modeling. *arXiv*, (1), 2–5. https://doi.org//10.48550/arXiv.1906.06622

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems. *Phys. Rev. Lett.*, *126*, 098302. https://doi.org/10.1103/PhysRevLett.126.098302

Bhouri, M. A., & Gentine, P. (2022). History-Based, Bayesian, Closure for Stochastic Parameterization: Application to Lorenz '96. https://doi.org/10.48550/arXiv.2210.14488

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., Meehl, G. A., Predoi, V., Roberts, M. J., & Eyring, V. (2020). Quantifying Progress Across Different CMIP Phases With the ESMValTool [e2019JD032321 2019JD032321]. *Journal of Geophysical Research: Atmospheres*, *125*(21), e2019JD032321. https://doi.org/10.1029/2019JD032321

Bony, S., Stevens, B., Frierson, D. M., Jakob, C., Kageyama, M., Pincus, R., Shepherd, T. G., Sherwood, S. C., Siebesma, A. P., Sobel, A. H., Watanabe, M., & Webb, M. J. (2015). Clouds, circulation and climate sensitivity. *Nature Geoscience*, *8*(4), 261–268. https://doi.org/10.1038/ngeo2398

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially Extended Tests of a Neural Network Parametrization Trained by Coarse-Graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019MS001711

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and Stabilizing Machine-Learning Parametrizations of Convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/JAS-D-20-0082.1

Buizza, R., Miller, M., & Palmer, T. (1999). Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. https://doi.org/10.21957/7kej0vcfs

Chinita, M. J., Witte, M., Kurowski, M. J., Teixeira, J., Suselj, K., Matheou, G., & Bogenschutz, P. (2023). Improving the representation of shallow cumulus convection with the simplified-higher-order-closure mass-flux (SHOC+MF v1.0) approach. *Geoscientific Model Development*, *16*(7), 1909–1924. https://doi.org/10.5194/gmd-16-1909-2023

Christensen, H. M., Moroz, I. M., & Palmer, T. N. (2015). Stochastic and Perturbed Parameter Representations of Model Uncertainty in Convection Parameterization. *Journal of the Atmospheric Sciences*, *72*(6), 2525–2544. https://doi.org/10.1175/JAS-D-14-0250.1

Christensen, H. M. (2020). Constraining stochastic parametrisation schemes using high-resolution simulations. *Quarterly Journal of the Royal Meteorological Society*, *146*(727), 938–962. https://doi.org/10.1002/qj.3717

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., McCaa, J. R., Williamson, D. L., Briegleb, B. P., Bitz, C. M., Lin, S.-J., & Zhang, M. (2006). The dynamical simulation of the Community Atmosphere Model version 3 (CAM3). *Journal of Climate*, *19*(11), 2162–2183. https://doi.org/10.1175/JCLI3762.1

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., ... Strand, W. G. (2020). The Community Earth System Model Version 2 (CESM2) [e2019MS001916 2019MS001916]. *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916. https://doi.org/10.1029/2019MS001916

Douville, H., Raghavan, K., Renwick, J., Allan, R., Arias, P., Barlow, M., Cerezo-Mota, R., Cherchi, A., Gan, T., Gergis, J., Jiang, D., Khan, A., Pokam Mba, W., Rosenfeld, D., Tierney, J., & Zolina, O. (2021). Water Cycle Changes. In V. Masson-Delmotte, P. Zhai,

A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 1055–1210). Cambridge University Press. https://doi.org/10.1017/9781009157896.010

Emanuel, K. (1994). *Atmospheric Convection*. Oxford University Press. https://books.google.de/books?id=VdaBBHEGAcMC

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Eyring, V., Gillett, N., Achuta Rao, K., Barimalala, R., Barreiro Parrillo, M., Bellouin, N., Cassou, C., Durack, P., Kosaka, Y., McGregor, S., Min, S., Morgenstern, O., & Sun, Y. (2021). Human Influence on the Climate System. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 423–552). Cambridge University Press. https://doi.org/10.1017/9781009157896.005

Frenkel, Y., Majda, A. J., & Khouider, B. (2012). Using the Stochastic Multicloud Model to Improve Tropical Convective Parameterization: A Paradigm Example. *Journal of the Atmospheric Sciences*, *69*(3), 1080–1105. https://doi.org/10.1175/JAS-D-11-0148.1

Frenkel, Y., Majda, A. J., & Khouider, B. (2013). Stochastic and deterministic multicloud parameterizations for tropical convection. *Climate Dynamics*, *41*(5-6), 1527–1551. https://doi.org//10.1007/s00382-013-1678-z

Frenkel, Y., Majda, A. J., & Stechmann, S. N. (2015). Cloud-radiation feedback and atmosphere-ocean coupling in a stochastic multicloud model. *Dynamics of Atmospheres and Oceans*, *71*, 35–55. https://doi.org//10.1016/j.dynatmoce.2015.05.003

Gagne II, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz '96 Model [e2019MS001896 10.1029/2019MS001896]. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896. https://doi.org/10.1029/2019MS001896

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could Machine Learning Break the Convection Parameterization Deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. https://doi.org/10.1029/2018GL078202

Gentine, P., Eyring, V., & Beucler, T. (2021). Deep Learning for the Parametrization of Subgrid Processes in Climate Models. In *Deep Learning for the Earth Sciences* (pp. 307–314). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781119646181.ch21

Gettelman, A., & Rood, R. B. (2016). Essence of a Climate Model. In *Demystifying Climate Models: A Users Guide to Earth System Models* (pp. 37–58). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-48959-8_4

Gneiting, T., & Raftery, A. E. (2005). Weather forecasting with ensemble methods. *Science*, *310*(5746), 248–249. https://doi.org/10.1126/science.1115255

Gneiting, T., & Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning* [http://www.deeplearningbook.org]. MIT Press.

Grabowski, W. W., Wu, X., & Moncrieff, M. W. (1996). Cloud-Resolving Modeling of Tropical Cloud Systems during Phase III of GATE. Part I: Two-Dimensional Experiments. *Journal of Atmospheric Sciences*, *53*(24), 3684–3709. https://doi.org/10.1175/1520-0469(1996)053<3684:CRMOTC>2.0.CO;2

Grabowski, W. W. (2001). Coupling cloud processes with the large-scale dynamics using the clouds-resolving convection parameterization (CRCP). *Journal of the Atmospheric Sciences*, *58*(9), 978–997. https://doi.org/10.1175/1520-0469(2001)058<0978:CCPWTL>2.0.CO;2

Grundner, A., Beucler, T., Gentine, P., & Eyring, V. (2024). Data-Driven Equation Discovery of a Cloud Cover Parameterization [e2023MS003763 2023MS003763]. https://doi.org//10.1029/2023MS003763

Guillaumin, A. P., & Zanna, L. (2021). Stochastic-Deep Learning Parameterization of Ocean Momentum Forcing [e2021MS002534 2021MS002534]. *Journal of Advances in Modeling Earth Systems*, *13*(9), e2021MS002534. https://doi.org/10.1029/2021MS002534

Hagihara, Y., Okamoto, H., & Luo, Z. J. (2014). Joint analysis of cloud top heights from CloudSat and CALIPSO: New insights into cloud top microphysics. *Journal of Geophysical Research: Atmospheres*, *119*(7), 4087–4106. https://doi.org/10.1002/2013JD020919

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A Moist Physics Parameterization Based on Deep Learning [e2020MS002076 2020MS002076]. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002076. https://doi.org/10.1029/2020MS002076

Han, Y., Zhang, G. J., & Wang, Y. (2023). An Ensemble of Neural Networks for Moist Physics Processes, Its Generalizability and Stable Integration [e2022MS003508 2022MS003508]. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2022MS003508. https://doi.org/10.1029/2022MS003508

Haynes, K., Lagerquist, R., McGraw, M., Musgrave, K., & Ebert-Uphoff, I. (2023). Creating and Evaluating Uncertainty Estimates with Neural Networks for Environmental-Science Applications. *Artificial Intelligence for the Earth Systems*, *2*(2), 220061. https://doi.org/10.1175/AIES-D-22-0061.1

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, *abs/1207.0580*. https://doi.org//10.48550/arXiv.1207.0580

Hoefler, T., Stevens, B., Prein, A. F., Baehr, J., Schulthess, T., Stocker, T. F., Taylor, J., Klocke, D., Manninen, P., Forster, P. M., Kölling, T., Gruber, N., Anzt, H., Frauen, C., Ziemen, F., Klöwer, M., Kashinath, K., Schär, C., Fuhrer, O., & Lawrence, B. N. (2023). Earth Virtualization Engines: A Technical Perspective. *Computing in Science & Engineering*, *25*(3), 50–59. https://doi.org/10.1109/MCSE.2023.3311148

Hohenegger, C., Kornblueh, L., Klocke, D., Becker, T., Cioni, G., Engels, J. F., Schulzweida, U., & Stevens, B. (2020). Climate Statistics in Global Simulations of the Atmosphere, from 80 to 2.5 km Grid Spacing. *Journal of the Meteorological Society of Japan. Ser. II*, *98*(1), 73–91. https://doi.org/10.2151/jmsj.2020-005

Hohenegger, C., Korn, P., Linardakis, L., Redler, R., Schnur, R., Adamidis, P., Bao, J., Bastin, S., Behravesh, M., Bergemann, M., Biercamp, J., Bockelmann, H., Brokopf, R., Brüggemann, N., Casaroli, L., Chegini, F., Datseris, G., Esch, M., George, G., . . . Stevens, B. (2023). ICON-Sapphire: simulating the components of the Earth system and their interactions at kilometer and subkilometer scales. *Geoscientific Model Development*, *16*(2), 779–811. https://doi.org/10.5194/gmd-16-779-2023

Holmlund, K., Grandell, J., Schmetz, J., Stuhlmann, R., Bojkov, B., Munro, R., Lekouara, M., Coppens, D., Viticchie, B., August, T., Theodore, B., Watts, P., Dobber, M., Fowler, G., Bojinski, S., Schmid, A., Salonen, K., Tjemkes, S., Aminou, D., & Blythe, P. (2021). Meteosat Third Generation (MTG): Continuation and Innovation of Observations from Geostationary Orbit. *Bulletin of the American Meteorological Society*, *102*(5), E990–E1015. https://doi.org/10.1175/BAMS-D-19-0304.1

Howard, L. (1894). *On the modifications of clouds*. Asher. https://digital.nmla.metoffice.gov.uk/IO_51ce11e6-5ca5-47a0-8fb4-6a6d251ff3d4/

Huaman, L., & Schumacher, C. (2018). Assessing the Vertical Latent Heating Structure of the East Pacific ITCZ Using the CloudSat CPR and TRMM PR. *Journal of Climate*, *31*(7), 2563–2577. https://doi.org/10.1175/JCLI-D-17-0590.1

Huffman, G. J., Adler, R. F., Behrangi, A., Bolvin, D. T., Nelkin, E. J., Gu, G., & Ehsani, M. R. (2023). The New Version 3.2 Global Precipitation Climatology Project (GPCP) Monthly and Daily Precipitation Products. *Journal of Climate*, *36*(21), 7635–7655. https://doi.org/10.1175/JCLI-D-23-0123.1

Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M., & Rosinski, J. (2008). A New Sea Surface Temperature and Sea Ice Boundary Dataset for the Community Atmosphere Model. *Journal of Climate*, *21*(19), 5145–5153. https://doi.org/10.1175/2008JCLI2292.1

Iglesias-Suarez, F., Gentine, P., Solino-Fernandez, B., Beucler, T., Pritchard, M., Runge, J., & Eyring, V. (2024). Causally-Informed Deep Learning to Improve Climate Models and Projections [e2023JD039202 2023JD039202]. https://doi.org//10.1029/2023JD039202

IPCC. (2021). Summary for Policymakers. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth*

Assessment Report of the Intergovernmental Panel on Climate Change (pp. 3–32). Cambridge University Press. https://doi.org/10.1017/9781009157896.001

Jansson, F., van den Oord, G., Pelupessy, I., Chertova, M., Grönqvist, J. H., Siebesma, A. P., & Crommelin, D. (2022). Representing Cloud Mesoscale Variability in Superparameterized Climate Models [e2021MS002892 2021MS002892]. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2021MS002892. https://doi.org/10.1029/2021MS002892

Jones, T. R., Randall, D. A., & Branson, M. D. (2019a). Multiple-Instance Superparameterization: 1. Concept, and Predictability of Precipitation. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3497–3520. https://doi.org/10.1029/2019MS001610

Jones, T. R., Randall, D. A., & Branson, M. D. (2019b). Multiple-Instance Superparameterization: 2. The Effects of Stochastic Convection on the Simulated Climate. *Journal of Advances in Modeling Earth Systems*, *11*(11), 3521–3544. https://doi.org/10.1029/2019MS001611

Khairoutdinov, M. F., & Randall, D. A. (2001). A cloud resolving model as a cloud parameterization in the NCAR community climate system model: Preliminary results. *Geophysical Research Letters*, *28*(18), 3617–3620. https://doi.org/10.1029/2001GL013552

Khairoutdinov, M. F., & Randall, D. A. (2003). Cloud Resolving Modeling of the ARM Summer 1997 IOP: Model Formulation, Results, Uncertainties, and Sensitivities. *Journal of the Atmospheric Sciences*, *60*(4), 607–625. https://doi.org/10.1175/1520-0469(2003)060<0607:CRMOTA>2.0.CO;2

Khairoutdinov, M., Randall, D., & DeMott, C. (2005). Simulations of the Atmospheric General Circulation Using a Cloud-Resolving Model as a Superparameterization of Physical Processes. *Journal of the Atmospheric Sciences*, *62*(7), 2136–2154. https://doi.org/10.1175/JAS3453.1

Khouider, B., & Majda, A. J. (2006). A simple multicloud parameterization for convectively coupled tropical waves. Part I: Linear analysis. *Journal of the Atmospheric Sciences*, *63*(4), 1308–1323. https://doi.org/10.1175/JAS3677.1

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*. https://doi.org/10.48550/arXiv.1412.6980

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, (1050), 1–14. https://doi.org//10.48550/arXiv.1312.6114

Kingma, D. P., & Welling, M. (2019). An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, *12*(4), 307–392. https://doi.org/10.1561/2200000056

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Hou, Y. T., Lord, S. J., & Belochitski, A. A. (2010). Accurate and Fast Neural Network Emulations of Model Radiation for the NCEP Coupled Climate Forecast System: Climate Simulations and Seasonal Predictions. *Monthly Weather Review*, *138*(5), 1822–1842. https://doi.org/10.1175/2009MWR3149.1

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013). Using Ensemble of Neural Networks to Learn Stochastic Convection Parameterizations for Climate and

Numerical Weather Prediction Models from Data Simulated by a Cloud Resolving Model. *Adv. Artif. Neu. Sys., 2013*. https://doi.org/10.1155/2013/485913

Krinitskiy, M. A., Zyulyaeva, Y. A., & Gulev, S. K. (2019). Clustering of polar vortex states using convolutional autoencoders. *CEUR Workshop Proceedings*, *2426*, 52–61.

Lauer, A., Bock, L., Hassler, B., Schröder, M., & Stengel, M. (2023). Cloud Climatologies from Global Climate Models—A Comparison of CMIP5 and CMIP6 Models with Satellite Data. *Journal of Climate*, *36*(2), 281–311. https://doi.org/10.1175/JCLI-D-22-0181.1

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J., Engelbrecht, F., Fischer, E., Fyfe, J., Jones, C., Maycock, A., Mutemi, J., Ndiaye, O., Panickal, S., & Zhou, T. (2021). Future Global Climate: Scenario-Based Projections and Near-Term Information. In V. Masson-Delmotte, P. Zhai, A. Pirani, S. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. Matthews, T. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou (Eds.), *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 553–672). Cambridge University Press. https://doi.org/10.1017/9781009157896.006

Liljequist, G. H., & Cehak, K. (2013). *Allgemeine Meteorologie*. Springer-Verlag. https://books.google.de/books?id=3tKeBgAAQBAJ

Lin, J., Yu, S., Beucler, T., Gentine, P., Walling, D., & Pritchard, M. (2023). Systematic Sampling and Validation of Machine Learning-Parameterizations in Climate Models. https://doi.org/10.48550/arXiv.2309.16177

Lohmann, U., Lüönd, F., & Mahrt, F. (2016). *An Introduction to Clouds: From the Microscale to Climate*. Cambridge University Press. https://books.google.de/books?id=FbpDDAAAQBAJ

Lorenz, E. N. (1996). Predictability: A problem partly solved. *Proc. Seminar on predictability*, *1*(1). https://www.ecmwf.int/en/elibrary/75462-predictability-problem-partly-solved

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. https://doi.org/10.48550/arXiv.1705.07874

Mamalakis, A., Ebert-Uphoff, I., & Barnes, E. A. (2022). Neural network attribution methods for problems in geoscience: A novel synthetic benchmark dataset. *Environmental Data Science*, *1*, e8. https://doi.org/10.1017/eds.2022.7

Mooers, G., Tuyls, J., Mandt, S., Pritchard, M., & Beucler, T. (2020). Generative Modeling of Atmospheric Convection. *arXiv*. https://doi.org/10.1145/3429309.3429324

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentine, P. (2021). Assessing the Potential of Deep Learning for Emulating Cloud Superparameterization in Climate Models With Real-Geography Boundary Conditions [e2020MS002385 2020MS002385]. *Journal of Advances in Modeling Earth Systems*, *13*(5), e2020MS002385. https://doi.org/10.1029/2020MS002385

Mooers, G., Pritchard, M., Beucler, T., Srivastava, P., Mangipudi, H., Peng, L., Gentine, P., & Mandt, S. (2023). Comparing Storm Resolving Models and Climates via Unsupervised Machine Learning. https://doi.org/10.48550/arXiv.2208.11843

Nadiga, B. T., Sun, X., & Nash, C. (2022). Stochastic parameterization of column physics using generative adversarial networks. *Environmental Data Science*, *1*, e22. https://doi.org/10.1017/eds.2022.32

Neelin, J. D., & Zeng, N. (2000). A quasi-equilibrium tropical circulation model—Formulation. *Journal of the atmospheric sciences*, *57*(11), 1741–1766. https://doi.org/10.1175/1520-0469(2000)057<1741:AQETCM>2.0.CO;2

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). A Fortran-Keras Deep Learning Bridge for Scientific Computing. *Scientific Programming*, *2020*, 1–13. https://doi.org/10.1155/2020/8888811

Pan, B., Anderson, G. J., Goncalves, A., Lucas, D. D., Bonfils, C. J. W., & Lee, J. (2022). Improving Seasonal Forecast Using Probabilistic Deep Learning [e2021MS002766 2021MS002766]. *Journal of Advances in Modeling Earth Systems*, *14*(3), e2021MS002766. https://doi.org/10.1029/2021MS002766

Parthipan, R., Christensen, H. M., Hosking, J. S., & Wischik, D. J. (2022). Using Probabilistic Machine Learning to Better Model Temporal Patterns in Parameterizations: a case study with the Lorenz 96 model. *EGUsphere*, *2022*, 1–27. https://doi.org/10.5194/egusphere-2022-912

Perezhogin, P., Zanna, L., & Fernandez-Granda, C. (2023). Generative Data-Driven Approaches for Stochastic Subgrid Parameterizations in an Idealized Ocean Model [e2023MS003681 2023MS003681]. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2023MS003681. https://doi.org/10.1029/2023MS003681

Pritchard, M. S., & Somerville, R. C. J. (2009). Assessing the diurnal cycle of precipitation in a multi-scale climate model. *Journal of Advances in Modeling Earth Systems*, *2*. https://doi.org/10.3894/james.2009.1.12

Pritchard, M. S., & Bretherton, C. S. (2014). Causal evidence that rotational moisture advection is critical to the superparameterized Madden–Julian oscillation. *Journal of the Atmospheric Sciences*, *71*(2), 800–815. https://doi.org//10.1175/JAS-D-13-0119.1

Pritchard, M. S., Bretherton, C. S., & DeMott, C. A. (2014). Restricting 32–128 km horizontal scales hardly affects the MJO in the Superparameterized Community Atmosphere Model v. 3.0 but the number of cloud-resolving grid columns constrains vertical mixing. *Journal of Advances in Modeling Earth Systems*, *6*(3), 723–739. https://doi.org//10.1002/2014MS000340

Randall, D. A., Khairoutdinov, M., Arakawa, A., & Grabowski, W. (2003). Breaking the Cloud Parameterization Deadlock. *Bulletin of the American Meteorological Society*, *84*(11), 1547–1564. https://doi.org/10.1175/BAMS-84-11-1547

Randall, D. A. (2013). Beyond deadlock. *Geophysical Research Letters*, *40*(22), 5970–5976. https://doi.org/10.1002/2013GL057998

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(39), 9684–9689. https://doi.org/10.1073/pnas.1810286115

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat, M. (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*, 195. https://doi.org/10.1038/s41586-019-0912-1

Rolinek, M., Zietlow, D., & Martius, G. (2019). Variational Autoencoders Pursue PCA Directions (by Accident). *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 12406–12415. http://openaccess.thecvf.com/content_CVPR_2019/papers/Rolinek_Variational_Autoencoders_Pursue_PCA_Directions_by_Accident_CVPR_2019_paper.pdf

Sakradzija, M., & Klocke, D. (2018). Physically Constrained Stochastic Shallow Convection in Realistic Kilometer-Scale Simulations. *Journal of Advances in Modeling Earth Systems*, *10*(11), 2755–2776. https://doi.org/10.1029/2018ms001358

Saranya, A., & Subhashini, R. (2023). A systematic review of Explainable Artificial Intelligence models and applications: Recent developments and future trends. *Decision Analytics Journal*, *7*, 100230. https://doi.org/10.1016/j.dajour.2023.100230

Satoh, M., Matsuno, T., Tomita, H., Miura, H., Nasuno, T., & Iga, S. (2008). Nonhydrostatic icosahedral atmospheric model (NICAM) for global cloud resolving simulations [Predicting weather, climate and extreme events]. *Journal of Computational Physics*, *227*(7), 3486–3514. https://doi.org/10.1016/j.jcp.2007.02.006

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-j., Putman, W. M., & Düben, P. (2019). Global Cloud-Resolving Models, 172–184. https://doi.org/10.1007/s40641-019-00131-0

Schlund, M., Lauer, A., Gentine, P., Sherwood, S. C., & Eyring, V. (2020). Emergent constraints on equilibrium climate sensitivity in CMIP5: do they hold for CMIP6? *Earth System Dynamics*, *11*(4), 1233–1258. https://doi.org/10.5194/esd-11-1233-2020

Schneider, T., Teixeira, J., Bretherton, C., Brient, F., Pressel, K., Schär, C., & Siebesma, A. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*, 3–5. https://doi.org/10.1038/nclimate3190

Shamekh, S., Lamb, K. D., Huang, Y., & Gentine, P. (2023). Implicit learning of convective organization explains precipitation stochasticity. *Proceedings of the National Academy of Sciences*, *120*(20), e2216158120. https://doi.org/10.1073/pnas.2216158120

Shin, J., & Baik, J.-J. (2022). Parameterization of Stochastically Entraining Convection Using Machine Learning Technique [e2021MS002817 2021MS002817]. *Journal of Advances in Modeling Earth Systems*, *14*(5), e2021MS002817. https://doi.org/10.1029/2021MS002817

Sohn, K., Yan, X., & Lee, H. (2015). Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems*, *2015-1*, 3483–3491. https://papers.nips.cc/paper_files/paper/2015/file/8d55a249e6baa5c06772297520da2051-Paper.pdf

Sonnewald, M., Wunsch, C., & Heimbach, P. (2019). Unsupervised Learning Reveals Geography of Global Ocean Dynamical Regions. *Earth and Space Science*, *6*(5), 784–794. https://doi.org/10.1029/2018EA000519

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., Klocke, D., Kodama, C., Kornblueh, L., Lin, S. J., Neumann, P., Putman, W. M., Röber, N., Shibuya, R., Vanniere, B., Vidale, P. L., ... Zhou, L. (2019). DYAMOND: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains. *Progress in Earth and Planetary Science*, *6*(1). https://doi.org/10.1186/s40645-019-0304-z

Stevens, B., Acquistapace, C., Hansen, A., Heinze, R., Klinger, C., Klocke, D., Rybka, H., Schubotz, Q., Windmiller, J., Adamidis, P., Arka, I., Barlakas, V., Biercamp, J., Brueck, M., Brune, S., Buehler, S. A., Burkhardt, U., Cioni, G., Costa-Suròs, M., ... Zängl, G. (2020). The Added Value of Large-eddy and Storm-resolving Models for Simulating Clouds and Precipitation. *Journal of the Meteorological Society of Japan. Ser. II*, *98*(2), 395–435. https://doi.org/10.2151/jmsj.2020-021

Tibau Alberdi, X.-A., Requena-Mesa, C., Reimers, C., Denzler, J., Eyring, V., Reichstein, M., & Runge, J. (2018). SupernoVAE : VAE based kernel PCA for analysis of spatio-temporal Earth data. https://www.researchgate.net/publication/330984865_SupernoVAE_VAE_based_kernel_PCA_for_analysis_of_spatio-temporal_Earth_data

Tomita, H., & Satoh, M. (2004). A new dynamical framework of nonhydrostatic global model using the icosahedral grid. *Fluid Dynamics Research*, *34*(6), 357. https://doi.org/10.1016/j.fluiddyn.2004.03.003

Tompkins, A. M., & Craig, G. C. (1998). Radiative–convective equilibrium in a three-dimensional cloud-ensemble model. *Quarterly Journal of the Royal Meteorological Society*, *124*(550), 2073–2097. https://doi.org/10.1002/qj.49712455013

Wang, P., Yuval, J., & O'Gorman, P. A. (2022a). Non-Local Parameterization of Atmospheric Subgrid Processes With Neural Networks [e2022MS002984 2022MS002984]. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS002984. https://doi.org/10.1029/2022MS002984

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022b). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development*, *15*(9), 3923–3940. https://doi.org/10.5194/gmd-15-3923-2022

Wing, A. A., Emanuel, K., Holloway, C. E., & Muller, C. (2018). Convective self-aggregation in numerical simulations: A review. *Shallow clouds, water vapor, circulation, and climate sensitivity*, 1–25. https://doi.org/10.1007/s10712-017-9408-4

Woelfle, M. D., Yu, S., Bretherton, C. S., & Pritchard, M. S. (2018). Sensitivity of Coupled Tropical Pacific Model Biases to Convective Parameterization in CESM1. *Journal of Advances in Modeling Earth Systems*, *10*(1), 126–144. https://doi.org/10.1002/2017MS001176

Xu, K.-M., & Randall, D. A. (1996). Explicit Simulation of Cumulus Ensembles with the GATE Phase III Data: Comparison with Observations. *Journal of Atmospheric Sciences*, *53*(24), 3710–3736. https://doi.org/10.1175/1520-0469(1996)053<3710:ESOCEW>2.0.CO;2

Yanai, M., Esbensen, S., & Chu, J.-H. (1973). Determination of Bulk Properties of Tropical Cloud Clusters from Large-Scale Heat and Moisture Budgets. *Journal of Atmospheric Sciences*,

30(4), 611–627. https://doi.org/10.1175/1520-0469(1973)030<0611:DOBPOT>2.0.CO;2

Yu, S., Hannah, W. M., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., **Behrens, G.**, Busecke, J. J. M., Loose, N., Stern, C., Beucler, T., Harrop, B. E., Hilman, B. R., Jenney, A. M., Ferretti, S. L., Liu, N., Anandkumar, A., … Pritchard, M. S. (2023). ClimSim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators. https://doi.org/10.48550/arXiv.2306.08754

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 1–10. https://doi.org/10.1038/s41467-020-17142-3

Zhang, G., & McFarlane, N. A. (1995). Sensitivity of climate simulations to the parameterization of cumulus convection in the Canadian climate centre general circulation model. *Atmosphere-Ocean*, *33*(3), 407–446. https://doi.org/10.1080/07055900.1995.9649539

Zhang, C. (2005). Madden-Julian Oscillation. *Reviews of Geophysics*, *43*(2). https://doi.org/10.1029/2004RG000158

Acknowledgments

1. Projects and Funding

This part is directly based on my published and my paper that is currently in review (**Behrens** et al. 2022; **Behrens** et al. 2024).

This thesis was funded from 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). Additionally it was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through the Gottfried Wilhelm Leibniz Prize awarded to Veronika Eyring (Reference No. EY 22/2-1). Moreover this thesis was funded from the EERIE project under the Horizon Europe research and innovation programme (Grant Agreement No. 101081383) by the European Union. The work related to this thesis used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID 1179 (USMILE) and 1083 (Climate Informatics), and the supercomputer JUWELS at the Jülich Supercomputing Centre (JSC) under the Earth System Modelling Project (ESM). Detailed information about the funding can be further found in the acknowledgements of the papers related to this thesis.

2. To the people

It is difficult to start acknowledging distinct persons in your thesis as you may have to prioritize. Therefore I start with an outstanding friend that helped me cope with all challenges during my PhD time. My biggest thanks go to Robert Zimmermann. He is an outstanding person and a true expert of the dynamics of the Earth system and in the stadium. You bolstered me a lot, gave brilliant feedback and ideas for my research, had basically every second an open ear whenever I needed it. Looking forward to more great research and conversations with you in the future! Next comes Erik Behrens, a true ocean modeller by heart from the super high-resolutions towards ocean general circulation models. I thank you for all the help and advice you are giving me in research and other topics. Moreover it helped me a lot to know that I am not the first one in my family that has to survive a long way towards graduation.

Coming now to professional acknowledgments, I thank my two supervisors Veronika Eyring and Pierre Gentine. I am grateful for their support and their guidance during my time as PhD student. Moreover I am glad that they shaped my research with a flavour of "big picture" science that I may have not recognized in some moments and aspects. A big thank you goes to Tom Beucler, for being such a great assistant supervisor. You helped me a lot towards my

graduation by navigating around potential pitfalls, prevent me from geese chasing and gave me a large role of duct tape at hand to stabilize my research projects. Basically the same applies to Fernando Iglesias-Suarez. You were an incredibly good assistant supervisor! Apart from an excellent taste with respect to football teams (although they may not be as dynamic as others), you helped to keep me on my toes everyday. Moreover you translated my super theoretical ideas into an human understandable form better than any machine learning algorithm. Who needs ChatGPT if you have Nando?!?

The next persons I would like to thank is the marvellous "Dream Team Number 1". Birgit Hassler, Axel Lauer and Lisa Bock, you combine an enormous knowledge about the Earth system with heart-warming human intelligence and a lot of PhD encouraging fun in one office. That is simply outstanding! You were the equivalent of a first aid kit both in physical and in emotional sense during my PhD, thanks a lot for that. The next person that I would like to thank is Manuel Schlund. You are a large inspiration when it comes to solving insurmountable research tasks and helped me as an East-German to cope with the US and the AGU. Moreover you are an excellent person to talk about science but also any other topic. I would like to move now to our DLR swimming crew. Kevin Debeire, you are an outstanding office mate, friend, hiker, swimmer, guitar hero and a lot more. I am really glad that you waited for me after your insane pace downhill in the Alps (another important reason why I survived my PhD). Aytac Pacal, you and me, we cannot compete against Kevin with respect to swimming or hiking that is for sure. But you are also an awesome office mate, a good friend and a person that is able to understand every bit of my "effervescent" irony, this is outstanding. So please excuse me for asking: "Where is your golden necklace?". The next person I would like to thank is Michaela Langer that is able to make even administrative things like Zeitnachweise funny. Thanks for all your time, your patience with me when you had to wait for my bullet points for the department meeting. Many thanks for all the funny moments during the Days of Hope and in the office. Homesickness for Northern Germany came often over me in hostile Bavaria, but thanks to Mierk Schwabe even Oberpfaffenhofen felt a little bit like Kiel or Altenholz. I am also grateful for all your advice and the nice discussions during my PhD.

Now coming to the Bremen group that I am a part of since almost a year. Thanks Tina Gier for being the emotional centre of action of our Bremen group. It is also great to graduate with you and it was great to be the second assistant auxiliary organizer of the PhD gaming evening under your lead. Moreover you are one of the most creative persons that I could think of and you are the only person that could handle Manni. I thank also Kathi Hafner for all the enjoyable conversation we have and for her iconic "impinging " stare to underline when I did something wrong. You are a rather small person, but you are a remarkable person and you know how to raise my mood. Big thanks goes out to Katja Weigel for stemming all the organization here in Bremen and helping with basically everything. Also I thank Evgenia Galystka for being a great person that cheered me up in difficult times and sharing a similar sense of irony like me. You and Soufiane can sell causal discovery so good that even I got convinced. Now coming to Soufiane Karmouche you are a great researcher and it is always fun to talk to you even about the most serious problems of the world. Many thanks further

to Jan-Hendrik Malles my new office mate. It is great to have in one moment some super scientific conversation and in the next moment we could talk about the golden era of Werder Bremen or some anecdotes about Scandinavian metal bands.

Furthermore a lot of thanks go out to the United States to Michael Pritchard, Griffin Mooers and Sungduk Yu. Without you my graduation would be simply impossible. I am thankful for all the enjoyable meetings, the large efforts of you to improve our collaboration. Moreover, only your massive technical help enabled us, with me as lead author, to explore "the eccentricities of modern machine learning" as Mike would call it. Thanks a lot and looking forward to see you at some point in the US in real life.

As a concluding part of the acknowledgements I would like to add a few words about some iconic person that I met during my phase from Bachelor towards PhD defence. Stefanie Schwarz, you are an outstanding friend for me and even managed to move from the beautiful Schondorf am Ammersee to Kiel. You made my life during Bachelor and Master studies really enjoyable. Thanks also for the endless hours of conversations and when meeting you it feels like in the good old times in Burse. Moreover it is impressive what you have achieved in geophysics and now transforming our world towards a renewable future. Great thanks also to Felix Clausen. You are the one that integrated shy greenish Gunnar in Burse and taught him quite a lot of valuable things how to handle people and to take yourself by far not serious. The same applies to Johannes Hollinderbäumer. Thank you for all the interesting and deep conversations that one could have with you over a coffee or two. Moreover I am thankful for your open ear whenever I had a really bad day during my PhD and for pushing me towards my limits with respect to running. Next Kiel Lauf, sure I will participate!

I am also deeply grateful to Richard Greatbatch. He is by far the greatest lecturer in climate science and theoretical general fluid dynamics that I know. You have such an outstanding background knowledge about dynamics of the atmosphere or ocean that it was quite a challenge for me to keep you always on your toes during your lectures. It was an honor to be one of your last Master students. Last but not least, I thank Christian Scharun for being a great friend and for giving me orientation what is possible if you can communicate your research in a crystal clear but also enjoyable way for society. It feels great to know such a rising star of German research. Especially if you know that you could ask him whatever you want and you will get an honest response. Thanks for all your aspirations in me and good vibes that helped a lot on the long way towards the graduation.

To whom I forgot to acknowledge personally, it is an honour to get so much support and to work in such a great environment. It was a long way to go but thanks to all your assistance I managed it to the finish line. "Thanks a lot, good luck and good night!"