Universität Bremen

Toward Physically Consistent, Accurate and Stable Machine Learning-Based Convection Parameterizations for ICON

Doctoral Dissertation of

Helge Heuer

October 2025

University of Bremen
Institute of Environmental Physics (IUP)

Toward Physically Consistent, Accurate and Stable Machine Learning-Based Convection Parameterizations for ICON

Doctoral Dissertation of
Helge Heuer

A thesis submitted in fulfillment of the requirements for the degree
Doktor der Naturwissenschaften (Dr. rer. nat.)

Primary Examiner: Prof. Dr. Veronika Eyring
Secondary Examiner: Prof. Dr. Pierre Gentine

Submission: 26 October 2025
Doctoral Colloquium: 06 February 2026

Abstract

Earth system models (ESMs) are important tools to project climate change, yet continue to have persistent systematic errors due to the representation of subgrid-scale processes, most notably atmospheric convection—a key driver of large-scale circulations such as the Hadley and Walker cells, as well as weather patterns such as thunderstorms. These errors contribute considerably to uncertainties in climate projections. Traditional convection parameterizations rely on physical assumptions and empirically derived relationships that fail to capture the full complexity of convective processes. This thesis contributes to showing that machine learning (ML) offers breakthroughs, leveraging high-fidelity simulations to learn data-driven parameterizations that better represent subgrid-scale dynamics. However, translating high offline ML performance into stable, physically consistent, and transferable online implementations in ESMs has proven challenging, often due to issues related to causality, scale separation, distributional shifts, and process separation.

This dissertation addresses these challenges through two complementary studies that advance the development, interpretation, and integration of ML-based convection parameterizations into the ICOsahedral Nonhydrostatic (ICON) model. The first study develops and benchmarks a suite of ML models, including deep learning and tree-based methods, trained on filtered and coarse-grained convective fluxes derived from storm-resolving ICON simulations over the tropical Atlantic. A filtering method to isolate convective contributions from other physical processes is meant to ensure that the ML models learn to represent deep convection. Offline, a U-Net architecture outperforms other models but exhibits non-causal dependencies on precipitating tracers, as revealed by explainable artificial intelligence (AI) analysis using SHapley Additive exPlanations (SHAP). Ablating these inputs yields a more physically interpretable and causally sound parameterization that demonstrates improvements in online stability, maintaining 180-day integrations in ICON while reducing biases in precipitation extremes compared to conventional schemes. However, a significant smoothing bias in the column water vapor distribution as well as biases in the mean temperature persist.

Building on these insights, the second study presents a proof-of-concept for cross-model transferability and long-term integrability. A bidirectional long short-term memory model trained on the global ClimSim dataset, derived from superparameterized Energy Exascale Earth System Model-Multiscale Modeling Framework (E3SM-MMF) simulations, is transferred to the ICON-A atmosphere model. Several innovations ensure physical consistency, accuracy, and robustness: removal of radiative tendencies to isolate the convective signal, physics-informed and vertical consistency losses, confidence-guided mixing with a conventional scheme, and additive input noise during training to enhance extrapolation and stability. This hybrid

AI-physics approach enables the stable multi-decadal (20-year) integration of an ML-based convection parameterization in ICON. Evaluating the resulting simulations against observations indicates improved precipitation statistics, including reduced root mean square error in the zonal mean, better spatial distribution, and improved precipitation extremes, as well as a more accurate spatial distribution of the near-surface temperature, relative to the reference ICON configuration. Furthermore, the developed scheme exhibits a physically interpretable regime behavior across column water vapor and stability metrics. However, the used training dataset has biases as well, e.g., the zonal mean precipitation does not match the observed climatology and conservation laws are not strictly enforced by the developed framework and would require a refined training dataset to do so. Moreover, the scheme is trained and evaluated at a relatively coarse horizontal resolution of ~160 km; implementing it in the currently developed version of the hybrid ICON model, which has a horizontal resolution of ~80 km, may require vertical interpolation and tuning.

Together, these studies demonstrate that ML-based parameterizations can become stable, interpretable, and transferable components of next-generation climate models. They highlight the critical importance of physical consistency, learning causal relationships, and robust training practices in achieving reliable long-term simulations, thereby demonstrating that multi-decadal stable hybrid simulations are achievable, paving the way toward more accurate and trustworthy climate projections.

Integrated Author's References

Parts of this thesis, including text, figures, and tables, have been published or submitted for publication in the following peer-reviewed studies. More details on this are given in Section 1.2 and at the beginning of the corresponding chapters. Two more co-authored publications are currently under review.

Heuer, **H.**, Schwabe, M., Gentine, P., Giorgetta, M. A., & Eyring, V. (2024). Interpretable Multiscale Machine Learning-Based Parameterizations of Convection for ICON. *Journal of Advances in Modeling Earth Systems*, *16*(8), e2024MS004398. https://doi.org/10.1029/2024MS004398

Heuer, **H.**, Beucler, T., Schwabe, M., Savre, J., Schlund, M., & Eyring, V. (2025). Beyond the Training Data: Confidence-guided Mixing of Parameterizations in a Hybrid AI-climate Model [Under Review for the Journal of Advances in Modeling Earth Systems]. *arXiv preprint arXiv:2510.08107*. https://doi.org/10.48550/arXiv.2510.08107

Yu, S., Hu, Z., Subramaniam, A., Hannah, W., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., Behrens, G., Busecke, J. J. M., Loose, N., Stern, C. I., Beucler, T., Harrop, B., **Heuer**, **H.**, Hillman, B. R., Jenney, A., Liu, N., White, A., Zheng, T., Kuang, Z., Ahmed, F., Barnes, E., Brenowitz, N. D., Bretherton, C., Eyring, V., Ferretti, S., Lutsko, N., Gentine, P., Mandt, S., Neelin, J. D., Yu, R., Zanna, L., Urban, N. M., Yuval, J., Abernathey, R., Baldi, P., Chuang, W., Huang, Y., Iglesias-Suarez, F., Jantre, S., Ma, P.-L., Shamekh, S., Zhang, G., & Pritchard, M. (2025). ClimSim-online: A Large Multi-Scale Dataset and Framework for Hybrid Physics-ML Climate Emulation. *Journal of Machine Learning Research*, *26*(142), 1–85. http://jmlr.org/papers/v26/24-1014.html

Contents

Abstract v

Integrated Author's References vii

1. Introduction 1
- 1.1. Key Science Questions . . . 3
- 1.2. Structure of the Thesis . . . 3

2. Scientific Background 5
- 2.1. Atmospheric General Circulation Models . . . 5
  - 2.1.1. Brief History of General Circulation Models . . . 5
  - 2.1.2. High-fidelity Atmospheric Models . . . 7
- 2.2. Conventional Convection Parameterizations . . . 12
  - 2.2.1. Adjustment Schemes . . . 12
  - 2.2.2. Moisture Convergence Schemes . . . 13
  - 2.2.3. Mass-Flux Schemes . . . 13
  - 2.2.4. Common Biases in Conventional Convection Parameterizations . . . 15
  - 2.2.5. The Tiedtke Convection Parameterization . . . 15
- 2.3. Selected Machine Learning Methods . . . 19
  - 2.3.1. Non-deep Learning Approaches . . . 19
  - 2.3.2. Deep Learning Approaches . . . 21
  - 2.3.3. Shapley Additive Explanations . . . 24
- 2.4. Machine Learning-Based Parameterizations . . . 25
  - 2.4.1. Emulation of Conventional Schemes . . . 27
  - 2.4.2. Learning from Multiscale Modeling Frameworks . . . 28
  - 2.4.3. Learning from High-Resolution Data . . . 30

3. Data and Methods 33
- 3.1. The ICOsahedral Nonhydrostatic Model . . . 33
- 3.2. The ClimSim dataset . . . 35
- 3.3. Relevant Data and Preprocessing for Chapter 4 . . . 36
  - 3.3.1. Computation of Output . . . 37
  - 3.3.2. Coarse-Graining . . . 39
  - 3.3.3. Filtering for Convection . . . 40
  - 3.3.4. Rescaling and Normalization . . . 42

3.4. Relevant Methods for Chapter 4 . . . 42
3.5. Relevant Data for Chapter 5 . . . 44
3.5.1. ClimSim and Cross-Validation Procedure . . . 44
3.5.2. "ClimSim Convection": Approximate Removal of Radiation for Training . . . 45
3.5.3. Datasets Used for Evaluation . . . 47
3.6. Relevant Methods for Chapter 5 . . . 47
3.6.1. Tiedtke Convection Scheme . . . 48
3.6.2. Machine Learning Scheme . . . 48
3.6.3. Loss Function . . . 50
3.6.4. Confidence-Guided Mixing . . . 52
3.6.5. Jointly Optimizing Performance and Inference Cost . . . 54
3.6.6. Additive Noise During Training for Improved Stability . . . 55
3.6.7. Online Coupling to ICON . . . 55

4. Interpretable multiscale Machine Learning-Based Parameterizations of Convection for ICON **57**
4.1. Introduction . . . 57
4.2. Results . . . 59
4.2.1. Machine Learning Model Benchmarking . . . 59
4.2.2. Explainability of U-Net and GBT . . . 65
4.2.3. Online Stability Tests . . . 70
4.3. Conclusions and Discussion . . . 76

5. Beyond the Training Data: Confidence-Guided Mixing of Parameterizations in a Hybrid AI-Climate Model **81**
5.1. Introduction . . . 81
5.2. Results . . . 83
5.2.1. Benchmarking with Observations . . . 83
5.2.2. Advantages of Physics-Informed Loss via Conservation Laws . . . 87
5.2.3. Process understanding: Why is the mixed model better than both the Tiedtke and pure ML model? . . . 87
5.2.4. Twenty-year AMIP run . . . 92
5.3. Summary . . . 94

6. Conclusion **99**
6.1. Overall Summary . . . 99
6.2. Outlook . . . 102

Appendix **105**
A. Supplementary Materials for Chapter 4 . . . 105
A.1. Section S2 . . . 114
A.2. Section S3 . . . 115

B. Supplementary Materials for Chapter 5 118
B.1. Non-dimensionalization of Residual Fluxes 118
B.2. The Hyperparameter Optimization Search Space and Offline R^2 Scores . 120
B.3. Additional Figures 121

Glossary **129**

List of Figures **133**

List of Tables **143**

References **145**

Acknowledgments **173**

1. Introduction

The accurate representation of atmospheric convection in Earth system models (ESMs) is arguably one of the most enduring and consequential challenges in climate modeling (Bony et al. 2015; Gentine et al. 2018; Schneider et al. 2017). Convection, particularly deep convection in the tropics, governs the vertical transport of heat, moisture, and momentum, driving large-scale circulations such as the Hadley and Walker cells (Neelin 2010). It plays a pivotal role in determining cloud formation, precipitation patterns, and the radiative balance of the planet. Despite its critical importance, convection occurs at spatial scales far below the grid resolution of typical global climate models (on the order of tens to hundreds of kilometers), necessitating the use of parameterizations: simplified, idealized, and often semi-empirical representations of subgrid-scale processes (Easterbrook 2023). These parameterizations have long been recognized as a dominant source of structural uncertainty in climate projections (Eyring et al. 2024b; Sherwood et al. 2014; Zelinka et al. 2020), particularly in estimates of the equilibrium climate sensitivity (ECS), a key metric that quantifies the long-term warming response to a doubling of atmospheric carbon dioxide (CO_2) (Council 1979; Manabe and Wetherald 1967). Remarkably, despite significant advances in computational power and physical understanding, the uncertainty range in ECS has changed little over the past decades, with values spanning 2–5 °C in the latest Coupled Model Intercomparison Project (Eyring et al. 2016; Forster et al. 2021). A substantial portion of this uncertainty can be traced back to differences in how models simulate tropical low-cloud feedbacks, in particular for cumulus and stratocumulus clouds over tropical oceans (Bony and Dufresne 2005; Brient and Schneider 2016; IPCC 2013; Schneider et al. 2017).

Traditional (often mass-flux-based) convection schemes, such as the widely used formulations by Tiedtke (1989) and Zhang and McFarlane (1995), are built upon simplified physical assumptions about convective triggering, entrainment, and closure. While these schemes have enabled meaningful progress in climate modeling, they are inherently limited by their structural simplicity and inability to capture the full complexity and nonlinearity of real-world convection. Common deficiencies include persistent biases in the position and structure of the intertropical convergence zone (ITCZ), such as the notorious "double-ITCZ" problem (Hwang and Frierson 2013; Satoh et al. 2019; Stevens et al. 2019b), an underestimation of extreme rainfall events (Christopoulos and Schneider 2021; Fosser et al. 2024; Stephens et al. 2010), a misrepresentation of convectively coupled waves like the Madden-Julian oscillation (Kuang et al. 2005; Lin et al. 2008), inaccuracies in reproducing teleconnection patterns such as the El Niño–Southern Oscillation (Mahajan et al. 2023), and a poor simulation of the diurnal cycle of precipitation (Anber et al. 2015; Christopoulos and Schneider 2021). These shortcomings are

not merely academic, they have direct implications for climate projections and may impede policy decisions and adaptation planning (Eyring et al. 2024b).

In recent years, machine learning (ML) has emerged as a promising approach to address the challenge of improving the representation of these processes. The emergence of storm-resolving models (SRMs) and multiscale modeling frameworks (MMFs), which explicitly simulate deep convection without relying on convection parameterizations at horizontal resolutions of 2–5 km, offers a powerful opportunity to generate high-fidelity training data for ML models. These simulations resolve individual convective cells and their interactions with larger-scale dynamics, providing a rich dataset from which subgrid processes can be diagnosed and emulated (Brenowitz and Bretherton 2018; Eyring et al. 2024a; Gentine et al. 2018). At the same time, ML, particularly deep learning, has demonstrated remarkable success in learning complex, nonlinear mappings from high-dimensional input-output relationships. This makes ML a promising tool for developing next-generation parameterizations that go beyond the rigid assumptions of traditional physics-based approaches.

In addition to emulating conventional convection schemes (Balogh et al. 2025; O'Gorman and Dwyer 2018; Zhong et al. 2024), ML-based convection and general moist subgrid physics parameterizations have been developed using data from MMF (Behrens et al. 2022, 2025; Beucler et al. 2021, 2024; Brenowitz et al. 2020; Chen et al. 2025; Gentine et al. 2018; Han et al. 2020, 2023; Hu et al. 2025; Iglesias-Suarez et al. 2024; Lin et al. 2025; Mooers et al. 2021a; Ott et al. 2020; Rasp et al. 2018; Wang et al. 2022b), and from SRMs (Beucler et al. 2024; Brenowitz and Bretherton 2018, 2019; Brenowitz et al. 2020; Krasnopolsky et al. 2013; Wang et al. 2022a; Watt-Meyer et al. 2024; Yuval and O'Gorman 2020; Yuval et al. 2021; Yuval and O'Gorman 2023), see Section 2.4. These studies have demonstrated success in parameterizing convection and subgrid cloud physics, but accurately representing subgrid convection remains challenging due to its complex and multiscale nature. Consequently, most of these studies have used idealized setups, like aquaplanets, and even under these simplifications, online stability (when the parameterization is coupled to the host ESM) is not guaranteed (Brenowitz et al. 2020; Lin et al. 2025; Rasp et al. 2018; Yuval and O'Gorman 2020). Stability crucially depends on technical details when training the ML model, the inclusion of specific variables and vertical levels (Brenowitz and Bretherton 2018, 2019), and the choice of training data (Rasp 2020). Moreover, high offline performance (i.e., when the model is evaluated in isolation, not coupled to the ESM) is not a sufficient condition for stable online integration (Lin et al. 2025; Yuval and O'Gorman 2020). Although methods exist to analyze the stability of ML-based schemes offline (Brenowitz et al. 2020), researchers typically rely on trial and error to test for how long their scheme runs stably when coupled online (Wang et al. 2022b).

Due to the aforementioned reasons, an online, accurate, and long-term stable ML-based convection parameterization suitable for operational hybrid ML-physics models has yet to be demonstrated. This dissertation contributes to this evolving paradigm by investigating the development, integration, and evaluation of ML–based convection parameterizations within the ICOsahedral Nonhydrostatic (ICON) modeling framework (Giorgetta et al. 2018; Zängl et al. 2015). It presents two research efforts that together aim to improve the representation of tropical

convection and its interactions with large-scale dynamics through data-driven approaches, while ensuring numerical robustness and physical consistency in coupled simulations.

1.1. Key Science Questions

This thesis advances the field of ML-based parameterizations for general circulation models by addressing the following three key science questions, with a focus on achieving skillful predictions, interpretability, robustness, and long-term stability:

1. How can we ensure that machine learning parameterizations learn physically consistent, causal, and interpretable relationships, rather than spurious correlations, when learning complex atmospheric convection?

2. Can machine learning parameterizations of subgrid convection improve the representation of convective processes in coarse-resolution models while maintaining long-term numerical stability?

3. To what extent can ML-based parameterizations be transferred across climate models, and how can training strategies be optimized to support robust hybrid climate modeling?

1.2. Structure of the Thesis

Parts of this thesis are already published (including text, figures, and tables) in two first-author studies, one already peer-reviewed (Heuer et al. 2024) and one currently under review by the *Journal of Advances in Modeling Earth Systems (JAMES)*. Another co-authored study (Yu et al. 2025) has been published in the *Journal of Machine Learning Research*. For this study, the author of this thesis contributed primarily by testing the new containerized online pipeline presented in the paper. A complete list of these publications is provided on Page vii. Two additional co-authored publications are currently in preparation.

Specifically, Chapter 4, Chapter 5, and parts of Chapter 3 are based on the two first-author publications (Heuer et al. 2024, 2025). The pronoun "we" is used in these chapters to enhance readability and to acknowledge the contributions of all co-authors. Unless stated otherwise, all content from these publications (text, figures, and tables) shown in this thesis originates from the author of this thesis.

This thesis is structured as follows: Chapter 2 introduces the scientific background and reviews the relevant literature upon which the presented studies build. Chapter 3 presents the data and methods relevant to the subsequent chapters. The data and methods introduced in Chapter 3, which directly reference Chapters 4 and 5, are based on the corresponding data and methodological frameworks detailed in Heuer et al. (2024, 2025). Chapter 4 is based on Heuer et al. (2024) and presents the development of ML-based convection schemes trained on storm-resolving regional ICON data. This chapter primarily addresses Key Science Question 1 while also presenting online results relevant to Key Science Question 2. Chapter 5, based on

Heuer et al. (2025), presents the development of an ML-based convection scheme trained on global MMF-based data and transferred to the ICON model. It addresses all three key science questions. A summary of the results presented in this thesis and an outlook are given in Chapter 6.

2. Scientific Background

This chapter presents the scientific background and reviews the relevant literature underpinning the research. It begins in Section 2.1 with a brief historical overview of atmospheric general circulation models, followed by a more detailed introduction of high-fidelity models. Section 2.2 discusses the traditional parameterization of convection, focusing on semi-empirical approaches commonly used in current modeling frameworks. In Section 2.3, key machine learning (ML) methods that are relevant to the results and methodologies developed in this thesis are introduced. Finally, Section 2.4 explores data-driven approaches to convection parameterization and reviews prior work in this field.

2.1. Atmospheric General Circulation Models

> You can never win competing with nature's complexity
>
> Syukuro Manabe

This section provides a brief historical overview of general circulation modeling and introduces two advanced high-fidelity modeling frameworks: storm-resolving models (SRMs) and multiscale modeling frameworks (MMFs) (also called superparameterized models). Both aim to explicitly resolve relatively small-scale dynamics, such as deep convection, thereby eliminating the need for conventional convection parameterizations. These next-generation models represent a paradigm shift in atmospheric simulation by explicitly capturing deep convective updrafts and downdrafts. Despite these advantages, coarse-resolution, non-storm-resolving general circulation models (GCMs) continue to play a vital role in climate research due to their computational efficiency, which enables long-term simulations spanning centuries, the representation of a broader range of processes in Earth system models (ESMs), and the generation of large ensembles (Eyring et al. 2024b). A promising approach to enhance the representation of key processes in such models using ML is discussed in Section 2.4.

2.1.1. Brief History of General Circulation Models

The foundation of modern climate modeling was laid in the mid-20th century with pioneering work in numerical weather prediction. Building on early successes by Jule Charney, Ragnar Fjørtoft, and John von Neumann, who conducted the first regional numerical forecasts over North America at a coarse resolution of approximately 736 km (Charney et al. 1950), Phillips (1956) published the first idealized GCM in 1956. This two-level, quasi-geostrophic model

simulated atmospheric dynamics on a single hemisphere with a horizontal resolution of 375 km × 625 km and already included parameterizations for non-adiabatic heating and friction. Later, the practice of parameterizing subgrid processes was introduced more formally by Smagorinsky (1963) for the diffusive lateral transfer of momentum and heat.

In 1965, Joseph Smagorinsky, Syukuro Manabe, and collaborators at the Geophysical Fluid Dynamics Laboratory developed a more sophisticated nine-level GCM of one hemisphere, albeit without realistic topography, and a horizontal resolution ranging from 320 km at the equator to 640 km at the pole (Edwards 2011; Manabe and Wetherald 1967; Smagorinsky et al. 1965). Around the same time, Yale Mintz and Akio Arakawa at the University of California, Los Angeles, introduced a two-level GCM incorporating realistic geography and global extent (Arakawa 1966; Edwards 2011) using a $7° \times 9°$ resolution, demonstrating the feasibility of simulating Earth's climate system on a planetary scale.

These pioneering models catalyzed rapid development across research institutions worldwide. As computational power expanded, GCMs improved quickly and grew in complexity. For example, coupling atmospheric models with ocean models, simple land models, and sea ice components (Manabe and Bryan 1969; Manabe et al. 1975) enabled more realistic representations of climate feedbacks. Over subsequent decades, these coupled systems evolved into comprehensive ESMs, integrating interactive carbon cycles, cryospheric processes, sulfate aerosols, dynamic vegetation, and more processes (Easterbrook 2023). Despite these advances, one of the main persistent limitations has remained: the inability to explicitly simulate convective and other small-scale processes, which has necessitated reliance on uncertain, empirical, and idealized parameterizations.

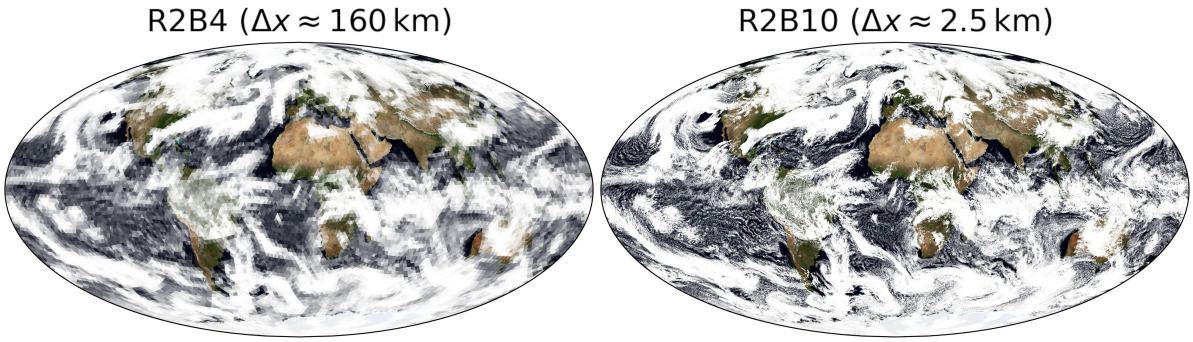

Figure 2.1.: Comparison of two horizontal resolutions (`R2B4` and `R2B10`) of an atmospheric state modeled by the ICOsahedral Nonhydrostatic (ICON) model (Giorgetta et al. 2018; Zängl et al. 2015). Shown is a snapshot of the cloud cover distribution from the 05.02.2020. Data is taken from the DYAMOND intercopmarison project (Stevens et al. 2019b). Background image: NASA Earth Observatory (https://neo.gsfc.nasa.gov/view.php?datasetId=BlueMarbleNG).

2.1.2. High-fidelity Atmospheric Models

Both SRMs and MMFs circumvent the ambiguities and structural uncertainties inherent to traditional parameterized convection schemes (Satoh et al. 2019; Stevens and Bony 2013) (see also Section 2.2.4) by employing horizontal resolutions typically finer than 5 km (Satoh et al. 2019; Stevens et al. 2019b). At these scales, deep convection (and other small-scale processes) can be explicitly resolved rather than parameterized. However, even at kilometer-scale resolutions, shallow clouds in remain largely unresolved and generally require parameterization (Schneider et al. 2017), which can introduce additional uncertainty and degrees of freedom. This partial resolution of convective processes gives rise to the so-called "gray zone" problem, where convection is neither fully resolved nor fully parameterized (Prein et al. 2015). To comprehensively resolve the full spectrum of convective motions and eliminate the "gray zone" problem, models would require grid spacings $\Delta x \lesssim 100\,\text{m}$ (Craig and Dörnbrack 2008; Prein et al. 2015). Improvements gained by explicitly resolving processes that would otherwise require parameterization come at a considerable increase in computational cost. For three-dimensional, lower-resolution models, a doubling of horizontal resolution typically increases computational demands by a factor of eight, following a $(\Delta x)^{-3}$ scaling due to the need for shorter time steps and increased grid points in the horizontal dimensions. In storm-resolving, non-hydrostatic models, where vertical resolution must also be refined to maintain numerical stability and physical accuracy, the scaling worsens to $(\Delta x)^{-4}$, implying a 16-fold increase in computational cost for each halving of grid spacing (Schneider et al. 2024). This unfavorable scaling causes long-term, global climate simulations at storm-resolving resolutions extremely computationally demanding (Eyring et al. 2024b), rendering, e.g., running ensembles of SRMs prohibitive given current supercomputing capabilities. For MMFs, the computational scaling is more favorable due to the idealized representation of subgrid processes, though significant computational resources are still required, as discussed below. The following two sections provide further details on both approaches, discussing their advantages over coarse-resolution GCMs as well as known biases.

Storm-Resolving Models

SRMs, also known as cloud-resolving models or convection-permitting models, represent a transformative advancement in atmospheric modeling, operating at horizontal resolutions of 5 km or less (Satoh et al. 2019; Stevens et al. 2019b; Weisman et al. 1997). At such scales below 10 km, the hydrostatic approximation, which is commonly used in coarser models, breaks down (Drake 2014). A comparison of a low-resolution snapshot with a resolution of $\Delta x \approx 160\,\text{km}$ and a high-resolution snapshot at $\Delta x \approx 2.5\,\text{km}$ is presented in Figure 2.1. At storm-resolving resolutions, non-hydrostatic dynamical cores are necessary, as vertical accelerations associated with convective updrafts and downdrafts become non-negligible compared to the gravitational acceleration. SRMs typically employ terrain-following hybrid sigma-altitude vertical coordinates, since pressure is no longer a monotonically decreasing function of altitude (Prill et al. 2022).

The first global SRM simulations were conducted using the Nonhydrostatic Icosahedral Atmospheric Model (NICAM), developed in Japan. Tomita et al. (2005) performed a one-week aquaplanet experiment at 3.5 km resolution (Satoh et al. 2019), followed two years later by Miura et al. (2007), who implemented a version incorporating realistic geography at the same resolution. Subsequently, several modeling centers developed or adapted global non-hydrostatic models capable of storm-resolving simulations (Satoh et al. 2019):

- In Germany, the ICON model was developed for global convection-permitting simulations (Giorgetta et al. 2018; Zängl et al. 2015).
- In the United States, multiple frameworks emerged, including the Model for Prediction Across Scales (Skamarock et al. 2012), the Finite-Volume Dynamical Core on the Cubed-Sphere (Lin 2004), the Goddard Earth Observing System Model (Putman and Suarez 2011), and the global version of the System for Atmospheric Modeling (SAM) (Khairoutdinov and Randall 2003).
- At the European Centre for Medium-Range Weather Forecasts (ECMWF), the Integrated Forecast System was extended into the kilometer-scale regime (Kühnlein et al. 2019; Smolarkiewicz et al. 2016; Wedi et al. 2015).

The DYAMOND intercomparison study by Stevens et al. (2019b) evaluates multiple SRMs under the same atmospheric initial conditions and prescribed sea surface temperatures (SSTs) and sea-ice concentrations provided by ECMWF. The simulations are compared both among models and against satellite observations. Despite the models being in relatively early stages of development, many having never been run in this configuration or been tuned for this application, the results show remarkable consistency in simulating key aspects of the climate system. The models reproduce realistic patterns of the general circulation, outgoing longwave radiation, precipitation, and precipitable water (Stevens et al. 2019b). Visually, cloud fields closely resemble those observed in satellite imagery. Furthermore, SRMs demonstrate skill in capturing tropical cyclone genesis and intensification statistics, as well as the diurnal cycle of precipitation, features notoriously difficult to simulate in conventional GCMs (Christopoulos and Schneider 2021; Judt et al. 2021).

Ongoing initiatives such as Deep Numerical Analysis in Japan, Next Generation Earth Modelling Systems (NextGEMS), and Destination Earth in Europe are advancing SRMs toward decadal and multi-decadal climate projections (Takasuka et al. 2024), marking a transition from process-oriented studies to operational climate simulation.

SRMs offer several compelling advantages (Satoh et al. 2019), including:

- The explicit resolution of mesoscale convective systems allows a direct analysis of storm dynamics and organization.
- The multiscale nature of convection is naturally represented, including realistic simulations of the Madden-Julian oscillation (MJO).

- Cloud-scale circulations are directly coupled to microphysical processes as opposed to coupling to convective schemes.
- High-resolution output enables more direct comparison with remote sensing and surface observations as scales are typically more comparable.
- High-resolution flow fields improve tracer transport, offering more accurate links between surface emissions of substances and their atmospheric concentrations.
- Improved precipitation patterns including the spatial distribution (mesoscale organization), more realistic propagation of precipitating systems, and an improved diurnal cycle (Stevens et al. 2020).
- Improved representation of extreme events and regional projections of climate change (Palmer 2014)
- Monsoons, tropical cyclones, and general precipitation are better captured (Schneider et al. 2017)

Despite these strengths, SRMs are not without limitations. They still require parameterizations for subgrid-scale processes such as turbulence, radiation, and cloud microphysics. Although they typically require less tuning than coarse GCMs (Stevens et al. 2019b), calibrating the remaining parameters is considerably more computationally expensive. Some common biases include:

- Although tropical cyclones are represented more realistically compared to coarse GCMs, there are still biases regarding the number, intensity, size, and structure of tropical cyclones (Judt et al. 2021).
- New turbulence parameterizations to accurately represent the planetary boundary layer, shallow convection, subgrid-scale cloud cover, and turbulent fluxes related to deep convective systems at kilometer scales have to be developed (Prein et al. 2015). In this regime, the assumptions for traditional parameterizations for coarse GCMs break down, while the specified processes are only partially resolved; this is also referred to as the convective "gray zone" (Honnert et al. 2020; Prein et al. 2015).
- For example, SRMs with typical horizontal resolutions of approximately 5 km operate directly within the convective "gray zone", where partially unresolved convection can lead to the simulation of overly intense rainfall events (Kendon et al. 2021).
- Challenges in representing land-surface processes were observed (Kendon et al. 2021).
- Imbalances in energy and water budgets, likely linked to the representation of shallow cloud, turbulence and microphysics (Stevens et al. 2019b).

- The well-known double-intertropical convergence zone (ITCZ) bias, with excessive precipitation over the tropical Pacific and Indian Oceans is still exhibited by storm-resolving models (Schneider et al. 2024).

These challenges underscore that while SRMs eliminate the need for deep convection parameterizations, they are not free of model errors and introduce new demands on the fidelity of other physical components.

Multiscale Modeling Frameworks

An alternative pathway toward high-fidelity representation of convection is multiscale modeling frameworks (MMFs). First proposed as the "cloud-resolving convection parameterization (CRCP)" by Grabowski and Smolarkiewicz (1999), this approach typically embeds two-dimensional SRMs with periodic boundary conditions within each column of a coarser global GCM. The embedded SRMs operate on a kilometer-scale grid and explicitly resolve storm-scale dynamics, such as updrafts, downdrafts, anvils, and mesoscale organization, thereby rendering conventional convection parameterizations unnecessary. The proposed framework was subsequently applied in an idealized setting by Grabowski (2001).

The development was quickly advanced by Khairoutdinov and Randall (2001), who implemented the CRCP framework with realistic topography within the Community Atmosphere Model, giving rise to the Superparameterized Community Atmosphere Model (SPCAM). SPCAM simulations produced reasonable simulations of tropical convection and large-scale dynamics with minimal tuning, a significant improvement over conventional parameterizations in coarse GCMs, which often require extensive calibration. This success sparked widespread interest and catalyzed further development of the MMF approach, particularly at Colorado State University's Center for Multiscale Modeling of Atmospheric Processes (Khairoutdinov et al. 2005; Randall et al. 2003; Randall 2013).

Other major modeling centers adopted MMF development, including the NASA Goddard Space Flight Center (Tao et al. 2009) and the U.S. Department of Energy's Energy Exascale Earth System Model-Multiscale Modeling Framework (E3SM-MMF) (Hannah et al. 2020). These efforts underscore the broad recognition of MMFs as a powerful tool for advancing the fidelity of climate simulations, particularly in representing convective processes.

Several extensions to the original MMF framework have since been proposed to address limitations associated with the use of 2D SRMs. One promising advancement is the so-called quasi-3D MMF (Q3D MMF), in which each GCM column hosts two perpendicular 2D SRMs of finite width that interact through shared lateral boundary conditions. This design more effectively approximates three-dimensional dynamics without the computational burden of a full 3D SRM (Arakawa 2004; Jung and Arakawa 2010). A key advantage of the Q3D MMF is its asymptotic consistency: as the host model's grid spacing decreases, the representation converges toward that of a global SRM.

The MMF framework offers numerous advantages over conventional parameterizations (Randall et al. 2003), including:

- It allows for the explicit (but idealized) representation of deep convection at the subgrid scale, including critical processes such as mesoscale convective organization, downdrafts, and anvils.
- MMFs have demonstrated success in simulating the MJO with realistic features
- Fractional cloudiness emerges naturally from the SRM grid cells
- The framework enables the explicit simulation of convectively generated gravity waves.
- Better representation of both light and intense precipitation, with a large improvement for extreme precipitation compared to conventionally parameterized models (Li et al. 2012)

From a computational standpoint, MMFs strike a pragmatic balance between accuracy and efficiency. As they are able to maintain a relatively coarse global grid while resolving clouds locally via embedded 2D SRMs, they require only a fraction of the computational resources needed for global SRMs (Satoh et al. 2019). For example, Randall et al. (2003) estimate that the costs of running MMF simulations are approximately 100 to 1,000 times more costly than conventionally parameterized GCMs. While substantial, this increase is modest compared to the computational costs of a global SRM, which would raise computational costs by a factor of roughly 10^6 (Randall et al. 2003). Furthermore, MMFs are almost *embarrassingly parallel* (Randall et al. 2003): the embedded SRMs operate independently across grid columns and typically do not exchange information directly, enabling efficient distribution across processor cores.

However, several limitations arise from the inherent idealizations typically used in MMFs:

- The use of 2D SRMs with limited spatial extent can affect the mean state response and lead to, e.g., a too moist lower atmosphere and an associated shortwave cloud effect (Pritchard et al. 2014).
- Difficulties representing (convective) momentum transport due to the 2D nature of the SRMs (Tulich 2015; Woelfle et al. 2018). This can affect, e.g., mean precipitation, winds, and spatiotemporal variability of tropical convection (Yang et al. 2022).
- The Representation of randomly distributed or clustered convection, especially in low wind environments, can be biased for 2D SRMs (Tompkins 2000).
- Biases in mean precipitation with too much rain in the ITCZ region and too little over the Amazon (Liu et al. 2023).
- Unphysical "checkerboard" patterns in cloud-related fields on the grid-scale have been observed in E3SM-MMF (and with much less frequency of occurrence in SPCAM), potentially related to the artificial scale gap between SRM and GCM dynamics (Hannah et al. 2022).

- The double-ITCZ bias can still be an issue for the MMF (Hannah and Pressel 2022; Kooperman et al. 2016).

Nonetheless, MMFs remain an important bridge between traditional GCMs and fully resolved SRMs, offering a computationally tractable platform for exploring convective-scale processes in a global context.

2.2. Conventional Convection Parameterizations

> Of the many subgrid-scale processes that must be represented in numerical models of the atmosphere, cumulus convection is perhaps the most complex and perplexing.
>
> Emanuel and Raymond (1993)

Atmospheric convection, particularly deep moist convection, plays a critical role in the vertical transport of heat, moisture, and momentum, thereby influencing large-scale circulation patterns, cloud formation, and precipitation. To explicitly resolve deep convection, horizontal resolutions of less than approximately 5 km have to be used (Drake 2014; Satoh et al. 2019; Stevens et al. 2019b; Weisman et al. 1997). However, due to the coarse spatial resolution of GCMs typically used in climate modeling (Eyring et al. 2016; Haarsma et al. 2016), individual convective elements such as updrafts and downdrafts cannot be resolved. As a result, convection must be represented through parameterizations, models which approximate the influence of subgrid-scale processes as a function of resolved-scale variables. These typically involve a number of empirical relationships and idealized assumptions about the modeled interactions.

Over the decades, various approaches to parameterize convection have been developed, broadly categorized into three main types: adjustment schemes, moisture convergence schemes, and mass-flux schemes. Each represents a different level of physical sophistication and computational complexity.

2.2.1. Adjustment Schemes

Adjustment schemes represent the earliest and simplest form of convective parameterization. The core concept is based on the assumption that when the atmosphere becomes convectively unstable, typically diagnosed when the lapse rate exceeds the moist adiabatic profile, it rapidly adjusts toward a more stable reference state, often following the moist adiabat. This adjustment occurs instantaneously or over a short timescale, redistributing temperature and moisture vertically to eliminate instability. One of the earliest examples is the scheme by Manabe et al. (1965), which applied moist convective adjustment to both temperature and water vapor profiles in a climate model, effectively preventing super-moist-adiabatic lapse rates. A refinement of this idea is the Betts-Miller scheme (Betts and Miller 1993), which introduces a relaxation of temperature and humidity profiles toward observationally informed reference equilibrium

profiles for shallow and deep convection, separately over a finite timescale, rather than enforcing an instantaneous adjustment. This allows for a more gradual and physically plausible response to convective instability.

The primary strengths of adjustment schemes lie in their simplicity and low computational cost, making them suitable for early-generation climate models. However, their weaknesses are significant: they lack a dynamical representation of convective updrafts and downdrafts, fail to capture the timing and lifecycle of convection, and do not account for convective memory or feedbacks with large-scale dynamics. As such, they provide only a crude approximation of real convective processes.

2.2.2. Moisture Convergence Schemes

Moisture convergence schemes take a different approach, linking convection directly to the large-scale dynamics through the horizontal convergence (and surface fluxes) of moisture. The fundamental premise is that precipitation results directly from the net accumulation of water vapor within a model column due to large-scale advection.

The most prominent example is the Kuo scheme (Kuo 1965, 1974), in which deep convection is activated when atmospheric conditions are sufficiently unstable to support deep convection and when the large-scale flow provides sufficient lift to trigger it. A key strength of this approach is its direct coupling between convection and large-scale dynamics. However, the scheme suffers from several limitations due to its simplicity. It assumes that water is consumed by convection at the rate it is supplied by the large-scale dynamics. Although this condition was later modified to allow fractional consumption of the large-scale moisture supply, the assumption fundamentally violates causality, as convection is not caused by the large-scale water supply (Emanuel and Raymond 1993). Furthermore, the scheme fails to reproduce realistic vertical profiles of convective heating (Emanuel and Raymond 1993).

2.2.3. Mass-Flux Schemes

Mass-flux schemes represent a major advancement in convective parameterization, offering a more physics-based treatment of convection by modeling discrete convective elements: updrafts and downdrafts. These actively transport heat, moisture, and momentum across atmospheric layers. Mass-flux schemes are grounded in the idea that the net effect of convection on the large-scale environment can be approximated by the vertical fluxes associated with these plumes, weighted by their fractional area coverage.

Two subcategories exist: spectral (Arakawa and Schubert 1974; Baba 2019; Grell and Dévényi 2002; Moorthi and Suarez 1992) and bulk mass-flux schemes (Bougeault 1985; Gregory and Rowntree 1990; Kain and Fritsch 1990, 1993; Tiedtke 1989; Zhang and McFarlane 1995). Spectral schemes model the full spectrum of convective elements using a continuous distribution of cloud types or plume properties, such as fractional entrainment rate and detrainment level. In contrast, bulk mass-flux schemes assume a single representative plume per grid column, significantly reducing complexity while retaining essential physical processes (Stensrud 2007;

Yanai et al. 1976). As a result, bulk schemes are more commonly used in operational atmospheric models (Giorgetta et al. 2018; Golaz et al. 2022; Neale et al. 2010; Roberts et al. 2018; Roehrig et al. 2020; Stevens et al. 2013; Walters et al. 2019).

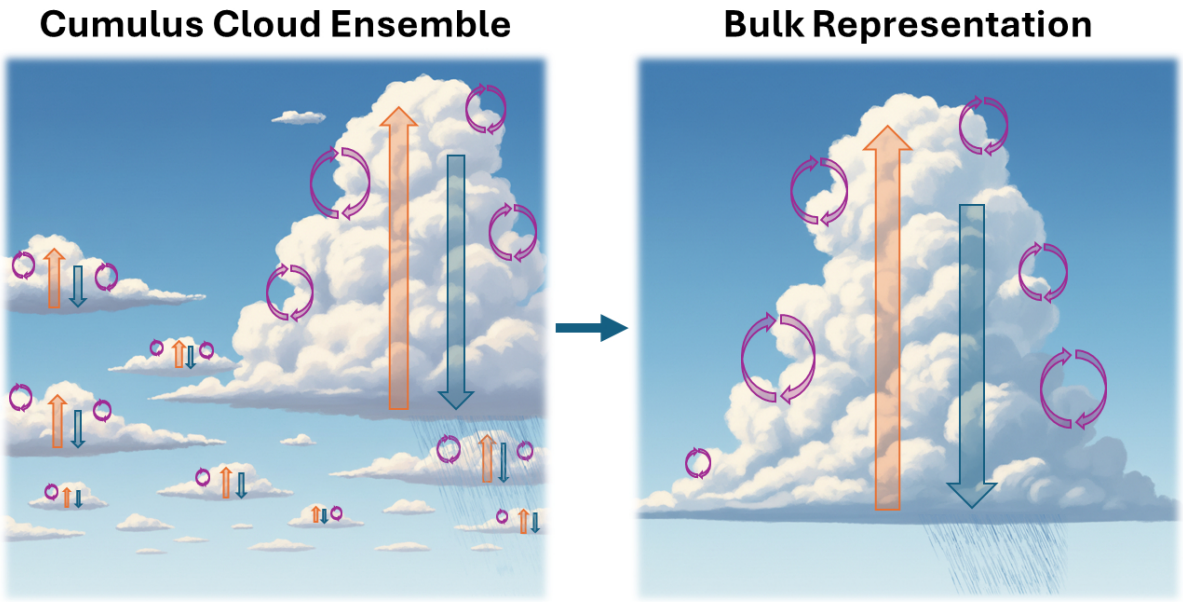

Figure 2.2.: Visualization of an ensemble of convective cumulus clouds (left) and the corresponding bulk mass-flux scheme representation (right). Updrafts and downdrafts are visualized as orange and blue arrows, respectively. Entrainment and detrainment are indicated by the violet arrows. The grid box is assumed to be much larger than the area depicted, with the environment subsidence region occupying most of it.

Mass-flux schemes are more sensitive to their local environments than adjustment schemes or the Kuo scheme, which do not try to model the vertical transport related to convection but instead prescribe a final state after the convective adjustment took place. However, mass-flux schemes typically rely on a number of parameters that are poorly constrained by observations yet have a large influence on the modeled convection (Stensrud 2007). The mass-flux schemes discussed here are all based on the assumption of convective quasi-equilibrium (CQE), meaning that convection reacts on sufficiently short timescales to rapidly remove instabilities generated by large-scale dynamics, thereby remaining almost in equilibrium with the large-scale environment (Arakawa and Schubert 1974; Bechtold et al. 2014; Yano and Plant 2012). This principle also underlies adjustment schemes, in which vertical profiles of temperature and humidity relax toward stable conditions when instability is diagnosed, also holding convective available potential energy (CAPE) approximately constant. Observationally, however, the CQE assumption remains contested (Neelin et al. 2008; Palmer 2019).

Notable examples among bulk mass-flux schemes include Bougeault (1985), Gregory and Rowntree (1990), the Zhang-McFarlane scheme (Zhang and McFarlane 1995), and the Kain–Fritsch scheme (Kain and Fritsch 1990, 1993). The latter employs a CAPE-based closure and includes both deep convective updrafts and downdrafts with entrainment and detrain-

ment (Stensrud 2007). Another important parameterization is the Tiedtke scheme (Tiedtke 1989), a bulk mass-flux scheme that distinguishes between penetrative, shallow, and mid-level convective regimes and incorporates physically motivated closure assumptions for each type. This scheme is used in the ICON model and is therefore of particularly important for this thesis. Further details will be provided, after a discussion of common errors associated with conventional convection parameterizations in Section 2.2.5.

2.2.4. Common Biases in Conventional Convection Parameterizations

Despite advances, the use of conventional convection parameterizations contributes to persistent errors across climate and weather models. These include systematic errors in the position and structure of the ITCZ, often simulated as a double-ITCZ with excessive precipitation in the tropical Southern Hemisphere (Hwang and Frierson 2013; Satoh et al. 2019; Stevens et al. 2019b). Precipitation patterns are frequently misrepresented in both magnitude and spatial distribution, with excessive light, drizzle-like rainfall and insufficient simulation of extreme events (Christopoulos and Schneider 2021; Fosser et al. 2024; Stephens et al. 2010). Many schemes fail to capture key modes of tropical variability, such as convectively coupled equatorial waves, including the MJO (Kuang et al. 2005; Lin et al. 2008), a shortcoming attributed, e.g., to the inadequate representation of convective organization (Moncrieff 2019). Teleconnection patterns associated with the El Niño–Southern Oscillation may also be misrepresented (Mahajan et al. 2023). Additionally, the diurnal cycle of convection, particularly over land, is often poorly simulated, with peak rainfall occurring earlier than observed (Anber et al. 2015; Christopoulos and Schneider 2021).

These persistent errors underscore the limitations of current parameterizations and motivate ongoing efforts to improve their physical realism, particularly through machine learning approaches, high-resolution modeling, and multiscale modeling frameworks.

2.2.5. The Tiedtke Convection Parameterization

Among the most widely used mass-flux schemes in modern atmospheric modeling is the convection parameterization developed by Tiedtke (1989). Designed for use in coarse GCMs, the Tiedtke scheme employs a bulk mass-flux approach, representing the net convective effect through simplified, one-dimensional updrafts and downdrafts. It was originally implemented in the ECMWF model and has since been adapted for use in other models, including the ICON model (Giorgetta et al. 2018; Zängl et al. 2015).

Core Framework

The Tiedtke scheme is based upon the principle that convection can be represented as a collection of organized updrafts and downdrafts, each transporting heat, moisture, and momentum vertically. Rather than resolving a full spectrum of cloud types, the scheme uses a simplified cloud ensemble modeled as a single bulk plume. This approach follows diagnostic studies

by Yanai et al. (1973), which successfully characterized observed convective systems using a bulk model framework. The scheme explicitly includes cumulus downdrafts which can be important for the large-scale heat and moisture budgets (Houze Jr and Betts 1981). Following Tiedtke (1989), the large-scale budget equations for dry static energy s and specific humidity q_v can be formulated as follows:

$$\begin{aligned}\frac{\partial \overline{s}}{\partial t} + \overline{\mathbf{v}} \cdot \nabla \overline{s} + \overline{w}\frac{\partial \overline{s}}{\partial z} = &- \frac{1}{\overline{\rho}}\frac{\partial}{\partial z}\left(M_u s_u + M_d s_d - (M_u + M_d)\,\overline{s}\right) \\ &+ L_v\left(c_u - e_d - \tilde{e}_l - \tilde{e}_p\right) - \frac{1}{\overline{\rho}}\frac{\partial}{\partial z}\left(\overline{\rho}\,\overline{w's'}\right)_\mathrm{tu} + \overline{Q}_R,\end{aligned} \tag{2.1}$$

$$\begin{aligned}\frac{\partial \overline{q}_\mathrm{v}}{\partial t} + \overline{\mathbf{v}} \cdot \nabla \overline{q}_\mathrm{v} + \overline{w}\frac{\partial \overline{q}_\mathrm{v}}{\partial z} = &- \frac{1}{\overline{\rho}}\frac{\partial}{\partial z}\left(M_\mathrm{u} q_\mathrm{v,u} + M_\mathrm{d} q_\mathrm{v,d} - (M_\mathrm{u} + M_\mathrm{d})\,\overline{q}_\mathrm{v}\right) \\ &+ \left(c_\mathrm{u} - e_\mathrm{d} - \tilde{e}_\mathrm{l} - \tilde{e}_\mathrm{p}\right) - \frac{1}{\overline{\rho}}\frac{\partial}{\partial z}\left(\overline{\rho}\,\overline{w'q'_\mathrm{v}}\right)_\mathrm{tu},\end{aligned} \tag{2.2}$$

with the horizontal velocity vector $\mathbf{v}$, vertical velocity w, time t, height z, density ρ, mass flux M, latent heat of vaporization L_v, condensation c, evaporation e, and radiative heating $\overline{Q}_R$. Overbars indicate grid-box averages, primes denote deviations from these horizontal averages, and subscripts u and d refer to updrafts and downdrafts, respectively. The tilde denotes environmental (non-cloudy) averages so that the evaporation of detrained cloud air is denoted as $\tilde{e}_l$ and the evaporation of precipitation in the unsaturated subcloud layer as $\tilde{e}_p$. The two terms with subscripts tu represent boundary layer turbulence, which is typically parameterized by a seperate scheme in GCMs. The environmental air occupies a much larger fraction of a given grid box compared to the area covered by updrafts and downdrafts in large-scale models. This justifies the already applied approximation of $\tilde{x} = \overline{x}$, where x is a transported quantity such as s or q_v.

To determine the mass flux, steady-state conditions are assumed for both updrafts and downdrafts. Under this assumption, the following bulk updraft equations can be derived for the stationary fluxes of mass, dry static energy, water vapor, and cloud liquid water:

$$\frac{\partial M_\mathrm{u}}{\partial z} = E_\mathrm{u} - D_\mathrm{u}, \tag{2.3}$$

$$\frac{\partial \left(M_\mathrm{u} s_\mathrm{u}\right)}{\partial z} = E_\mathrm{u}\overline{s} - D_\mathrm{u} s_\mathrm{u} + L_\mathrm{v}\overline{\rho} c_\mathrm{u}, \tag{2.4}$$

$$\frac{\partial \left(M_\mathrm{u} q_\mathrm{v,u}\right)}{\partial z} = E_\mathrm{u}\overline{q}_\mathrm{v} - D_\mathrm{u} q_\mathrm{u,v} - \overline{\rho} c_\mathrm{u}, \tag{2.5}$$

$$\frac{\partial \left(M_\mathrm{u} q_\mathrm{l,u}\right)}{\partial z} = -D_\mathrm{u}\overline{q}_\mathrm{l} - \overline{\rho} c_\mathrm{u} - \overline{\rho} G_\mathrm{p}, \tag{2.6}$$

with the rates of mass entrainment and detrainment per unit length E resp. D, the cloud liquid water content q_l, and the conversion rate from cloud water to rain G_p. Equations for the transport of momentum and the corresponding downdraft transport equations can be derived in analogously; further details are provided in Nordeng (1994) and Tiedtke (1989).

The remaining challenge lies in determining the entrainment and detrainment rates and in closing the problem, i.e., relating the problem back to large-scale conditions to provide boundary conditions at cloud base for Equations (2.3)–(2.6). In the Tiedtke scheme, entrainment and detrainment are partitioned into organized and turbulent components. The organized entrainment E_u^{org} for deep and midlevel convection is directly related to moisture convergence, while the turbulent parts E_u^{tu} are set directly proportional to the mass flux:

$$E_u^{org} = -\frac{\overline{\rho}}{\overline{q}_v}\left(\overline{\mathbf{v}}\cdot\nabla\,\overline{q}_v + \overline{w}\frac{\partial\overline{q}_v}{\partial z}\right), \quad E_u^{tu} = \epsilon_u M_u. \tag{2.7}$$

Where ϵ_u is the fractional entrainment and is chosen as a constant $\epsilon_{u,0}$ for deep and midlevel convection, and as $(3\,\epsilon_{u,0})$ for shallow convection. The closures used to determine the cloud base mass flux are closely related to moisture convergence and will be elaborated upon in the following section, as slightly different closures are applied for the considered convective types.

Representation of Convective Types

One of the defining features of the Tiedtke scheme is its distinction among three physically distinct types of convection, each governed by different triggering mechanisms and closure assumptions:

1. **Penetrative (Deep) Convection:**

   This type occurs in regions of large-scale convergence, typically associated with low-level synoptic scale convergence (Tiedtke 1989). The closure is based on a moisture convergence hypothesis, inspired by Kuo (1965, 1974) and Lindzen (1981), in which the entrainment of environmental air through the cloud base and sides is directly proportional to the large-scale moisture convergence in the subcloud layer (Tiedtke 1989). This closure ensures that deep convection is sustained by the large-scale supply of moisture. It can be formulated in the following way:

$$\begin{aligned}&\left(M_u(q_{v,u}-\overline{q}_v)+M_d(q_{v,d}-\overline{q}_v)\right)\Big|_B\\ &=-\int_0^B\left(\overline{\mathbf{v}}\cdot\nabla\,\overline{q}_v+\overline{w}\frac{\partial\,\overline{q}_v}{\partial z}+\frac{1}{\overline{\rho}}\frac{\partial}{\partial z}\left(\overline{\rho}\,\overline{w'q'_v}\right)_{tu}\right)\overline{\rho}\,dz,\end{aligned} \tag{2.8}$$

   where B denotes the cloud base height and $|_B$ indicates that the corresponding term is evaluated at cloud base.

2. **Shallow Convection:**

   This type occurs in regions with weak or negligible large-scale convergence, such as trade wind cumuli beneath a subsidence inversion or fair-weather cumulus clouds. Shallow convection is primarily driven by surface fluxes and subcloud-layer turbulence (Tiedtke 1989). The scheme assumes that moisture supply from surface evaporation maintains the cloud layer, with a closure based on the same mechanism as for deep convection but

driven mostly by surface evaporation. Therefore, Equation (2.8) is also used for shallow convection. Importantly, the scheme accounts for overshooting cumuli, allowing shallow clouds to penetrate into the inversion layer. This improves the simulation of the inversion layer strength and the distribution of moisture within the cloud layer (Tiedtke 1989).

3. **Midlevel Convection:**

   This type refers to convection initiated above the boundary layer, commonly observed in rainbands at warm fronts and within the warm sectors of extratropical cyclones (Browning et al. 1973; Herzegh and Hobbs 1980; Houze Jr et al. 1976; Tiedtke 1989). It typically occurs when large-scale ascent lifts moist air to its level of free convection (LFC), where instability exists aloft (Blanchard et al. 2021; Tiedtke 1989). The upward mass flux at the LFC $M_u|_B$ is set equal to the vertical mass transport by the large-scale flow at that level $\overline{\rho}|_B\overline{w}|_B$, thereby linking midlevel convection directly to synoptic-scale dynamics:

$$M_u|_B = \overline{\rho}|_B\overline{w}|_B. \tag{2.9}$$

The closure assumptions outlined above are grounded in observational evidence regarding the synoptic conditions that favor specific convection types (Tiedtke 1989). Specifically, penetrative and midlevel convection are sustained by large-scale moisture convergence and ascent, respectively. Shallow convection is maintained by the supply of moisture due to surface evaporation. This approach ensures that convection is tied to the broader synoptic environment, enhancing the spatial and temporal coherence of simulated rainfall.

In addition to heat and moisture, the Tiedtke scheme accounts for the transport of momentum by convective updrafts, downdrafts, and cumulus-induced environmental subsidence. This enables the scheme to represent the deceleration of zonal winds in the upper troposphere as observed by diagnostic studies (Sui and Yanai 1986; Tiedtke 1989).

Despite the overall improvements over the previously used Kuo scheme at ECMWF, Tiedtke (1989) acknowledges several limitations of the scheme. The vertical profiles of convective heating and drying, and thus the quality of the simulation, are very sensitive to the choice of numerical discretization scheme. Moreover, considerable uncertainty introduced by observationally weakly constrained parameters such as entrainment and detrainment rates. The scheme also tends to produce excessive convective heating in the middle and upper troposphere, an overly dry lower troposphere, excessive moistening in the upper troposphere, and underestimated convection in the extratropics.

Implementation in the ICON Model

In the ICON model, the Tiedtke scheme has been modified and updated to incorporate improvements proposed by Nordeng (1994). A key enhancement is the introduction of a CAPE-based closure for deep convection, which replaces the original moisture convergence closure. In this formulation, the cloud base mass flux is proportional to the CAPE divided by a newly introduced relaxation timescale, τ. Furthermore, organized detrainment is defined as the

loss of total mass flux due to the detrainment of clouds that have lost their buoyancy (Nordeng 1994). This process is parameterized in terms of the vertical derivative of the convective cloud area fraction. Another modification introduced by Nordeng (1994) links entrainment to buoyancy by recognizing that vertical acceleration induces inflow into updrafts due to mass continuity. Further details on the implementation of the scheme in ICON can be found in Möbis and Stevens (2012) and Giorgetta et al. (2018).

2.3. Selected Machine Learning Methods

> [Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.
>
> Arthur Samuel, 1959

ML has emerged as a fundamental component of modern artificial intelligence research, enabling systems to identify patterns, make predictions, and extract insights from complex datasets through algorithmic models rather than rule-based programming. Broadly, ML methods can be categorized into **supervised learning**, where models learn a mapping from inputs to known outputs; **unsupervised learning**, which discovers hidden structures in unlabeled data; **semi-supervised learning**, combining both labeled and unlabeled data; and **reinforcement learning**, where agents learn optimal behaviors through interaction with an environment via rewards and penalties (Géron 2022).

In this dissertation, I employ supervised learning methods, deploying a diverse set of classical (non-deep) and deep learning models across two studies. These models were selected based on their interpretability, scalability, performance, and suitability for the specific tasks at hand.

2.3.1. Non-deep Learning Approaches

Linear Regression

As models with the lowest algorithmic complexity, linear regression techniques were used as baseline predictors. These models assume a linear relationship between an input vector $\boldsymbol{x} \in \mathbb{R}^d$ and a target vector $\boldsymbol{y} \in \mathbb{R}^k$, approximating it as

$$\hat{\boldsymbol{y}} = \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}, \tag{2.10}$$

with the prediction $\hat{\boldsymbol{y}} \in \mathbb{R}^k$, weight matrix $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ and bias vector $\boldsymbol{b} \in \mathbb{R}^k$. Despite their simplicity, linear models offer strong interpretability and serve as important benchmarks for evaluating more complex architectures.

To address issues such as overfitting and multicollinearity (Chan et al. 2022), ordinary least squares regression can be extended with regularization techniques:

- **Lasso regression** (Least Absolute Shrinkage and Selection Operator Regression; Tibshirani (2018)) adds an L_1 penalty term $\lambda \|\boldsymbol{W}\|_1$ to the loss function, promoting sparsity by

shrinking some coefficients exactly to zero, thereby performing implicit feature selection. The regularization coefficient λ is a hyperparameter set prior to training time to control the strength of regularization.

- **Ridge regression** (Hoerl and Kennard (1970)) introduces an L_2 penalty $\lambda \|\boldsymbol{W}\|_2^2$, which shrinks coefficients toward zero without eliminating them, improving model stability in the presence of correlated predictors.

In general, regularization refers to any modification made to a learning algorithm to reduce generalization error without necessarily reducing training error (Goodfellow et al. 2016). A key concept in machine learning is that the generalization error of any model can be decomposed into (a) bias, (b) variance, and (c) irreducible error (Géron 2022). Regularization reduces variance at the cost of increased bias, aiming for improved performance on unseen data.

Tree-Based Methods

Tree-based methods were selected for their ability to model non-linear relationships, robustness to noise (e.g., through bagging (Breiman 1996)), relatively low computational complexity, and interpretability. Ensemble methods based on decision trees leverage the principle that aggregating multiple predictions often yields better performance than any individual prediction (Géron 2022).

- **Decision trees** (Breiman et al. 2017) recursively partition the feature space into regions that are as homogeneous as possible based on optimal splits that minimize impurity (e.g., variance for regression tasks). While individual trees are prone to overfitting, they form the foundation for powerful ensemble techniques.

- **Random forests (RFs)** (Breiman 2001) construct an ensemble of decorrelated decision trees by training each tree on bootstrapped samples of the data and considering a random subset of features at each split. This bootstrap aggregating (bagging) strategy reduces variance and enhances generalization performance.

- **Extremely randomized trees (ETs)** (Geurts et al. 2006) extend RFs by introducing additional randomness: instead of optimizing split points, candidate splits are drawn at random for randomly selected features, and the best among them is selected. This further de-correlates individual trees and can enhance generalization through even lower variance compared to RFs, particularly in noisy datasets. Notably, bagging of training samples for each tree is typically not used in ETs (Geurts et al. 2006).

- **Gradient boosting trees (GBTs)** (Friedman 2001, 2002) belong to the family of boosting algorithms, which iteratively fit weak learners (typically shallow decision trees) to the residuals of previous models, gradually minimizing a differentiable loss function via gradient descent in function space. Each new tree corrects the errors of its predecessor, leading to strong predictive performance. Algorithms making use of GBTs include the

Histogram-based Gradient Boosting Regression Tree (Ke et al. 2017), which significantly accelerates training on large datasets by discretizing continuous features into bins (Alsabti et al. 1998). This algorithm is used in Chapter 4.

2.3.2. Deep Learning Approaches

Deep learning models, particularly neural networks (NNs), have revolutionized many domains due to their ability to learn complex representations from raw data and their scalability. Their success stems from key algorithmic and architectural innovations, as well as advances in hardware that enable the processing of increasingly large datasets.

The backpropagation algorithm (LeCun 1985; Linnainmaa 1970; Parker 1985; Rumelhart et al. 1986; Werbos 2005) lies at the core of training deep networks. It efficiently computes gradients of the loss function with respect to network parameters using the chain rule of calculus, enabling gradient descent (Cauchy et al. 1847) to optimize millions of parameters in high-dimensional spaces.

Modern neural network design incorporates inductive biases, i.e. assumptions about the structure of the data tailored to specific problem domains (Bronstein et al. 2021). For instance, convolutional architectures encode translational symmetry, while graph neural networks exploit permutation symmetry.

Multilayer Perceptrons

Multilayer perceptrons (MLPs), also known as feedforward neural networks, represent the most fundamental class of deep learning models. An MLP learns a function $f^* : \mathcal{X} \to \mathcal{Y}$ by approximating $y = f(\boldsymbol{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ denotes the learnable parameters (weights and biases).

An MLP consists of an input layer, one or more hidden layers, and an output layer, as shown in Figure 2.3. Each layer i applies an affine transformation followed by a non-linear activation function:

$$\boldsymbol{l}_i = \sigma_i(\boldsymbol{W}_i \boldsymbol{l}_{i-1} + \boldsymbol{b}_i)$$

where $\boldsymbol{W}_i$ is the weight matrix, $\boldsymbol{b}_i$ is the bias vector, and σ_i is a non-linear activation such as ReLU (Nair and Hinton 2010) or sigmoid (LeCun et al. 2002). As illustrated in Figure 2.3, this layered composition enables MLPs to approximate highly non-linear functions, although they lack built-in priors for grid-like, sequential, or otherwise structured data.

Convolutional Neural Networks

Convolutional neural networks (CNNs) can be understood as special cases of MLPs that incorporate local connectivity and weight sharing, making them especially effective for grid-structured data such as images. Each neuron in a convolutional layer connects only to a local receptive field (a small patch) of the previous layer's feature map. The same filter (kernel) is slid across the spatial dimensions, computing dot products and generating activation maps. This mechanism enforces translational equivariance: if the input is shifted, the output

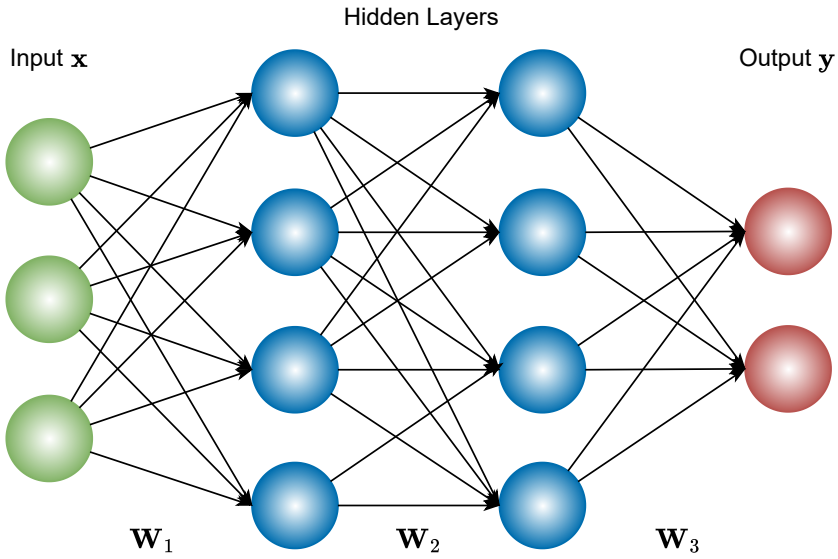

Figure 2.3.: Visualization of an MLP with two hidden layers. The network has three input and two output neurons. Connection weights are represented by the W_i matrices.

shifts accordingly. Formally, a function $f(x)$ is equivariant under a transformation g if $f(g(x)) = g(f(x))$ (Goodfellow et al. 2016). In this case, g would be a translational operator. This inductive bias drastically reduces the number of trainable parameters and enhances generalization when object location is irrelevant to classification.

CNNs (LeCun et al. 1989) typically stack convolutional, pooling, and normalization layers to hierarchically extract features at increasing levels of abstraction; for example, from edges to basic patterns to semantic parts in image recognition tasks.

Residual Neural Networks

Residual neural networks (ResNets) (He et al. 2016) address a critical challenge in training very deep networks: degradation and vanishing or exploding gradients. As networks grow deeper, performance often plateaus or deteriorates due to these optimization difficulties (He et al. 2016).

ResNets introduce skip connections (or shortcuts) that allow the signal to bypass one or more layers. Specifically, each residual block (consisting of arbitrary computations) computes a function $B_i(\boldsymbol{x})$, and the output of the residual layer becomes:

$$\boldsymbol{l}_i = \boldsymbol{x} + B_i(\boldsymbol{x}) \tag{2.11}$$

This formulation enables the network to learn residual functions relative to the identity mapping, thereby facilitating gradient flow during backpropagation. The success of residual learning has transcended image classification, influencing architectures in natural language

processing such as the Generative Pre-trained Transformer (Radford et al. 2018)) and in protein structure prediction (Jumper et al. 2021).

U-Nets

Originally proposed by Ronneberger et al. (2015) for biomedical image segmentation, the U-Net features a symmetric, U-shaped encoder-decoder architecture. The encoder path uses convolutional and pooling layers to downsample the input, capturing contextual information at multiple scales. The decoder path employs transpose convolutions to upsample feature maps, gradually restoring spatial resolution. Crucially, skip connections link corresponding encoder and decoder layers, allowing the network to merge fine-grained spatial details with high-level semantic features. This multiscale processing is essential for pixel-wise prediction tasks, where both local receptive fields and global context are required. More information on U-Nets and a visualization of the specific architecture implemented in Heuer et al. (2024) will be provided in Chapter 4.

While initially designed for segmentation, U-Nets have been adapted to diverse applications, including denoising (Gurrola-Ramos et al. 2021), super-resolution (Lu and Chen 2022), and general inverse problems in scientific computing (Ray et al. 2022). Further architectural details and their adaptations will be discussed in Chapter 4.

Bidirectional Long Short-Term Memory

Long short-term memory (LSTM) networks, introduced by Hochreiter and Schmidhuber (1997), are a type of gated recurrent neural network (RNN) designed to model long-range temporal dependencies. Standard RNNs suffer from vanishing or exploding gradients, which limit their memory span. LSTMs mitigate this issue through a memory cell and three adaptive gates that regulate information flow: input, forget, and output gate. This architecture enables LSTMs to retain relevant information over extended sequences while discarding outdated states, a property known as time warping invariance (Bronstein et al. 2021; Tallec and Ollivier 2018) (or more generally, sequence warping invariance). This property gives LSTMs the ability to process sequences with variable temporal dynamics, such as varying speaking rates in speech recognition.

In my work, I employ bidirectional long short-term memory (BiLSTM) networks, which process sequences in both forward and backward directions using two separate hidden states. The final representation at each time step combines past and future context, making BiLSTMs especially effective for tasks requiring full-sequence understanding. This architecture was instrumental in modeling vertical spatial dependencies in the atmosphere, as detailed in Chapter 4.

2.3.3. Shapley Additive Explanations

SHapley Additive exPlanations (SHAP) (Lundberg and Lee 2017) is a model-agnostic explainable artificial intelligence (XAI) framework that quantifies the contribution of individual input features to a specific prediction made by an ML model. These SHAP values are grounded in cooperative game theory, specifically leveraging the concept of *Shapley* values introduced by Shapley (1951). Given a game with multiple players, *Shapley* values assign each player their marginal contribution to the total outcome, averaged over all possible combinations of player coalitions. In the context of ML, SHAP applies this principle to explain the predictions of an ML model: for a given prediction, the "game" corresponds to the model output, and the "players" are the input features. The explanation model $g(\mathbf{z}')$ used by SHAP is a linear function of binary variables indicating the presence or absence of features (Lundberg and Lee 2017):

$$g(\mathbf{z}') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i', \tag{2.12}$$

where ϕ_0 represents the baseline (i.e., the mean model output over the dataset), ϕ_i denotes the SHAP value for the i-th feature, M is the number of input features, and $\mathbf{z}' \in \{0,1\}^M$ is the binary coalition vector quantifying which features i are included in the coalition ($z_i' = 1$) (Molnar 2025).

The computation of *Shapley* values, and by extension, SHAP values, is the only attribution method that fulfills four fundamental properties of a fair attribution mechanism (Molnar 2025):

- **Efficiency:** The sum of all SHAP values ϕ_i equals the difference between the prediction and the average,
- **Symmetry:** Features contributing equally to every possible prediction receive the same SHAP value,
- **Dummy:** Features that do not alter the prediction in any coalition get assigned a SHAP value of zero, and
- **Additivity:** For a prediction based on the addition of two other predictions $f = f_a + f_b$, the SHAP values for the overall prediction are the sum of the SHAP values for the individual predictions.

These properties make SHAP a theoretically well-founded approach to local interpretability, i.e., explaining individual predictions. Due to its model-agnostic nature, SHAP can be applied to any predictive model. However, for computational efficiency, specialized approximations have been developed for different model classes. In Chapter 4, I employ three such implementations from the SHAP library (Lundberg and Lee 2017): `DeepExplainer` for deep NNs, `TreeExplainer` for tree-based methods, and `KernelExplainer` as a model-agnostic method based on weighted linear regression (Lundberg and Lee 2017).

The results presented in Chapter 4 aggregate SHAP values across individual predictions to estimate the global impact of input features on model outputs, using a method inspired

by Beucler et al. (2024). Further details on the respective implementation are provided in Section 4.2.2.

2.4. Machine Learning-Based Parameterizations

> The hope is that machine learning doesn't replace all of the physically based work. . . but that we find a way to use the power of the two together.
>
> Scientist P6 of Wu and Easterbrook (2025)

In recent years, ML has emerged as a transformative tool for improving the representation of subgrid-scale processes in GCMs. While alternative approaches such as end-to-end learning from reanalysis data (Alet et al. 2025; Bi et al. 2023; Lam et al. 2023), ML-based correction methods (Bretherton et al. 2022; Sanford et al. 2023; Watt-Meyer et al. 2021), or online learning frameworks (Christopoulos et al. 2024; Ouala et al. 2024; Rasp 2020) have shown promise, this section focuses specifically on ML-based parameterizations of atmospheric subgrid processes designed for integration into hybrid GCMs. These hybrid models retain a physics-based dynamical core while replacing or augmenting traditional parameterizations with ML emulators trained to represent unresolved physical processes, such as convection, radiation, turbulence, cloud microphysics, and gravity wave drag, with improved fidelity or efficiency.

The use of NNs in atmospheric modeling dates back to the late 1990s, with early applications focused on accelerating computationally expensive radiative transfer calculations. Cheruy et al. (1995) and Chevallier et al. (1998, 2000) demonstrated that NNs could accurately emulate radiation schemes with significant speed-ups, laying the foundation for future ML applications in climate modeling. Since then, the field has evolved substantially, driven by advances in deep learning, increased availability of high-resolution simulation data, and growing recognition of the limitations of conventional parameterizations. Figure 2.4 shows a simplified schematic of how ML-based parameterizations work alongside conventional schemes in hybrid models. As illustrated, parameterizations, commonly referred to as model "physics", typically operate on 1D atmospheric columns and provide source and sink terms (tendencies) to the dynamical core, which integrates the Navier-Stokes equations (Gettelman and Rood 2016).

To organize the rapidly expanding body of literature, here I group the studies on ML-based parameterizations along a complexity spectrum defined by the source of training and the level of physical realism it represents. The three primary categories are:

1. Emulation of conventional parameterization schemes,
2. Learning from multiscale modeling frameworks (MMFs), and
3. Learning directly from high-resolution simulations.

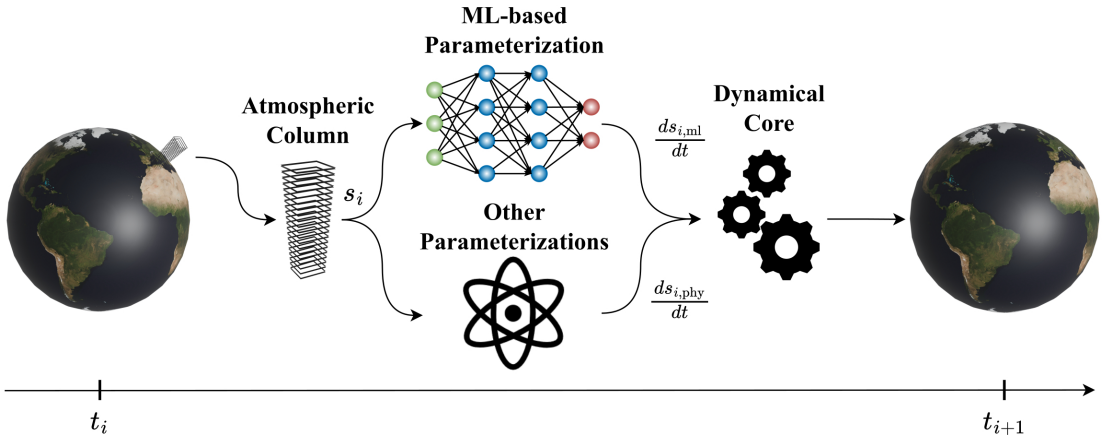

Figure 2.4.: Simplified visualization of one timestep with a coupled ML-based parameterization in a GCM for a representative atmospheric column. The column state at time t_i is called s_i. The tendencies computed by the ML scheme and other conventional parameterization are called $\frac{ds_{i,\text{ml}}}{dt}$ and $\frac{ds_{i,\text{phy}}}{dt}$, respectively. The actual implementation in ICON is more intricate, see Zängl et al. (2015) for details on, e.g., the distinction between *fast-physics* and *slow-physics* and *operator splitting*. Earth background image: NASA Earth Observatory (https://neo.gsfc.nasa.gov/view.php?datasetId=BlueMarbleNG).

These categories reflect a progression from simplified, deterministic representations toward more complex, chaotic, and emergent behaviors derived from realistic or idealized high-fidelity simulations. Category 1 can be seen as distinct to the other two categories, as its primary the goal is computational acceleration rather than improved physical representation of the subgrid processes. More skillful projections may still be achieved, e.g., by enabling larger ensembles or being able to call the emulator more frequently, through the use of a more efficient emulator. In contrast, approaches 2 and 3 aim to improve the physical fidelity of subgrid representations.

For approach 2 and 3, parameterizations can be learned either for all subgrid physics simultaneously or by targeting specific subgrid processes. One advantage of the latter method is that the resulting parameterization can be validated more easily against the known physical principles. Furthermore, by the same logic, physical priors, such as known functional dependencies or conservation laws, can be implemented to constrain an otherwise fully data-driven learning process. This approach has the potential to uncover previously unrecognized aspects of the targeted subgrid physics through the application of XAI techniques. However, this capability critically depends on careful preprocessing and filtering of training data to isolate the specific subgrid process of interest from other subgrid phenomena, which is a nontrivial task in most practical scenarios. In nature, as in high-resolution numerical simulations, subgrid-scale processes in GCMs are not inherently separable but represent emergent phenomena from fundamental physical laws. Consequently, inadequate separation of processes in the training data may lead to double-counting of physical effects, particularly if the processes are not defined as strictly disjoint. Ensuring a clear separation of physical processes during data preprocessing is therefore essential for maintaining physical consistency

and interpretability of the resulting models. Another advantage of the process-specific approach is its modularity, which enables, e.g., retraining an underperforming ML-based scheme without having to retrain others. However, this modular approach may require tuning of the resulting ML models when coupled together, similar to how conventional parameterizations are tuned together. Moreover, by parameterizing individual processes in isolation, interactions between them are neglected, potentially compromising the physical consistency and overall stability of the coupled model. Additionally, the (computational) complexity of the hybrid modeling framework may increase significantly when integrating multiple independent ML parameterizations, compared to employing a unified ML scheme that accounts for coupled processes simultaneously. In summary, neither approach is universally superior. Both joint and process-specific parameterization strategies have their advantages and limitations. As will be demonstrated in Sections 2.4.1–2.4.3, a range of recent studies have successfully employed either method, underscoring the value of both paradigms in advancing ML-based parameterization development.

A fourth possible approach involves learning directly from observational data. While theoretically appealing, this remains largely impractical due to the sparsity, noise, and indirect nature of most atmospheric observations, which complicate the construction of robust training datasets (Eyring et al. 2024b; Rasp 2020). Consequently, most ML parameterizations rely on synthetic or simulated data, where inputs and outputs can be precisely defined and conveniently sampled. One approach that combines simulated and observational data is learning from reanalysis datasets (Rasp 2020), such as the ECMWF Reanalysis v5 (ERA5) (Hersbach et al. 2020). This is particularly useful for end-to-end learning approaches, as each atmospheric state represents the current best estimate of the true atmospheric state. However, when learning individual subgrid processes, one directly inherits the flaws of the model used to generate the reanalysis, as discussed in the following section. Examples of this approach applied specifically for learning the parameterization of gravity wave drag include Amiramjadi et al. (2023), Gupta et al. (2024, 2025), and Matsuoka et al. (2020). A further notable advance in this direction is NeuralGCM (Kochkov et al. 2024), a fully differentiable GCM that can be trained online. It includes a fully connected NN with residual connections to learn subgrid physics within coarse GCM columns (resolutions of 0.7°, 1.4°, or 2.8°). NeuralGCM thus provides a framework for learning subgrid physics from reanalysis datasets. In a follow-up study, Yuval et al. (2024) extend the framework to learn directly from satellite observations of precipitation.

2.4.1. Emulation of Conventional Schemes

A prominent strategy in developing ML-based parameterizations involves training ML models to emulate the input-output mapping of existing physics-based parameterization schemes. This approach, often referred to as *surrogate modeling* or *emulation*, aims to preserve the form of established schemes while reducing their computational cost. Moreover, because ML emulators are fully differentiable, they enable new capabilities such as gradient-based parameter optimization and uncertainty quantification (Mansfield and Sheshadri 2024). Emulators have

been developed for a range of atmospheric processes, including those listed below. This non-exhaustive list specifies the horizontal resolution used to develop each ML-based scheme and indicates whether the emulator was coupled online or only evaluated offline:

- Radiation (Cheruy et al. (1995) and Chevallier et al. (1998) (resolution not specified as inputs are taken from radiosonde observations, not coupled); Chevallier et al. (2000), resolution: 1.875°, coupled; Krasnopolsky et al. (2005), resolution: 3°, coupled; Song et al. (2021), resolution: 5 km, coupled; Ukkonen (2022), resolution: 310 km, not coupled; Hafner et al. (2025a), resolution: 80 km, not coupled; Hafner et al. (2025b), resolution: 80 km, coupled).
- Convection (O'Gorman and Dwyer (2018), resolution: ~330 km, coupled to aquaplanet; Zhong et al. (2024), resolution: ~5 km, coupled; Balogh et al. (2025), resolution: ~55 km, coupled).
- Gravity wave drag (Chantry et al. (2021), resolution: ~25 km, coupled; Espinosa et al. (2022) and Mansfield and Sheshadri (2024), resolution: ~2.8 km, coupled; Hardiman et al. (2023) resolution: $0.556° \times 0.833°$, coupled; Sun et al. (2024), resolution: $0.95° \times 1.25°$, not coupled).
- Turbulence (Wang et al. (2019), resolution: 12 km, coupled).
- Microphysics (Gettelman et al. (2021), resolution: 1°, coupled; Harder et al. (2022), resolution: 150 km, coupled; Perkins et al. (2024), resolution: ~200 km, coupled).

These efforts underscore the utility of emulation as a pragmatic pathway toward more efficient GCMs.

The resulting parameterizations are seamlessly integrable into operational modeling systems, as they replicate the interface and behavior of the original schemes. Furthermore, because the emulated conventional schemes typically have a relatively low complexity compared to, e.g., learning directly from high-resolution data, the resulting ML model can be expected to less likely produce unphysical responses during online inference and tend to yield more stable coupled simulations. Any emulation errors can be traced back by comparing to the original scheme, simplifying debugging and validation.

However, because the goal of this approach is to merely emulate conventional parameterizations, their inherent weaknesses, such as reliance on empirical closures, tunable parameters, and simplifying assumptions (e.g., closure assumptions in convection schemes or bulk approximations in microphysics), are inherited. As a result, ML emulators learn not only the strengths but also the biases and structural errors of their targets schemes. They do not discover new physics and instead learn to reproduce potentially flawed approximations more efficiently.

2.4.2. Learning from Multiscale Modeling Frameworks

An intermediate-complexity approach between emulating conventional schemes and learning directly from storm-resolving global simulations is the emulation of storm-resolving models

(SRMs) embedded within MMFs. These SRMs, typically two-dimensional as explained in Section 2.1.2, explicitly simulate key subgrid processes such as deep convection and cloud dynamics. This setup provides a physically richer training target than conventional parameterizations, as emergent phenomena like mesoscale organization, cold pools, and anvils are naturally represented. The primary objective is to accelerate the computationally expensive SRM components of MMFs. Additionally, because this approach enables learning unknown emergent processes not represented in conventional parameterizations, the trained ML-based schemes can be coupled to other GCMs (see Chapter 5). Several studies have successfully applied ML methods to learn general (or convection-focused, as presented in this thesis) subgrid physics. A non-exhaustive list of these studies is presented below, categorized by horizontal resolution:

- 2° resolution under aquaplanet conditions (Gentine et al. (2018), not coupled; Rasp et al. (2018), coupled; Ott et al. (2020), coupled; Brenowitz et al. (2020), coupled; Beucler et al. (2021), not coupled).
- 2.8° resolution in aquaplanet setups (Behrens et al. (2022), not coupled; Behrens et al. (2025), coupled; Iglesias-Suarez et al. (2024), coupled).
- 1.9° × 2.5° resolution with real geography (Han et al. (2020), coupled to single column model (SCM); Mooers et al. (2021a), only one-way coupled to land model; Wang et al. (2022b), coupled; Han et al. (2023), coupled; Beucler et al. (2024), not coupled and additionally using MMF-based aquaplanet data; Chen et al. (2025), coupled).

  For all of the studies listed until this point, the training data are based on MMF simulations using embedded 2D SRMs with a grid spacing of 4 km.
- 11.5° resolution with real geography (Hu et al. (2025), coupled; Lin et al. (2025), coupled).

  For these studies, the training data are based on MMF simulations using embedded 2D SRMs with a grid spacing of 2 km.
- 1.5° resolution with real geography targeting convective tendencies by subtracting radiative tendencies (Heuer et al. (2025), coupled to different GCM: ICON, presented in this thesis).

  For these studies, the training data are based on MMF simulations using embedded 2D SRMs with a grid spacing of 2 km.

These studies illustrate the practicability of MMF-based emulation and its potential as an intermediate-complexity approach, offering a balance between practical applicability and physical realism.

A key strength of this paradigm compared to learning from high-resolution data is the clear scale separation inherent in MMFs. The host GCM provides large-scale forcing (e.g., temperature and moisture advection, large-scale horizontal winds), while the SRM simulates

subgrid-scale tendencies (e.g., convective heating and moistening). This well-defined input-output structure simplifies the design of ML models, as the predictors (large-scale state) and targets (tendencies) are unambiguously defined.

Despite its promise, this approach has limitations. MMFs themselves are subject to biases due to the idealized nature of the embedded 2D SRMs, which have limited spatial extent (see Section 2.1.2). These simplifications propagate to the ML parameterization, potentially leading to inaccuracies when the parameterization is implemented in a full GCM. Additionally, although MMFs are significantly more computationally efficient than global storm-resolving models (by a factor of approximately 10^3 - 10^4, as discussed in Section 2.1.2), they remain expensive to run for long durations or at high spatial resolution, limiting the volume and diversity of training data. Furthermore, as the 2D SRMs dynamically resolve various subgrid processes simultaneously, isolating individual processes to develop targeted parameterizations for specific atmospheric processes is challenging and requires careful data postprocessing and method design.

2.4.3. Learning from High-Resolution Data

The arguably most ambitious and potentially transformative approach involves training ML parameterizations directly on (coarse-grained) outputs of global or regional high-resolution simulations. This paradigm aims to overcome the limitations of both conventional parameterizations and MMFs by learning from data that most closely approximate real-world atmospheric behavior. The central goal is to develop parameterizations that capture realistic multiscale interactions as they occur in nature, free from the simplifying assumptions of conventional closures or the idealizations of MMFs. By training on data from simulations with explicitly resolved convection and realistic geography, ML models can, in principle, learn complex emergent behaviors. These include mesoscale organization of convection, cold pools, and convective memory, which are poorly represented in current models.

Notable examples demonstrating the feasibility of this approach include the following studies, categorized by targeted process and degree of realism. The horizontal resolution reported for each study is expressed as "`coarse-grained-resolution` & `training-data-resolution`":

- General subgrid physics under aquaplanet conditions (Brenowitz and Bretherton (2018), resolution: 160 km & 4 km, coupled to SCM; Brenowitz and Bretherton (2019), resolution: 160 km & 4 km, coupled; Brenowitz et al. (2020), resolution: 160 km & 4 km, coupled; Yuval and O'Gorman (2020), resolution: mainly 96 km & 12 km (h.p.), coupled; Yuval et al. (2021), resolution: mainly 96 km & 12 km (h.p.), coupled; Wang et al. (2022a), resolution: 192 km & 12 km (h.p.), not coupled; Yuval and O'Gorman (2023), resolution: mainly 160 km & 12 km (h.p.), coupled).

  h.p. denotes that hypohydrostatic rescaling was applied, which increases the horizontal length scale of convection and allows using a coarser horizontal grid spacing to resolve convective motions (Boos et al. 2016; Garner et al. 2007)

- General subgrid physics with realistic geography (Beucler et al. (2024), resolution: 96 km & 12 km (h.p.), not coupled; Watt-Meyer et al. (2024), resolution: 200 km & 3 km, coupled).
- Convection with realistic geography (Krasnopolsky et al. (2013), resolution: 256 km & 1 km, coupled in a regional setting of the tropical Pacific; Heuer et al. (2024), resolution: ~80 km & 2.5 km, coupled, presented in this thesis).
- Cloud cover (Grundner et al. (2022), resolution: ~80 km/160 km & 2.5 km/5 km, not coupled; Grundner et al. (2024), resolution: ~80 km & 2.5 km, not coupled; by Grundner et al. (2025), resolution: ~80 km & 2.5 km, coupled; Morcrette et al. (2025), resolution: $1.25° \times 1.875°$ & 1.5 km, coupled).
- Turbulence by Wang and Tan (2023) (resolution: mainly 2 km & 100 m, not coupled) and by Shamekh and Gentine (2023) (resolution: 1.5 km & 25 m, not coupled).
- Microphysics (and subgrid dynamics) (Sarauer et al. (2024), resolution: 80 km & 5 km, not coupled; Sarauer et al. (2025), resolution: 80 km & 5 km, not coupled).
- Radiation (Hafner et al. (2025c), resolution: 80 km & 5 km, not coupled)

These efforts illustrate the potential of ML-based parameterizations trained on storm-resolving data, showing that such models can learn the complex relationships modeled by SRMs and reduce errors when coupled to GCMs, as illustrated, e.g., by Grundner et al. (2025) for cloud cover.

Storm-resolving simulations represent a broad spectrum of atmospheric processes with minimal parameterization, enabling ML models to learn from a richer and more diverse set of physical states. This may lead to parameterizations that are more accurate across climate regimes and improve long-standing model biases.

However, this approach introduces significant challenges. The most critical is the absence of a clean scale and process separation. In high-resolution simulations, atmospheric processes occur across a continuum of scales, making it difficult to unambiguously define what constitutes "subgrid" physics at a given coarse resolution. This necessitates the use of filtering and coarse-graining operators to define and compute subgrid tendencies. Different filtering (e.g., spatial averaging or spectral filtering) and coarse-graining (e.g., box averaging, subsampling, or spectral truncation) methods can yield different subgrid tendencies, introducing ambiguity in the training data (Brenowitz et al. 2020; Ross et al. 2023). Additionally, the resulting subgrid tendencies may not be fully predictable from the large-scale mean state of a single column, unlike in the other two approaches. Instead, full predictability may necessitate including non-local variables from adjacent columns as predictors (Wang et al. 2022a) or motivate the development of stochastic parameterization frameworks (Christensen et al. 2024). Furthermore, computational cost and large data volume remain major barriers. Due to the $(\Delta x)^{-4}$ scaling of computational demand with respect to grid spacing Δx (see Section 2.1.2), running global storm-resolving simulations for even a few months remains extremely expensive (Giorgetta

et al. 2022; Hohenegger et al. 2023). Publicly available datasets often lack the necessary temporal output frequency, spatial coverage, or completeness of variables required for the comprehensive development of ML-based parameterizations.

3. Data and Methods

This chapter begins with a brief general introduction of the ICON model and the ClimSim dataset, both of which are central to the subsequent analysis. The ICON model is highly relevant for both following chapters, while the ClimSim dataset serves as the training dataset for Chapter 5. Following this introduction, this chapter provides more detailed descriptions of the data and methods split by the results presented in Chapters 4 and 5. Sections 3.4 and 3.5 are already published in Heuer et al. (2024) and Sections 3.5 and 3.6 are already published as a preprint in Heuer et al. (2025).

As indicated in Section 1.2, the author of this thesis created all the content, including text, figures, and tables, that is presented from both publications and implemented the code[1] to reproduce Sections 3.3 and 3.4 and the code[2] to reproduce Sections 3.5 and 3.6.

3.1. The ICOsahedral Nonhydrostatic Model

This thesis presents ML-based parameterizations coupled to the ICOsahedral Nonhydrostatic (ICON) model. In Chapter 4, training data generated by the ICON model, specifically the NARVAL dataset (Klocke et al. 2017; Stevens et al. 2019a) over the tropical Atlantic, are used. Therefore, this section presents a brief introduction to the ICON modeling framework.

The ICON framework is a collaborative development initiative led by the German Weather Service (DWD) and the Max Planck Institute for Meteorology, with contributions from additional partner institutions. It is designed as a unified platform for both numerical weather prediction (NWP) and global climate modeling (Zängl et al. 2015). A core innovation of the model lies in its non-hydrostatic dynamical core, which enables the explicit resolution of vertical accelerations associated with convective updrafts and downdrafts, as detailed in Section 2.1.2. ICON utilizes an icosahedral grid structure (see Figure 3.1), which provides quasi-uniform spatial resolution globally and avoids the grid distortions inherent in traditional longitude-latitude grids. Such distortions would otherwise severely limit the maximum allowable timestep (Heikes et al. 2013; Satoh et al. 2014). The horizontal grids used by the ICON model are denoted as RnBk grids. Starting from an initial icosahedron (R1B0), n represents the number of segments into which each edge is divided, and k denotes the number of subsequent edge bisections (Prill et al. 2022). The number of faces (cells) of the final grid is then: $n_{\text{cells}} = 20n^2 4^k$.

[1] published under https://github.com/EyringMLClimateGroup/heuer23_ml_convection_parameterization (last access: 14.10.2025) and preserved (helgehr 2024)

[2] published under https://github.com/EyringMLClimateGroup/heuer25james_ml_convection_climsim (last access: 14.10.2025) and preserved (helgehr 2025)

The default vertical coordinate in ICON is altitude-based, with terrain-following close to the surface transitioning to smooth, eventually horizontal levels at higher altitudes (Leuenberger et al. 2010).

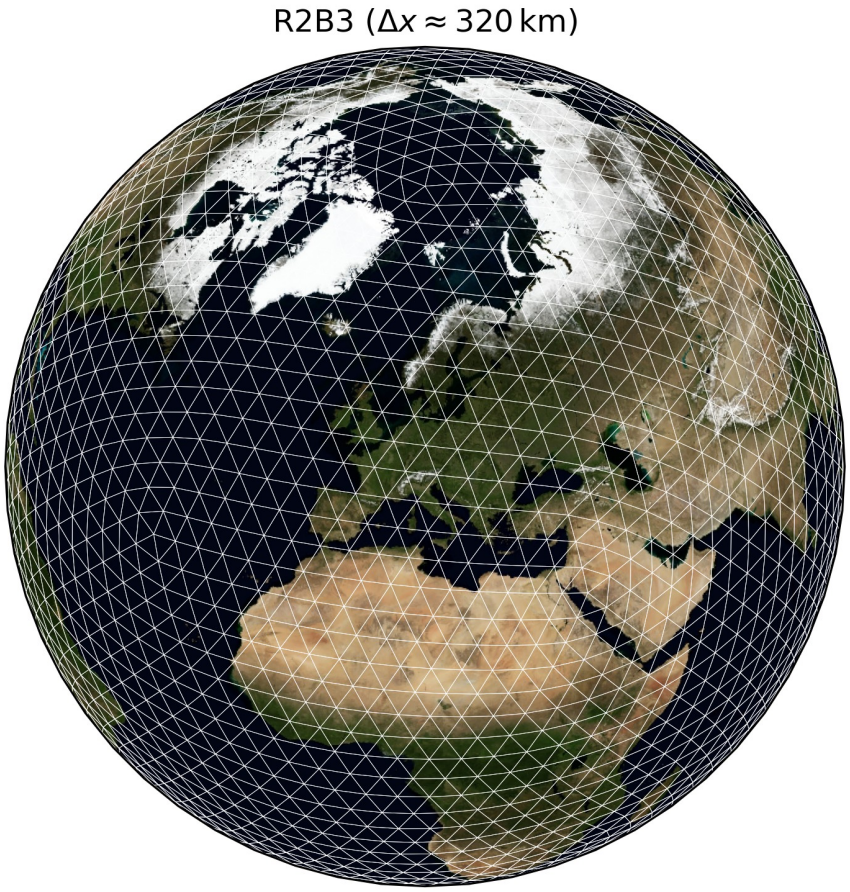

Figure 3.1.: The icosahedral grid structure of the ICON model with an R2B3 grid resolution, translating to a grid spacing of approximately 320 km. The chosen ortographic projection is centered at the location of DLR in Oberpfaffenhofen. Background image: NASA Earth Observatory (https://neo.gsfc.nasa.gov/view.php?datasetId=BlueMarbleNG).

Comparative tests published in Zängl et al. (2015) have demonstrated that ICON achieves higher forecast skill than its operational predecessor at DWD, the hydrostatic Global Model Europe (GME), despite the physics package still being under optimization at the time of initial evaluation. Notably, ICON also exhibits superior computational efficiency, requiring up to four times fewer computational resources than GME for a seven-day forecast. This improvement stems primarily from a more efficient time-stepping algorithm that permits larger timesteps, along with technical improvements targeting the communication between parallel processes (Zängl et al. 2015).

In addition to its NWP applications, ICON has been extended into climate research through the development of ICON-A, the atmospheric component of the ICON Earth System Model (ICON-ESM) (Jungclaus et al. 2022). ICON-A builds upon the physical parameterizations of the ECHAM6 model (Stevens et al. 2013) but benefits from ICON's advanced nonhydrostatic

dynamical core, improved tracer conservation, and high scalability (Giorgetta et al. 2018). Initial setups use a low-resolution configuration to ensure comparability with ECHAM6 in Atmospheric Model Intercomparison Project (AMIP) experiments. The tuning of physical parameterizations in ICON-A required only minor adjustments from the ECHAM6.3 configuration, indicating a high degree of consistency in the underlying physics (Giorgetta et al. 2018). Together, these features position ICON as a next-generation modeling system capable of bridging the gap between high-resolution weather prediction and long-term climate simulation.

The currently developed configuration, ICON-XPP (where XPP stands for eXtended Predictions and Projections) (Müller et al. 2025), is a coupled ESM comprising ICON-NWP (Zängl et al. 2015) for the atmosphere, the ICON ocean and sea-ice model (Korn et al. 2022), and the land component JSBACH (Reick et al. 2021). More information on these and other model components of the model are provided by Müller et al. (2025). Two baseline ESM configurations are available: one with a 160 km atmosphere and 40 km ocean resolution, and another with 80 km atmosphere and 20 km ocean resolution (Müller et al. 2025). ICON-XPP is intended to serve as the foundation for the German contribution to the Coupled Model Intercomparison Project phase 7 (Dunne et al. 2025).

3.2. The ClimSim dataset

The ClimSim dataset represents a comprehensive global dataset spanning ten years, offering multiple resolutions and configurations for the development and evaluation of hybrid climate-ML models. It is generated using the E3SM-MMF. The resolution of the MMF configuration enables for the explicit representation of critical subgrid-scale processes such as deep convection and larger turbulent eddies, which are traditionally parameterized in conventional global climate models and represent a major source of uncertainty in climate projections (Yu et al. 2025). E3SM-MMF uses 2D SRMs with a grid spacing of 2 km, based on the SAM (Khairoutdinov and Randall 2003), to resolve the subgrid dynamics. The high-resolution version of ClimSim used in Chapter 5 features a horizontal grid spacing of approximately $1.5° \times 1.5°$, covering the decade from 2005 to 2014 with outputs every 20 minutes, resulting in a dataset of approximately 41.2 TB (Yu et al. 2025). Simulations were conducted over realistic geography with prescribed sea surface temperatures and sea-ice concentrations, while boundary conditions such as ozone and aerosol levels were held at their 2005-2014 climatological averages, ensuring a stable and representative baseline climate state.

However, certain idealizations are inherent in the setup. As discussed in Section 2.1.2, the model assumes scale separation between the host grid and the embedded SRMs. Additionally, processes such as gravity wave drag and boundary layer mixing are still parameterized by the host model outside the SRMs; subgrid-scale topographic and land-surface heterogeneity are neglected; and atmosphere-ocean coupling and aerosol radiative effects are not included (Yu et al. 2025). Despite these limitations, ClimSim provides an unprecedented opportunity to explore the emulation of high-fidelity climate dynamics. As demonstrated in recent studies

(Hu et al. 2025; Lin et al. 2025) and in this thesis, the dataset supports robust evaluation of machine learning models using different offline metrics, as well as a containerized pipeline for assessing model performance when coupled back to the host model (Yu et al. 2025).

3.3. Relevant Data and Preprocessing for Chapter 4

As training data we use short storm-resolving simulations of the tropical Atlantic that accompanied the NARVAL expeditions performed with ICON (Klocke et al. 2017; Stevens et al. 2019a). Focusing on the deep convective systems of the ITCZ and the explicit representation of convection, this data set serves as an ideal starting point to learn convective subgrid processes. There were two related research campaigns, one from the boreal winter (Dec 2013 / Jan 2014), and one from the boreal summer (Aug 2016). We use simulation data accompanying both expeditions. The horizontal resolution of the used simulations is $\Delta x \approx 2.5\,\text{km}$ (`R2B10` grid), and is available with an hourly output frequency. The simulations were performed with the ICON model (Giorgetta et al. 2018; Zängl et al. 2015), and for each day of the 2-month data set the simulations were initialized at 0000 UTZ and run for 36 h. For this simulation the ICON model was used in its NWP setup without parameterizations for convection and subgrid-scale orography. Parameterizations for radiation, cloud microphysics, and turbulence were active (Klocke et al. 2017). The ICON model solves the fully compressible Navier-Stokes equations with the density ρ as a prognostic variable. ICON uses an icosahedral-triangular C grid and has a non-hydrostatic dynamical core (Zängl et al. 2015).

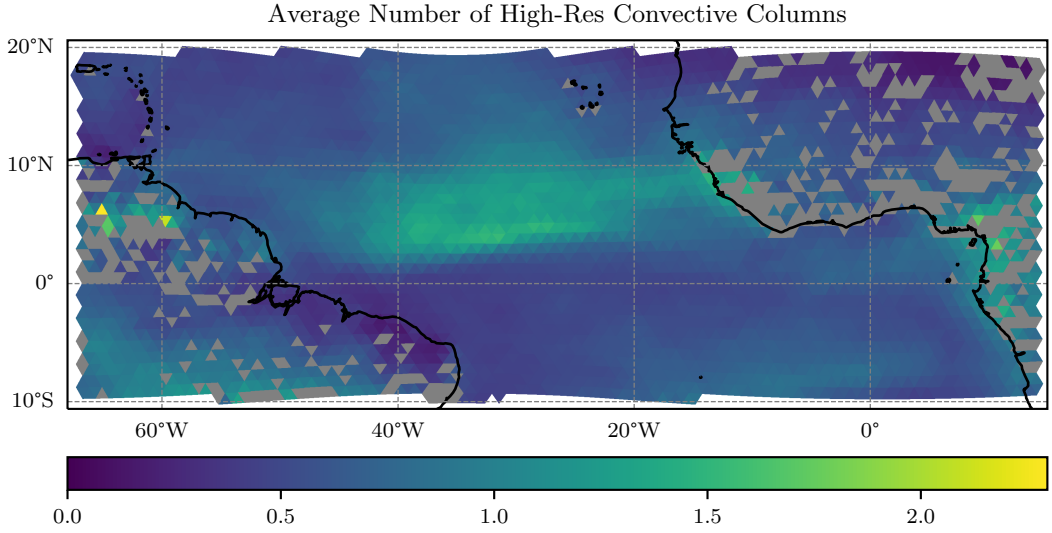

Figure 3.2.: Average number of high-resolution convective cells per displayed low-resolution column and time frame as defined in Equation (3.4) in the studied tropical Atlantic region over the entire considered period of time. In the west the coastline of South America and some Caribbean islands can be seen and in the east the coastline of Africa. The low resolution grid has an approximate horizontal resolution of $\Delta x \approx 80\,\text{km}$. Excluded columns are marked in grey. Adapted with permission from Heuer et al. (2024).

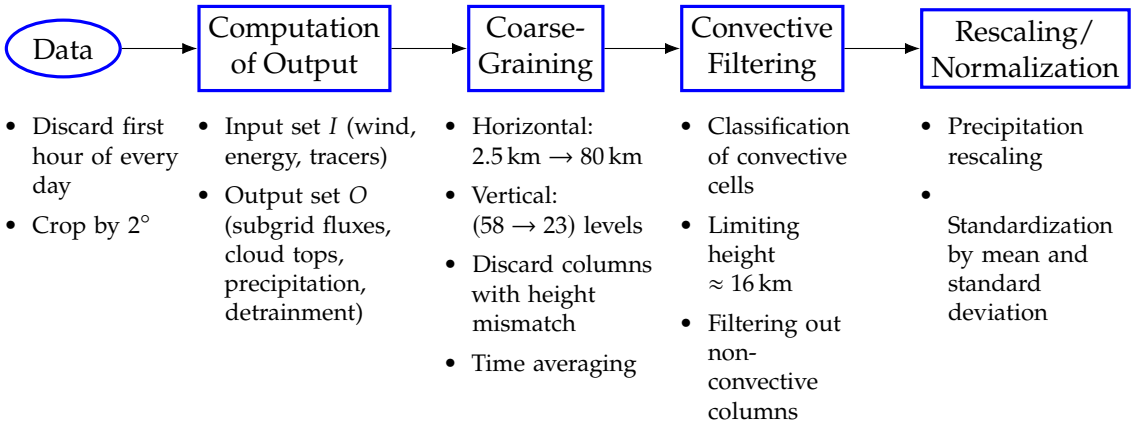

Figure 3.3.: Summary of preprocessing steps. Starting from the original data, first the subgrid fluxes as well as 2D outputs, such as precipitation, were computed. After this, the data was coarse-grained and filtered for active convection. As a final preprocessing step, the data was rescaled and normalized. Adapted with permission from Heuer et al. (2024).

These simulations are well suited for learning a coarse-resolution data-driven convection scheme, as a high number of convective cases are present in the tropical Atlantic region. In Figure 3.2 the spatial distribution of the average number of convective cells per column (as defined below) in the studied region is shown. Columns excluded from the training dataset as described later in the coarse-graining section (Section 3.3.2) are marked in grey. The figure shows a clear pattern of the ITCZ (compare Stevens et al. (2019a, Fig. 2)) with an increased number of convective cells. Additionally, many convective cells can be found along the coast and over mountainous terrain. While many columns over mountainous terrain are filtered out from the data set, there are still many datapoints to learn from over these areas, as seen in Figure 3.2.

As a first preprocessing step we discarded the first hour of every day in the dataset because of some discontinuous behavior at the start of each day related to the initialization/spin-up phase of the simulations. Additionally, we also cropped the original NARVAL region by 2° on all sides since we noticed some boundary effects in the spatial patterns as well. The region seen in Figure 3.2 was already cropped by the mentioned 2°.

To give a short overview of the preprocessing steps described below, Figure 3.3 depicts an overview of the various steps used, beginning with the original data set.

3.3.1. Computation of Output

The selection of input and output variables for the ML models are based on the implementation of the cumulus scheme in the ECHAM6 model (Nordeng 1994; Stevens et al. 2013; Tiedtke 1989). They correspond to the physical quantities transported by convective processes and a few related quantities such as precipitation. If not stated differently, we used the following set of variables for the input of the convective scheme

$$I = \{u, v, w, h, q_v, q_l, q_r, q_i, q_s\}.$$

This set consists of the zonal, meridional, and vertical wind components (u, v, w), as well as the liquid/ice water static energy (h) and five different tracer species. These tracer species are the specific humidity (q_v) and specific cloud water, cloud ice, rain, and snow content (q_l, q_i, q_r, q_s). The liquid/ice water static energy is defined here as

$$h = c_p T + zg - L_v \cdot (q_c + q_r) - L_s \cdot (q_i + q_s + q_g), \tag{3.1}$$

with temperature T, altitude z, the specific heat at constant pressure c_p, specific graupel content q_g, and the latent heat of evaporation and sublimation L_v and L_s. We chose to not give the ML models any information about their spatial location or solar insolation in order to force them to learn from the dynamical state. This also enables the application of the trained models outside of their limited training domain.

Correspondingly, the output fields are

$$O = \{F_u^{\text{sg}}, F_v^{\text{sg}}, F_h^{\text{sg}}, F_{q_v}^{\text{sg}}, F_{q_l}^{\text{sg}}, F_{q_r}^{\text{sg}}, F_{q_i}^{\text{sg}}, F_{q_s}^{\text{sg}}, z_{cltop}, p_{cltop}, q_{l,detr}, q_{i,\text{detr}}, P\}.$$

The first eight variables with notation "F_{var}^{sg}" are 3D fields and correspond to the subgrid flux component of the input variables I (excluding w). The remaining variables in the output set are 2D fields, namely cloud top height (z_{cltop}), cloud top pressure (p_{cltop}), integrated liquid/ice detrainment ($q_{l,\text{detr}}$, $q_{i,\text{detr}}$), and precipitation (P). For the cloud top level we chose to predict the altitude as well as the pressure, although they contain very similar information, because our goal was to provide the same output as the ECHAM6 cumulus scheme.

We focused on predicting subgrid fluxes instead of the direct tendencies because this allowed abiding conservation laws by applying appropriate boundary conditions (no-flux at the top and a flux which is consistent with the surface forcing at bottom). We decomposed variables such as the density (ρ) into a horizontal spatial average (on the same model level) over the coarse resolution, denoted by an overline, and a fluctuating component, denoted by a prime, as $\rho = \overline{\rho} + \rho'$. The fluctuating component therefore represents the departure from the coarse grid average. This enabled us to calculate the subgrid (i.e., unresolved) vertical advective flux of, say, the variable u, F_u^{sg}, for a given coarse resolution as follows:

$$F_u^{\text{sg}} = \overline{\rho w u} - \overline{\rho}\,\overline{w}\,\overline{u} = \overline{\rho}\,\overline{w'u'} + \overline{w}\,\overline{\rho'u'} + \overline{u}\,\overline{\rho'w'} + \overline{\rho'w'u'}. \tag{3.2}$$

This subgrid momentum flux F_u^{sg} was calculated as the difference between the coarse-grained flux $\overline{\rho w u}$ obtained by first calculating the flux with the high-resolution resolved variables, then coarse-graining it to the coarser resolution, and the flux calculated with the low-resolution variables $\overline{\rho}\,\overline{w}\,\overline{u}$ (see Equation (3.2)). The term on the right hand side in Equation (3.2) results from the fact that averages over fluctuations are by definition zero. This method is similar to the one of Yuval et al. (2021), but without neglecting the horizontal density fluctuations between high-resolution cells within a coarse resolution target cell of the coarse-graining procedure. This is especially important for models with terrain-following vertical coordinates, such as the height based terrain following vertical coordinate of the ICON model (Giorgetta et al. 2018), because horizontally neighbouring cells (same vertical level) in the lower troposphere over

land with steep topography can have strongly different height, thus different pressure and density. By looking into the subgrid variations of ρ we found that, especially in the lowest levels over heterogeneous terrain, there are fluctuations of up to 25 % of the mean value within a single coarse grid cell. As we are calculating the subgrid flux from a single snapshot of the dynamics and do not consider differences between timesteps, the subgrid flux represents the flux difference between the coarsened high-resolution state and coarse state due to resolved processes. Here, these resolved processes are cumulus convection and gravity waves since we only learn from convective columns (method is shown later in this section). Gravity wave drag mainly impacts higher levels (Kim et al. 2003) and the here developed parameterizations are limited in height (see Figure 3.4). The momentum flux due to gravity waves, excited by convection, is a second order effect which we neglect here.

For the cloud top height/cloud top pressure (z_{cltop}/p_{cltop}) we took the height/pressure of the highest cell with convective clouds found according to the condition formulated in the next section (Equation (3.4)). While there are different ways to estimate the detrainment of liquid/ice (Arakawa and Schubert 1974; Zhang and McFarlane 2019) we decided to follow Baba and Giorgetta (2020) and Nordeng (1994) and calculated the fractional detrainment as

$$\delta = -\frac{1}{\sigma}\frac{\partial \sigma}{\partial z}, \tag{3.3}$$

where z is the altitude and σ the fractional cloud area. As such, it was possible to calculate the integrated detrainment of water and ice by multiplication with the vertical mass flux and integrating along the column. Before integration, the column was masked according to its temperature (above or below 0 °C) (Stevens et al. 2013) to differentiate between liquid and ice detrainment. For precipitation we cannot assume that it stems entirely from convective precipitation in convective columns as stratiform and convective precipitation often occur simultaneously (Houze 1997; Schumacher and Funk 2023). Therefore, when coupling the ML parameterization to the ICON model we will set the large-scale precipitation from the model to zero in regions where the ML parameterization is active. Another approach would be to classify the precipitation in the high-resolution data as convective or not, based on thresholds on e.g. vertical velocity, precipitation rate or based on the spatial structure of precipitation clusters. Here, we decided to predict both precipitation types together as the before mentioned approaches would introduce additional degrees of freedom into the method and therefore complexity.

3.3.2. Coarse-Graining

The coarse-graining was done first in the horizontal and afterwards in the vertical direction as described in Grundner et al. (2022) for a data-driven cloud cover parameterization. The horizontal coarse-graining from the `R2B10` ($\Delta x \approx 2.5$ km) to an `R2B5` ($\Delta x \approx 80$ km) grid was performed with the help of the `remapcon` function from the Climate Data Operators (Schulzweida 2022). At this scale individual convective clouds and smaller convective systems are coarse-grained, allowing us to parameterize their average impact on the large-scale dynamics. In the

vertical, we reduced the resolution from 58 to 23 levels up to the mentioned limiting height of ~ 15.9 km in Figure 3.4. The vertical coarse-graining operator works in a similar way as the horizontal averaging. The high-resolution cells were averaged weighted by their fractional proportion in the coarse cell (Grundner et al. 2022). Some low-resolution columns have a significantly lower base than the high-resolution cells because of the more detailed topography in the high-resolution data. Therefore, it was not possible to compute reasonable averages with the above described coarse-graining operator in the lowest model levels. Here, we also adopted the method from Grundner et al. (2022) and excluded columns with a significant difference between the vertical extent of low and high-resolution columns of the dataset.

In the high-resolution data, cells on the same vertical level can be on different geometric heights due to the terrain-following coordinate system. An approximation applied here is that the coarse-graining is first performed in the horizontal and afterwards in the vertical. Therefore the result can be different from coarse-graining over the low-resolution volume (Grundner et al. 2022).

Additionally, we introduce time-averaging to reduce the noise from instantaneous snapshots of the dynamics as it was found to reduce model overfitting in Ramadhan et al. (2020). For a column in the data set at time t_i, we average the column variables and fluxes over the time steps t_{i-1}, t_i, t_{i+1}, corresponding to a moving window of a three-hour duration. Physically, the three-hour temporal averaging should still allow to resolve the life cycle of the tropical deep convective clouds with a diurnal cycle (Chen and Houze Jr 1997). A 3 h window is just about short enough to resolve the life cycle of such clouds and still allow a minimal smoothing of higher frequency variability.

3.3.3. Filtering for Convection

In order to learn mainly from columns in which convection has a dominant impact on the overall dynamics we introduced a filtering of the data. First, individual high-resolution cells were classified as convective if the following conditions (Kirshbaum 2022; Romps and Charn 2015) are met:

$$q_l + q_i > 0.01\,\mathrm{g\,kg^{-1}}, \quad w > 0, \quad B \propto \theta_v - \overline{\theta_v} > 0, \tag{3.4}$$

where w is the vertical velocity, θ_v is the virtual potential temperature, and q_l/q_i are the specific cloud liquid water/cloud ice content, respectively. Additionally, the buoyancy B has been introduced in the conditions (3.4). In this case the overline denotes horizontal averaging over approximately 10 km. We chose this averaging scale as convection becomes partly resolved by grid scale dynamics for resolutions higher than approximately 10 km (Ahn and Kang 2018; Arakawa et al. 2011). The averaging was performed with the `remapcon` function (Schulzweida 2022) to an R2B8 resolution. Next, we classified entire low resolution columns as convective or non-convective. For this, the number of convective cells per high-resolution column was summed up along the height dimension and coarse-grained horizontally (as explained above). If the so-calculated 2D field was equal (or higher than) 1 for a given column, so that on average

all high-resolution columns inside the coarse column had at least one convectively classified cell, this coarse column was classified as convective and was added to the training data set. These columns are henceforth referred to as "convective" columns. A time average over the entire observed period of this so computed low resolution data is displayed in Figure 3.2. Furthermore, we added 10 % of the non-convective columns for training so that we ended up with slightly more than 2 million coarse sample columns. Before the filtering, there were about 5 million low-resolution and approximately 455 million high-resolution columns in the whole data set.

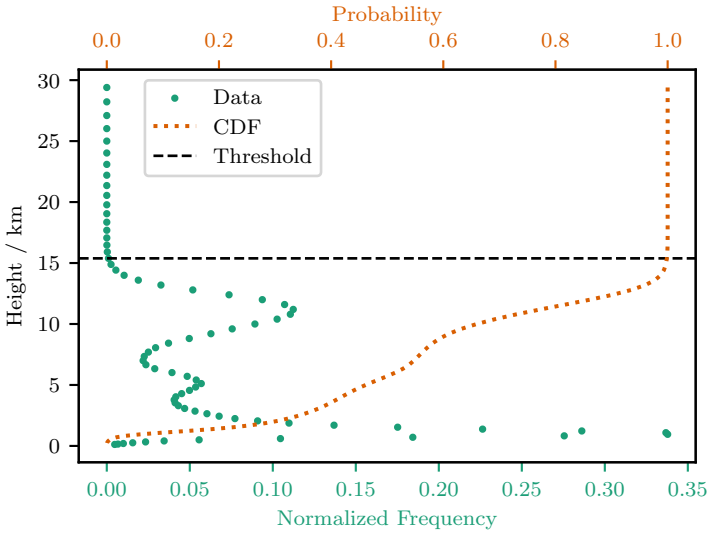

Figure 3.4.: Probability distribution of convectively classified cells over altitude (green large dots) in the high-resolution data with ICON over the NARVAL region. The orange dashed line shows the cumulative distribution function (CDF) and the black dashed line represents the height up to which 99.9 % of the convective cells are found. The bottom scale corresponds to the probability and the top to the cumulative distribution. Adapted with permission from Heuer et al. (2024).

In order to find a limit in altitude to predict unresolved convective effects, we considered that convection in the atmosphere under normal conditions is limited by the tropopause (Shenk 1974). Therefore, we checked up to which height we find convectively classified cells in the data set. The result can be seen in Figure 3.4. The limiting height in the figure is drawn at the height up to which 99.9 % of the convectively classified cells are found (compare dashed orange line). This height is at ~ 15.9 km, which is reasonable considering the tropical tropopause height of roughly 12 km to 17 km (Gettelman et al. 2002). Only values below this height are considered as input and output to the machine learning algorithms.

The general form of the data observed in Figure 3.4 resembles the expected trimodal distribution of convective clouds in the tropics (Johnson et al. 1999). The lowest peak corresponds to (shallow) cumulus, the peak at ~ 5 km to cumulus congestus and the highest clouds found are deep cumulonimbus clouds.

3.3.4. Rescaling and Normalization

For higher numerical stability of the machine learning models and to have the variables on the same scale, we standardize the 2D fields by subtracting the mean across samples from all 2D variables and dividing by the standard deviation. The same procedure is done for all 3D variables, but in this case mean and standard deviation are calculated across the height dimension as well. We also tested normalizing variables by their mean and standard deviations level by level but observed a decrease in model skill.

Furthermore, before applying the standardization, we use the following nonlinear rescaling for the accumulated precipitation P per hour:

$$P' = \ln\left(1 + \frac{P}{1\,\mathrm{kg\,m^{-2}\,h^{-1}}}\right). \tag{3.5}$$

The reason for this is that precipitation intensities are typically represented by a heavily skewed (gamma) distribution (Martinez-Villalobos and Neelin 2019). This distribution is characterized by a comparatively large number of low values and very few heavy precipitation events. Without a proper rescaling, ML models would achieve a low prediction error by predicting zero precipitation regardless of the input (Rasp and Thuerey 2021). Additionally, it is well known that coarse GCMs have a bias towards low intensity precipitation events (Moseley et al. 2016; Rasp et al. 2018). The rescaling should help mitigate some of this problem.

3.4. Relevant Methods for Chapter 4

ML-based convection parameterizations have been developed using different kinds of methods. These include RFs (Limon and Jablonowski 2023; O'Gorman and Dwyer 2018; Yuval and O'Gorman 2020), MLPs (Gentine et al. 2018; Iglesias-Suarez et al. 2024; Rasp et al. 2018; Yuval et al. 2021), ensembles of MLPs (Krasnopolsky et al. 2013), Residual Convolutional Neural Networks (Han et al. 2020, 2023), ResNets (Wang et al. 2022b), generative adversarial networks (Nadiga et al. 2022), and variational auto encoders / variational encoder decoders (Behrens et al. 2022; Mooers et al. 2021b). One goal of this study is to evaluate various kinds of machine learning models on the same data set. Therefore, we first introduce the used models. All models use a vertical column (23 height levels and nine variables) from the sample data set as input and the column fluxes (23 height levels and eight variables) plus five 2D variables as output, see above.

We tested four different deep learning architectures: MLP, CNN, RNN (He et al. 2016), and a convolutional neural network with a U-shaped architecture (U-Net) (Ronneberger et al. 2015). The MLP family consists of several fully connected layers with additional optional batch normalization layers and activation functions (see section A.2). Furthermore, we introduced a linear model (LinMLP) which is based on the best found architecture of the MLP class but all nonlinear activation functions are replaced by linear ones. For the CNN class we decided to consider networks with a first convolutional layer connected to some number of fully connected

layers thereafter. All convolutions are 1D convolutions in the vertical as the data set consists of variables on different levels due to the typical neglect of horizontal interactions and variability for parameterized processes in climate models. The ResNet architecture is inspired by Wang et al. (2022b), the network consists of several different blocks with some number of fully connected layers and optional batch normalization. The input of each block is added to its output to form the final output set. This helps prevent vanishing gradients and degradation (He et al. 2016). For the gradient-based optimization of the networks we chose to use the Adam algorithm (Kingma and Ba 2014). For the implementation of all deep learning models we relied on the Pytorch library (Paszke et al. 2019).

Furthermore, we decided to use a U-Net architecture, see Figure 3.5. This network is similar to the ResNet in the sense that it contains residual connections and that it is constructed out of structurally similar blocks. In contrast to the ResNet, these blocks use two convolutional layers each instead of an arbitrary number of fully connected layers. Additionally, this architecture utilizes max pooling and transpose convolution layers to compress and expand the input in the height dimension. This allows the network to process the input information on multiple spatial scales. During the compression process (left part of Figure 3.5) the channel dimension (width in the figure) grows. The kernel size of the convolutions stays constant but the height dimension shrinks, this effectively increases the receptive field for each consecutive layer in the network. The U-Net is therefore able to detect patterns on scales between the models vertical level spacing (~30 m at the lowest level or up to ~500 m for the highest predicted level) and the column height (~16 km). In the expansion process (right part) the channel dimension shrinks again. We propose this architecture, which is particularly suited for multiscale modeling, for the given parameterization problem because of the multiscale nature of moist convection (Majda 2007). The U-Net has favorable properties for our problem as local features can be picked up by the network on a variety of different scales throughout the downscaling process, and the residual connections help to communicate this information to the upscale branch of the network. This capability is crucial for tasks that require understanding both local and global context within the input data, such as in image segmentation (Ronneberger et al. 2015) where the target output can depend on patterns of varying sizes and resolutions. In the context of convection, the initial layers are capable of capturing more small-scale convective systems/flows and the more compressed layers are responsible for representing deep convection/large-scale systems.

Besides these deep learning architectures, we trained five different non-deep learning models. For the implementation of these we used Scikit-Learn (Pedregosa et al. 2011). As lowest complexity models we used linear methods such as Lasso (Tibshirani 2018) and Ridge (Hoerl and Kennard 1970) regression. Additionally, we used three tree-based models. These include RF (Breiman 2001), ET (Geurts et al. 2006), and GBT (Friedman 2002). Further information about the different ML models can be found in section A.1.

To select an appropriate set of hyperparameters we chose to split the data non-consecutively into a training/validation/test set with a fraction of 80 %/10 %/10 % of the data. This corresponds to $\sim 1.6 \cdot 10^6$ sample columns for training and $\sim 2 \cdot 10^5$ columns for validation/testing. Depending on the architecture we treated the different input variables as separate channels (for

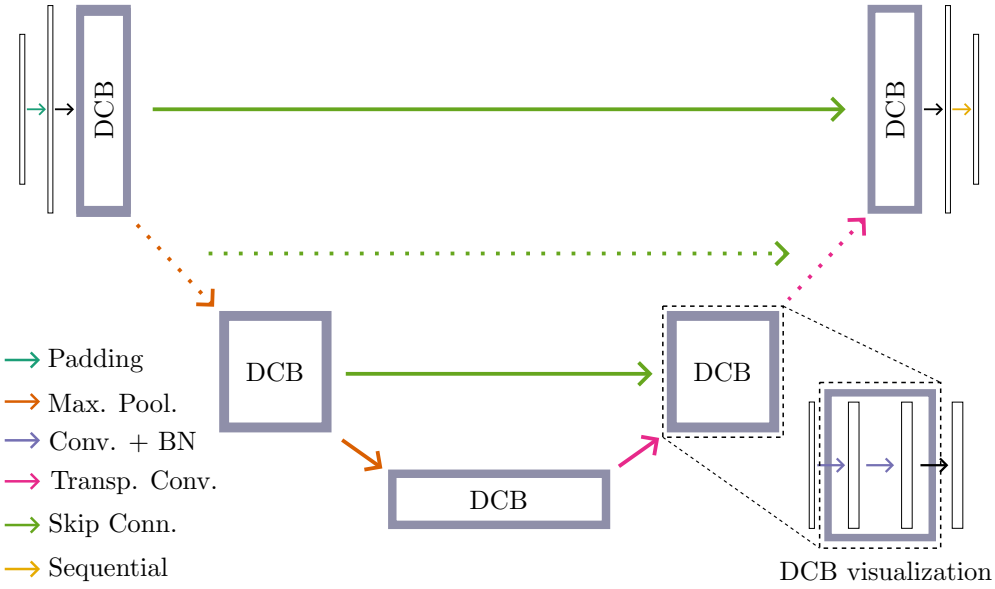

Figure 3.5.: Visualization of the used U-Net architecture. The abbreviations DCB, Conv., Transp. Conv., and BN stand for double convolution block, convolutional layer, transpose convolutional layer, and batch normalization layer, respectively. The dotted lines mark the possibility for more blocks depending on the result of the hyperparameter optimization (HPO). The horizontal lines indicate skip connections. In the lower right of the figure, a more detailed visualization of the double convolution block is given. Adapted with permission from Heuer et al. (2024).

CNN and U-Net) and otherwise concatenated them in one vector. The output variables were always concatenated in one vector. For the non-deep learning algorithms we first did the HPO on a subset of the data from five random days ($\sim 1.6 \cdot 10^5$ samples) because most of the models have difficulties with handling vast amount of data. The models identified as best in the HPO where then trained on the whole data set. An explanation of the different hyperparameters involved in all models can be found in section A.2.

3.5. Relevant Data for Chapter 5

3.5.1. ClimSim and Cross-Validation Procedure

We used the "high-resolution version" of the ClimSim dataset (LEAP 2023; Yu et al. 2023) with a horizontal resolution of approximately $1.5° \times 1.5°$. The data are produced over realistic geography with E3SM-MMF (E3SM Project 2018), span 2005–2014 with 20 min output, and total about 41.2 TB (Yu et al. 2025). Sea surface temperatures and sea-ice amount were prescribed. Boundary conditions such as ozone and aerosol concentrations were set to the climatological average of 2005–2014 (Yu et al. 2025). In this multiscale modeling framework, subgrid-scale dynamics are resolved by 2D SRMs embedded within each grid column of the coarse atmospheric model. These SRMs have a horizontal resolution of 2 km and are two-way coupled to the coarse atmospheric model (Hannah et al. 2022). The SRMs replace the coarse model's parameterizations for convection and boundary-layer turbulence (Lee et al.

2023) and are used for the calculation of radiative fluxes. The SRMs are mostly based on the System for Atmospheric Modeling (SAM; Khairoutdinov and Randall (2003)), use SAM's single-moment microphysics, and close sub-SRM-grid-scale turbulent fluxes with a diagnostic Smagorinsky-type closure. Gravity wave drag and vertical diffusion are parameterized by the coarse atmospheric model outside the SRMs (Yu et al. 2025). We refer the curious reader to Yu et al. (2023) for more details on ClimSim and Hannah et al. (2022) for the E3SM-MMF setup.

This dataset offers several advantages compared to training data from other high-resolution models that enhance its utility for research. Notably, it features a well-defined scale separation between subgrid-scale and grid-scale dynamics, as it is generated through a superparameterized modeling framework. Additionally, the dataset is readily accessible to the research community and was utilized in a Kaggle competition (Lin et al. 2024) that attracted over 690 finalized submissions. The collaborative efforts of participants in this competition have yielded highly competitive machine learning models and baselines, providing a valuable benchmark for future studies and inspiring innovative approaches to data-driven modeling.

Potential drawbacks of learning from the superparameterized ClimSim data set are the usage of 2D SRMs with limited extent for the embedded subgrid dynamics and the useful but artificial scale separation. Therefore, the subgrid dynamics are highly idealized and can, e.g., influence the mean state response affecting moisture and associated shortwave cloud effects (Pritchard et al. 2014). Additionally, as shown later in Section 5.2.1 and Section 5.2.4, the zonal precipitation distribution of the high-res version of ClimSim shows too high mean precipitation with respect to the Global Precipitation Climatology Project (GPCP), especially in the mid to high latitudes as well as for the ITCZ.

To train ML models efficiently while utilizing the temporal variability of the data we only used the first two days of every month over the span of the ten years with a timestep of 20 min. This resulted in approximately $217 \times 10^6/37 \times 10^6/37 \times 10^6$ training/validation/test samples. For more efficient training of the NNs we further subsampled the data, ending up with a $25 \times 10^6/5 \times 10^6/5 \times 10^6$ training/validation/test split.

3.5.2. "ClimSim Convection": Approximate Removal of Radiation for Training

While ClimSim facilitates process separation, it does not cleanly isolate convective processes. To use ClimSim to train a drop-in replacement for ICON's convection scheme, we must avoid double-counting radiation. We therefore constructed ClimSim Convection by subtracting radiative temperature tendencies from ClimSim. Because the E3SM-MMF subgrid state is unavailable, we approximated the radiative contribution by recomputing column radiation offline with the RTE+RRTMGP scheme (Pincus et al. 2019, 2023), the scheme used in our ICON setup, and subtracting the resulting radiative heating from the superparameterized temperature tendencies (see Figure 3.6).

RTE+RRTMGP is driven by per-column inputs from ClimSim: temperature; tracers for specific humidity, cloud liquid, and cloud ice; solar insolation and the solar zenith angle's cosine; ozone, N_2O, and CH_4; shortwave/longwave albedos; surface pressure; and outgoing

longwave radiation. Mid- and half-level pressures are reconstructed from the time-independent coefficients `hyam`/`hybm` and `hyai`/`hybi` provided by ClimSim:

$$P_{m,k} = \texttt{hyam}_k P_0 + \texttt{hybm}_k P_{\text{sfc}}, \qquad P_{h,k} = \texttt{hyai}_k P_0 + \texttt{hybi}_k P_{\text{sfc}}, \tag{3.6}$$

with $P_0 = 1000\,\text{hPa}$, and k representing the height level index. Cloud effective radii are computed as in ICON.

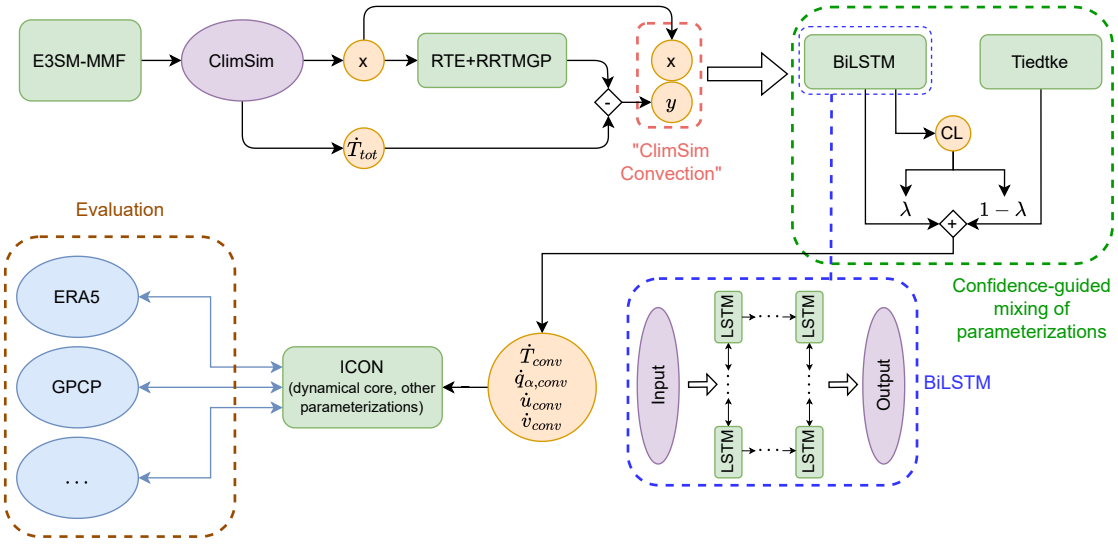

Figure 3.6.: Overall training and evaluation pipeline of our hybrid model. x and y represent inputs and outputs of the ClimSim dataset, based on the E3SM-MMF model. $\dot{T}_{tot}$ is the total temperature tendency, and RTE+RRTMGP the ICON radiation scheme. The ClimSim dataset is first modified to separate radiative and convective subgrid tendencies, forming a new dataset, "ClimSim Convection". Afterward, we trained a BiLSTM model including a confidence loss (CL). Using CL, this model is mixed with the conventional "Tiedtke" cumulus convection scheme to predict convective tendencies as well as precipitation. In the mixing process, λ represents the fraction provided by the BiLSTM and $1-\lambda$ is the fraction from the conventional "Tiedtke" scheme, respectively. This mixed scheme predicts the tendencies due to convection in temperature $\dot{T}_{conv}$, water vapor, cloud liquid water, cloud ice ($\dot{q}_{\alpha,conv}$, $\alpha = v, l, i$), zonal wind $\dot{u}_{conv}$, and meridional wind $\dot{v}_{conv}$. Finally, we coupled the mixed scheme with the ICON model and evaluate these runs' emergent statistics with respect to observational datasets, including ERA5 and GPCP. Adapted with permission from Heuer et al. (2025).

This subtraction yields tendencies dominated by convective heating, which we aim to learn, with residual contributions from microphysics and turbulence. Explicitly separating convection from microphysics/turbulence would be ad hoc and arguably unphysical (Arakawa and Jung 2011; Arakawa 2004; Randall et al. 2003). Accordingly, in coupled runs we replaced only deep convection in ICON and keep its native vertical diffusion scheme active; once radiation was removed, we found no evidence of residual double-counting (e.g., anomalous diffusion signatures; not shown). Furthermore, we set up ICON simulations without vertical diffusion and/or without microphysics schemes. These simulations diverged almost immediately.

Figure B.1 in Appendix B.3 shows that removing radiative tendencies preserves the distributional shape across the column and yields a net convective heating (left) that balances the removed longwave cooling (middle), with shortwave heating as expected (right), consistent with the atmospheric energy budget. Overall, ClimSim Convection keeps assumptions minimal while acknowledging ClimSim's imperfections when training convective parameterizations. Learning from ClimSim Convection is therefore treated as a transfer-learning exercise that requires online validation.

3.5.3. Datasets Used for Evaluation

For the evaluation of the coupled ICON online runs, we mainly employed two datasets: GPCP (Adler et al. 2018) and the ERA5 reanalysis (Hersbach et al. 2020). The GPCP dataset provides a comprehensive, long-term record of global precipitation, combining various satellite observations, rain gauge measurements, and other remote sensing data. GPCP offers a spatial resolution of 2.5° × 2.5° and temporal coverage spanning several decades with monthly temporal resolution, making it ideal for validating simulated precipitation patterns against observational benchmarks. ERA5, on the other hand, is the fifth-generation ECMWF reanalysis dataset, which provides atmospheric data at a higher resolution than GPCP (about 0.25° × 0.25°) at hourly intervals. It incorporates a wide range of variables, including temperature, wind, humidity, and surface pressure, and is widely used for evaluating climate models due to its high accuracy and consistency with physical laws. These datasets were chosen for their broad applicability, high quality, and availability, enabling a direct and meaningful evaluation of the model's performance in real-world scenarios.

For the bulk of the evaluation, we used the Earth System Model Evaluation Tool (ESMValTool) (Andela et al. 2025; Righi et al. 2020). ESMValTool is a community diagnostic and performance metrics tool for the evaluation of ESMs (Righi et al. 2020). Besides the ERA5 and GPCP references, ESMValTool offers the possibility to evaluate against a multi-observational mean for certain variables. These datasets additionally include, e.g., MERRA2 (Gelaro et al. 2017), ESACCI-WATERVAPOUR (Schröder et al. 2023), and ISCCP-FH (Zhang and Rossow 2023). We used the multi-observational mean to evaluate the spatial distribution of column integrated water vapor. For evaluating precipitation statistics, we utilized the GPCP dataset, while ERA5 is used for the near-surface temperature T_{2m}.

3.6. Relevant Methods for Chapter 5

This section describes how the conventional Tiedtke scheme compares to our newly developed ML-based scheme. After introducing the Tiedtke scheme, we outline the methodology behind training the ML model, including constructing its loss function, selecting hyperparameters, and implementing confidence-guided mixing in ICON.

As seen in Figure 3.6, we used the ClimSim Convection data to train NNs (with a physics informed loss) to predict the convective tendencies and convective precipitation with the

atmospheric state variables as input. The NNs are based on a BiLSTM architecture and trained with a CL inspired by the first place entry "greySnow" to the ClimSim Kaggle competition (Lin et al. 2024). This enables the networks to judge their own prediction error during inference. We leveraged these error predictions for a mixed convection parameterization in which the NNs' predictions are mixed with those from the conventional cumulus convection scheme when the NNs exhibit low confidence, as explained in Section 3.6.4.

3.6.1. Tiedtke Convection Scheme

As described in Giorgetta et al. (2018) and Möbis and Stevens (2012), the conventional cumulus convection scheme used in the ICON model is based on a mass flux formulation by Tiedtke (1989) with modifications by Nordeng (1994). It differentiates between shallow, mid-level, and deep convection. Deep convection occurs in disturbed environments with synoptic scale convergence whereas undisturbed environments allow for shallow convection (Tiedtke 1989). Mid-level convection originates at levels above the boundary layer and is often formed by lifting of low level air until saturation (Blanchard et al. 2021; Tiedtke 1989). For deep convection, an adjustment-type closure based on the Convective Available Potential Energy is used. Shallow convection uses a moisture convergence closure and a large scale vertical momentum closure which determines the cloud base mass-flux for mid-level convection. The scheme represents all subgrid convective cloud processes by one updraft and one downdraft, respectively.

The bulk convection scheme works by defining a vertical profile for the mass-flux $M(z)$ which varies by the amount of entrainment and detrainment happening in the up-/downdrafts (for downdrafts only turbulent entrainment/detrainment is considered (Nordeng 1994)). To determine the magnitude of the mass-flux and relate the subgrid convection process to the resolved large-scale flow, the three different closures are used. Tendencies for temperature, water vapor, cloud liquid water, cloud ice, and zonal/meridional wind are calculated with this scheme. The convective rain and snow rates are also computed and analyzed. We refer to this scheme as "Tiedtke scheme" in this study.

3.6.2. Machine Learning Scheme

The backbone architecture for the selected NN is a BiLSTM. Our implementation is a BiLSTM based on the winner of the 5$^{\text{th}}$ place in the Kaggle competition, "YA HB MS EK" (Lin et al. 2024), and considers sequences along the model height dimension for each column. We selected this approach due to its accessibility and the demonstrated effectiveness of BiLSTMs in capturing vertical profiles for atmospheric parameterization tasks (Hafner et al. 2025a; Ukkonen and Chantry 2024; Yao et al. 2023). Furthermore, in the Kaggle competition, the solution of the 5$^{\text{th}}$ placed team only had a difference of 0.0037 in its coefficient of determination R^2 compared to the 1$^{\text{st}}$ place ("greySnow") on the private leaderboard, which we do not expect to make a significant difference for the coupled online skill. Our architecture is shown in Figure 3.7.

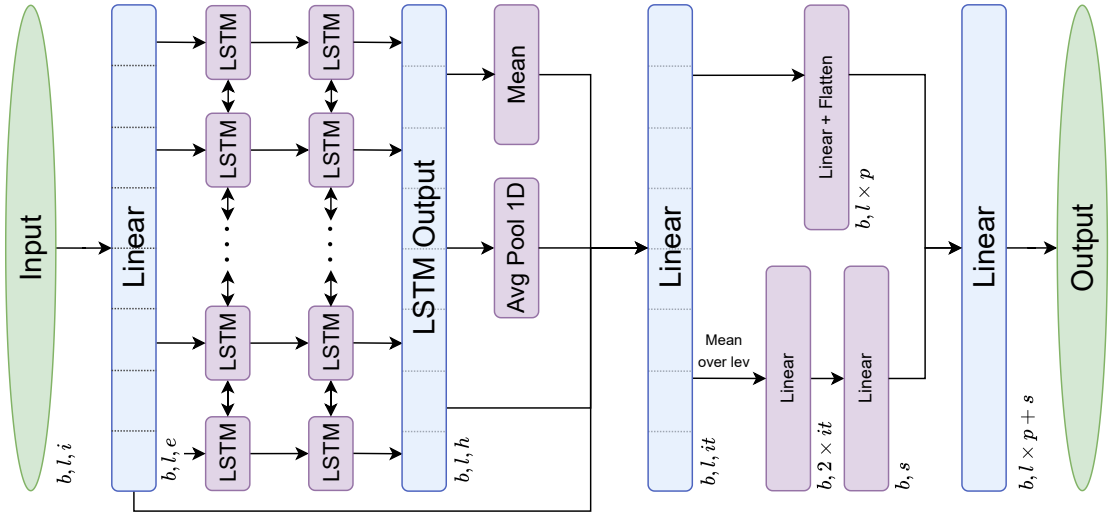

Figure 3.7.: The BiLSTM architecture developed by the 5th place Kaggle competition winner "YA HB MS EK", and used in the work presented in this article. Tensor dimensions are visualized in the lower right corner of the individual layers. The tensor dimensions shown in the figure are the batch dimension b, the column height level dimension l, the input dimension i, the encoding dimension e, hidden dimension h, iter dimension it, output scalar dimension s, and the output profile dimension p. In the blue-marked layers, the horizontal dotted lines indicate operations restricted to the last dimension, thereby preserving "vertical locality". Adapted with permission from Heuer et al. (2025).

The numerical values of the various dimensions shown in Figure 3.7 are given in Table B.1. The inputs to the ML model used in this work were inspired by Hu et al. (2025) and consist of the input variable set:

$$I = \{T, RH, q_\mathrm{l}, q_\mathrm{i}, \chi_\mathrm{liq}, u, v, \dot{T}_{t-1}, \dot{q}_{\mathrm{v},t-1}, \dot{q}_{\mathrm{l},t-1}, \dot{q}_{\mathrm{i},t-1}, \dot{u}_{t-1}, \dot{T}_{t-2}, \dot{q}_{\mathrm{v},t-2}, \dot{q}_{\mathrm{l},t-2}, \dot{q}_{\mathrm{i},t-2}, \dot{u}_{t-2}\},$$

with temperature T, relative humidity RH, cloud liquid water q_l, cloud ice q_i, liquid partition χ_liq, zonal wind u, meridional wind v, and water vapor q_v. All variables with a dot superscript are convective tendencies from the last ($t-1$ subscript) or second to last ($t-2$ subscript) timestep. The liquid partition χ_liq is a function of the temperature and has a value of 1 for temperatures above 0 °C and 0 for temperatures below −20 °C. Between −20 °C and 0 °C the function varies linearly as shown in Figure 2 of Hu et al. (2025).

This input set is similar to the inputs the conventional Tiedtke scheme uses, but also includes atmospheric variables from the two previous timesteps. The choice to include inputs from the timesteps $t-1$ and $t-2$ was also inspired by Hu et al. (2025) and can be motivated by the fact that by suppressing access to the high-resolution state, the evolution of the low-resolution state is conditionally dependent on the low-resolution states of previous timesteps as argued in Beucler et al. (2025). Furthermore, by incorporating information from previous time steps,

especially from the thermodynamic variables temperature and water vapor, the scheme gains the capability to capture convective memory effects (Colin et al. 2019).

The model outputs the following set of variables:

$$O = \{\dot{T}, \dot{q_v}, \dot{q_l}, \dot{q_i}, \dot{u}, \dot{v}, \mathcal{P}_{\text{rain}}, \mathcal{P}_{\text{snow}}\},$$

with the two 2D variables convective rain rate $\mathcal{P}_{\text{rain}}$ and convective snow rate $\mathcal{P}_{\text{snow}}$. The other variables are 3D tendencies for temperature, water vapor, cloud liquid water, cloud ice, and zonal/meridional wind.

We implemented our NNs in PyTorch (Ansel et al. 2024) and PyTorch Lightning (Falcon and The PyTorch Lightning team 2019). Inspired by the Kaggle competition, we generally chose AdamW as the optimizer (Loshchilov and Hutter 2019).

3.6.3. Loss Function

Total loss

The total per-sample loss during training ℓ_{tot} combines the Huber loss ℓ_{Huber} with the CL ℓ_{conf}, the "difference" loss ℓ_{diff}, and a physics-informed loss ℓ_φ grouping the residual of the enthalpy, mass, and momentum budgets. These terms are explained in the following subsections, and the overall loss is computed as

$$\ell_{\text{tot}}(\hat{y}, y) = \alpha\, s \cdot \ell_\varphi\left(x, \hat{y}\right) + (1-\alpha) \cdot \left[\ell_{\text{Huber}}(\hat{y}, y) + \ell_{\text{diff}}(\hat{y}, y) + \ell_{\text{conf}}\left(\hat{y}_{\text{loss}}, \hat{y}, y\right)\right]. \tag{3.7}$$

The parameter α serves as a tunable hyperparameter that governs the relative weight of the physically informed loss terms. To ensure an approximately equal contribution from both the data-driven and the physics-based components, we introduced another hyperparameter s. We initially trained the model without minimizing the physical residuals, instead quantifying their magnitude during this phase. Empirical analysis revealed that a scaling factor of approximately $s = 385$ effectively balances the magnitudes of these terms. This factor was subsequently applied to the summed physical residuals prior to their integration into the overall loss function, thereby enabling stable and effective backpropagation during subsequent training iterations.

Huber loss

As the Huber loss and other combinations of the L_1 and L_2 loss terms were used successfully by many teams in the Kaggle competition, we chose the Huber loss with hyperparameter $\delta = 1$ as our base loss. An L_2 loss is applied for absolute biases between predictions $\hat{y}$ and targets y smaller than δ, and an L_1 loss otherwise:

$$\ell_{\text{Huber}}(\hat{y}, y) = \begin{cases} 0.5 \cdot (y - \hat{y})^2, & \text{if } |y - \hat{y}| < \delta \\ \delta \cdot \left(|y - \hat{y}| - 0.5 \cdot \delta\right), & \text{otherwise.} \end{cases} \tag{3.8}$$

Physics-informed loss

The physical loss ℓ_φ is introduced to reduce enthalpy, mass, and momentum conservation errors in the ML scheme during training. Note that the conventional scheme in ICON strictly conserves these quantities in the vertical or converts atmospheric water phases to precipitation. For numerical stability and ease of implementation, we implemented the calculation of the physical terms in the BiLSTM architecture in non-dimensional form. The constants we chose for non-dimensionalization are the following:

$$\begin{aligned} g_0 &= 9.806\,65\,\mathrm{m\,s^{-1}}, \\ t_0 &= 10\,\mathrm{s}, \\ \rho_{\mathrm{h2o}} &= 1000\,\mathrm{kg\,m^{-3}}, \\ c_p &= 1004.64\,\mathrm{J\,K^{-1}\,kg^{-1}}. \end{aligned}$$

The choice of these scales was physically motivated and their numerical values were taken from the ICON model, except for the timescale, which was chosen so that the derived scales for, e.g., length, energy, temperature, and pressure are reasonably close to statistical average values of the dataset. Non-dimensional variables are henceforth denoted with tildes and more details about the non-dimensionalization of the physical terms can be found in Appendix B.1.

Our physics-informed loss $\ell_\varphi = \widetilde{H}_{\mathrm{res}} + \widetilde{m}_{\mathrm{res}} + \widetilde{u}_{\mathrm{res}} + \widetilde{v}_{\mathrm{res}}$ sums the non-dimensional residual fluxes of conserved variables, which were calculated as follows:

$$\widetilde{H}_{\mathrm{res}} = \int_{\tilde{p}_{\mathrm{top}}}^{\tilde{p}_{\mathrm{bot}}} \left(\frac{\partial \widetilde{T}}{\partial \widetilde{t}} - \frac{\partial \widetilde{q_{\mathrm{l}}}}{\partial \widetilde{t}} \cdot \widetilde{L}_{\mathrm{v}} - \frac{\partial \widetilde{q_{\mathrm{i}}}}{\partial \widetilde{t}} \cdot \widetilde{L}_{\mathrm{s}} \right) d\tilde{p} - \widetilde{L}_{\mathrm{v}} \cdot \widetilde{\mathcal{P}}_{\mathrm{rain}} - \widetilde{L}_{\mathrm{s}} \cdot \widetilde{\mathcal{P}}_{\mathrm{snow}}, \tag{3.9}$$

$$\widetilde{m}_{\mathrm{res}} = \int_{\tilde{p}_{\mathrm{top}}}^{\tilde{p}_{\mathrm{bot}}} \left(\frac{\partial \widetilde{q_{\mathrm{v}}}}{\partial \widetilde{t}} + \frac{\partial \widetilde{q_{\mathrm{l}}}}{\partial \widetilde{t}} + \frac{\partial \widetilde{q_{\mathrm{i}}}}{\partial \widetilde{t}} \right) d\tilde{p} + \widetilde{\mathcal{P}}_{\mathrm{rain}} + \widetilde{\mathcal{P}}_{\mathrm{snow}}, \tag{3.10}$$

$$\widetilde{u}_{\mathrm{res}} = \int_{\tilde{p}_{\mathrm{top}}}^{\tilde{p}_{\mathrm{bot}}} \frac{\partial \widetilde{u}}{\partial \widetilde{t}}\, d\tilde{p}, \tag{3.11}$$

$$\widetilde{v}_{\mathrm{res}} = \int_{\tilde{p}_{\mathrm{top}}}^{\tilde{p}_{\mathrm{bot}}} \frac{\partial \widetilde{v}}{\partial \widetilde{t}}\, d\tilde{p}. \tag{3.12}$$

$\widetilde{L_v}$ and $\widetilde{L_s}$ are the non-dimensionalized latent heat of vaporization and sublimation. The residual fluxes for the conserved quantities (enthalpy $\widetilde{H}_{\mathrm{res}}$, mass $\widetilde{m}_{\mathrm{res}}$, zonal momentum $\widetilde{u}_{\mathrm{res}}$, and meridional momentum $\widetilde{v}_{\mathrm{res}}$), were calculated following Equations (3.9)–(3.12) by integration over the pressure coordinate, necessitating the inclusion of mid-level and surface pressure as inputs to the neural network. In the integrals, the pressure coordinate ranges from the pressure level of the highest predicted level $\tilde{p}_{\mathrm{top}}$ to the surface pressure $\tilde{p}_{\mathrm{bot}}$. These pressure variables were utilized solely for computing differences between pressure half-levels within the model code, which were then employed in the residual flux calculation and were not used in the forward pass of the network itself. Equations (3.9) and (3.10) do contain terms for q_{v}, q_{l},

and q_i only, as rain and snow are not treated as 3D resolved tracers in the setup of ICON and the convective parameterization respectively.

Adding these residual fluxes to the loss function in Equation (3.7) effectively encouraged the model to redistribute the conserved quantities in a column instead of introducing non-physical sources or sinks. As a result, the NNs trained in this manner are no longer purely data-driven, but rather physics-informed.

Improving the output's vertical structure via the "difference loss"

Inspired by the 2nd place ("Z Lab") solution of the Kaggle competition (Lin et al. 2024), to help the model learn the vertical structure of each predicted profile, we included an additional loss term $\ell_{\text{diff}}(\hat{y}, y)$ that quantifies the error between real and predicted differences of vertically adjacent levels:

$$\ell_{\text{diff}}(\hat{y}, y) = \sum_{i=1}^{N_{\text{lev}}-1} \ell_{\text{Huber}}(\hat{y}_{i+1} - \hat{y}_i, y_{i+1} - y_i), \tag{3.13}$$

where i indexes the vertical level and N_{lev} is the total number of vertical levels.

Confidence loss

Finally, inspired by the first place solution of the Kaggle competition from "greySnow" and the AlphaFold loss function (Jumper et al. 2021), we implemented a technique in which the NN estimates its own prediction error. The method introduces a second prediction head by doubling the number of output neurons in the final layer, where the second half of the output layer predicts the error of the predictions $\hat{y}_{\text{loss}}$. Combining these loss predictions and minimizing the resulting "confidence-loss" term defined as:

$$\ell_{\text{conf}}\left(\hat{y}_{\text{loss}}, \hat{y}, y\right) = \ell_{\text{Huber}}\left(\hat{y}_{\text{loss}}, \ell_{\text{Huber}}\left(\hat{y}, y\right)\right) \tag{3.14}$$

ensures that the network learns to estimate its own loss as accurately as possible. In practice, the model is able to anticipate when its predictive skill is reduced because of high variability in the output due to, e.g., latent drivers, or when predictions are made in regions of the input feature space containing few training samples.

3.6.4. Confidence-Guided Mixing

On the validation set, we estimate the empirical CDF F_{val} of the predicted-loss averaged over all outputs. In practice, we store 101 equally spaced percentiles (0 % to 100 %), which are used to approximate F_{val}. In coupled runs, each online predicted error $\hat{y}_{\text{loss}}$ is mapped to its percentile rank

$$q = 100\, F_{\text{val}}(\hat{y}_{\text{loss}}) \in [0, 100]. \tag{3.15}$$

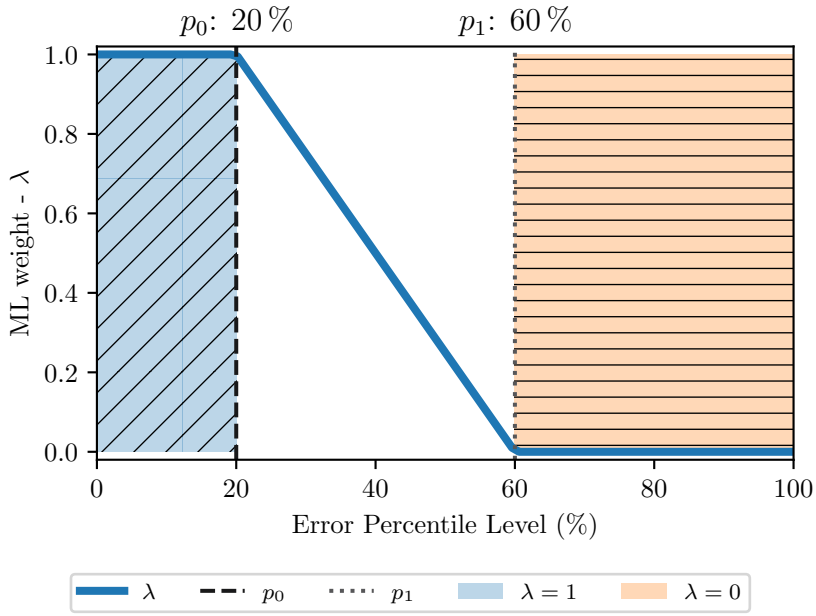

Figure 3.8.: ML weight λ as function of the predicted error percentile level. The tuning parameters p_0 and p_1 (here 20 % and 60 %) are marked by dashed and dotted lines, respectively. In blue and with slanted hatching, the area with $\lambda = 1$ (pure ML) is shown. $\lambda = 0$ (pure Tiedtke) is shown in orange and with horizontal hatching. Adapted with permission from Heuer et al. (2025).

To ensure a smooth transition between the pure ML and conventional schemes, confidence-guided mixing uses two user-set percentile levels $p_0 < p_1$ (e.g., 20 and 60), defined with respect to F_{val} (Fig. 3.8). Expressing thresholds in percent makes them scale-free and comparable across models. Given q, the ML weight λ is then defined as a linear ramp:

$$\lambda(q) = \max\left\{0,\ \min\left[1,\ 1 - \frac{q - p_0}{p_1 - p_0}\right]\right\}. \tag{3.16}$$

Predicted tendencies are then mixed component-wise as

$$\hat{y}_{\text{mixed}} = \lambda\, \hat{y} + (1 - \lambda)\, \hat{y}_{\text{Tiedtke}}. \tag{3.17}$$

Importantly, F_{val} (the mapping from error to percentile rank) is fixed from the validation set, while p_0 and p_1 offer the possibility to tune the coupled hybrid ICON model in order to better match observations; this avoids conflating the empirical percentiles with the mixing thresholds.

This confidence-guided mixing is coupled online to ICON, and the resulting tendencies are integrated with the model's other parameterized and dynamical tendencies in the dynamical core (Zängl et al. 2015).

3.6.5. Jointly Optimizing Performance and Inference Cost

The original BiLSTM used by the 5th place winner "YA HB MS EK" in the Kaggle competition has around 18 million trainable parameters. To find a balance between model skill and computational efficiency, we first used `Ray Tune` (Liaw et al. 2018) on a smaller data subset of 3 million training and 1.5 million validation samples. We varied the encoding dimension, the hidden dimension, the iteration dimension, the number of LSTM layers, and the dropout rate within the NN architecture. For the optimizer/ scheduler we additionally varied the learning rate, the weight decay parameter, the batch size, and the type of scheduler. The model marked as "Trade-off" in Figure 3.9 has about 540 k trainable parameters. This hyperparameter setting is used in the remainder of this study. More information on the search space and the optimal parameters is given in Appendix B.2.

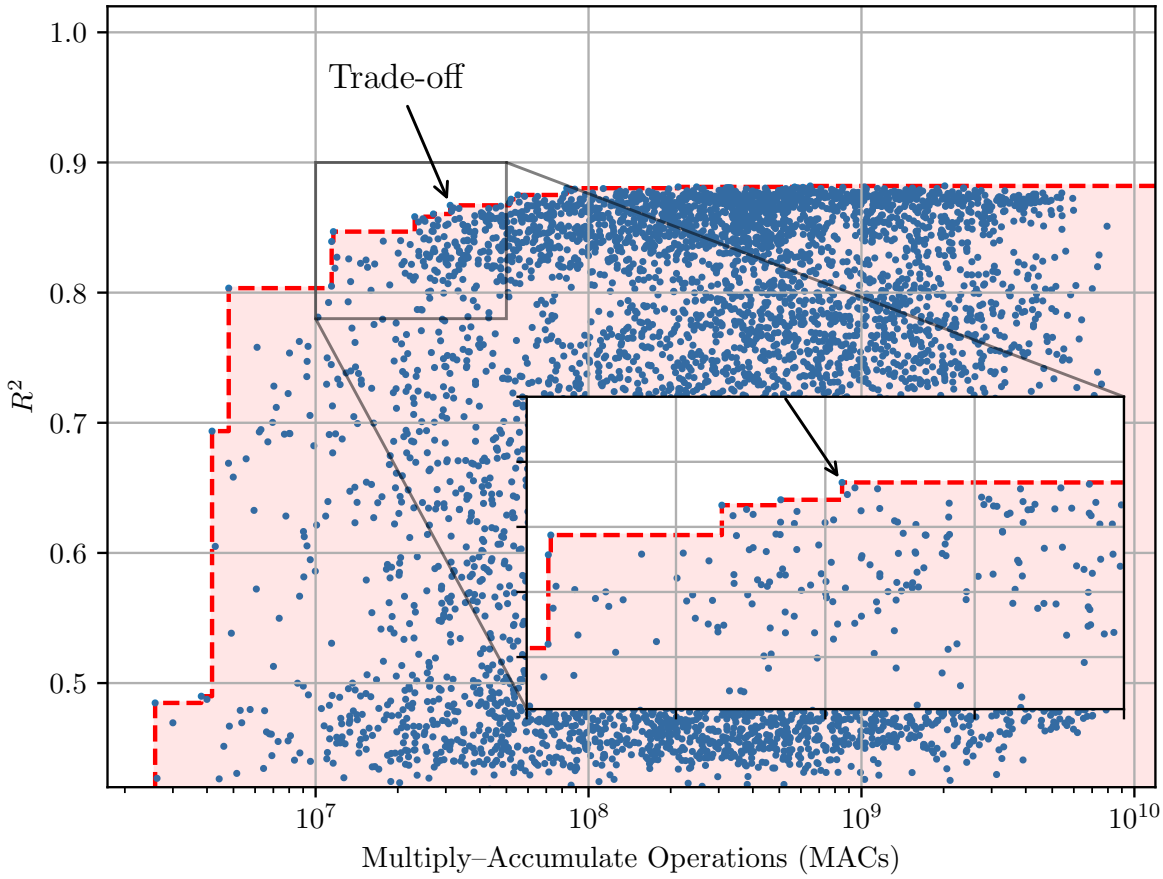

Figure 3.9.: Offline skill-complexity plane for various combinations of nine chosen hyperparameters of the BiLSTM on a smaller subset of the dataset with 3 million training and 1.5 million validation samples. The red dashed line shows the Pareto Front between the coefficient of determination R^2 and the number of multiply-accumulate operationss (MACs). The highlighted NN is selected for the remainder of this study because it strikes a suitable balance between skill and computational performance. Adapted with permission from Heuer et al. (2025).

Figure 3.9 shows all tested configurations and their coefficient of determination, as well as the number of MACs. We also measured the inference time on the CPU directly for each of the models shown and found a correlation of $\sim 99\,\%$ between MACs and inference time on CPUs, thus demonstrating that MACs are an appropriate measure of computational performance. The correlation of MACs with the GPU inference time is only $\sim 9\,\%$, meaning that if we were doing such a skill-complexity comparison for a coupled model running on a GPU, we should look at the GPU inference time directly. We decided to perform this comparison on the CPU instead of the GPU, as the NN is later coupled to the ICON model on the CPU. This might change in the future as ICON can be run on GPUs (Giorgetta et al. 2022).

The usefulness of Pareto fronts for ML models in climate modeling has been demonstrated in Beucler et al. (2025). Given multiple metrics, Pareto fronts are defined as the set of NNs for which no other NN $\mathcal{M}$ exists such that $\mathcal{M}$ shows an improvement in one metric without a worsening in any other metric relative to the original NN. Testing a limited number of other NN architectures along the Pareto front revealed that our results did not seem to be very sensitive to the specific architecture chosen (not shown).

3.6.6. Additive Noise During Training for Improved Stability

Inspired by the "Engression" framework by Shen and Meinshausen (2024) and by the results of Brenowitz et al. (2020), we made the ML schemes more robust with respect to the transfer to a new domain with slightly different distributions. We did this by including additive noise to the inputs during training of the BiLSTMs:

$$y = \mathrm{NN}\left(x + \eta\right), \quad \eta \sim \mathcal{N}\left(0, \sigma^2\right), \tag{3.18}$$

where η is a noise vector sampled from a Gaussian distribution $\mathcal{N}$ with zero mean and a tunable variance σ^2. As x and y are standardized using a Z-score, the variance is constant across variables in x. This preadditive noise can reveal some information about the true function outside the domain it was trained on, which can enable data-driven extrapolation (Shen and Meinshausen 2024).

To add noise during training, we performed a warm restart from an optimized, noise-free NN. Algorithmically, we implemented a Python class initialized with four hyperparameters: the initial noise level $\sigma_0 > 0$; the tolerated R^2 drop compared to its value before any noise addition, $\Delta R^2 > 0$; and multiplicative growth/decay factors $m_\uparrow > 1$ and $m_\downarrow \in (0, 1)$ for the noise. In the first epoch, we add Gaussian input noise with standard deviation σ_0. After each epoch, we compute the change in R^2: if the drop exceeds ΔR^2, we reduce the noise by $m_\downarrow$; otherwise, we increase it by $m_\uparrow$. After a manual search, we adopted $(\sigma_0, \Delta R^2, m_\uparrow, m_\downarrow) = (0.05, 0.1, 1.1, 0.9)$.

3.6.7. Online Coupling to ICON

We used the ICON 2.6.4 model version with a horizontal resolution of approximately 158 km × 158 km, corresponding to an `R2B4` ICON grid. The model incorporates a range of parameterized

processes, including radiation, cloud microphysics, orographic and non-orographic gravity wave drag, boundary layer turbulence, and convection. Since our approach consists in mixing the pure ML and physical convection parameterizations, our ML-based model did not replace the original Tiedtke scheme but was run alongside it. In order to initialize the convective tendencies of the two most recent timesteps needed by the ML convection scheme, we utilized the two last timesteps from the Tiedtke scheme as initial conditions.

To ensure compatibility between our ICON setup and the ClimSim data, we configured ICON with 60 vertical levels, adjusting their heights to approximately match those of the ClimSim dataset. The ML schemes' tendencies were then coupled within the troposphere, and only the lowest 42 levels (corresponding to an approximate upper pressure level of 95 hPa) were used as inputs/outputs for the scheme.

For the coupling of the ICON model implemented in FORTRAN and the ML models in Python/PyTorch, we used the FTorch library (Atkinson et al. 2025). This library enables running the ML models in inference mode during the time integration of the ICON model. After training and before exporting the ML models, we added normalization layers before and after the `forward` method of the model to take care of the preprocessing and postprocessing of the inputs and outputs during inference.

4. Interpretable multiscale Machine Learning-Based Parameterizations of Convection for ICON

4.1. Introduction

This chapter was already published in Heuer et al. (2024). As indicated in Section 1.2, the author of this thesis created all the content, including text, figures, and tables, that is presented from this publication and implemented the code[1] to reproduce this study. This chapter focuses on addressing Key Science Question 1, while also providing some insights into Key Science Question 2.

GCMs have been used since the late 1960s to answer scientific questions about our climate (Manabe and Wetherald 1967; Phillips 1956) and to project its expected changes, which are already felt across the globe (Eyring et al. 2021a). Over time, these models gradually included more and more aspects and processes of the climate system and have evolved into ESMs, including the carbon cycle and biogeochemical processes. However, the uncertainty of the simulated equilibrium climate sensitivity (ECS), i.e. the response of global surface air temperature to a doubling of carbon dioxide (CO_2) at equilibrium, has not reduced significantly in the last decades (Schlund et al. 2020). For the latest generation of ESMs, the ECS is estimated by the Intergovernmental Panel on Climate Change Sixth Assessment Report (Forster et al. 2021) at 2 °C to 5 °C. This uncertainty is about twice the uncertainty for the estimated ECS including all other scientific evidence such as emergent constraints and paleoclimates of 2.5 °C to 4 °C (Forster et al. 2021).

A large portion of this uncertainty is attributed to cloud feedbacks (Ceppi and Nowack 2021; Schneider et al. 2017), the change in cloud types and distributions in response to warming climate. Therefore, it is highly important to have a good representation of the effects of convection, which is typically a subgrid-scale process in climate models (Sherwood et al. 2014). Parameterizations based on physical process understanding, normally relying on mass-flux approaches (Arakawa and Schubert 1974; Tiedtke 1989), have been used extensively for approximating the effect of subgrid convection on the large scale. These parameterizations, however, cause some common problems in climate models (Eyring et al. 2021b), such as biases in precipitation patterns (Christopoulos and Schneider 2021; Fosser et al. 2024; Stephens et al.

[1] published under https://github.com/EyringMLClimateGroup/heuer23_ml_convection_parameterization (last access: 14.10.2025) and preserved (helgehr 2024)

2010), in the position and shape of the ITCZ (Stevens et al. 2019b), the missing representation of convectively coupled waves, and the Madden-Julian Oscillation (Kuang et al. 2005), or teleconnections (Mahajan et al. 2023) and the incorrect diurnal cycle of convection (Anber et al. 2015). These biases are reduced in storm-resolving models (Bock et al. 2020; Klocke et al. 2017; Stevens et al. 2019b, 2020).

Accurately representing convection in climate models remains a challenge due to its complex and multiscale nature. In light of recent advances in deep learning, many data-driven machine learning-based parameterizations have been developed to reduce the above-mentioned biases (Brenowitz and Bretherton 2018; Gentine et al. 2018; Iglesias-Suarez et al. 2024; Krasnopolsky et al. 2013; Otness et al. 2023; Rasp et al. 2018). These studies first used MLP neural networks in a simplified aquaplanet setup to replace the superparameterized physics in the SuperParameterized Community Atmosphere Model (SPCAM3) (Collins et al. 2006). RFs have been used as well (O'Gorman and Dwyer 2018; Yuval and O'Gorman 2020) with the advantage of guaranteeing conservation properties and physical consistency, via constraints in the sign of quantities such as precipitation, as well as on its magnitude (reducing coupled model instability). A disadvantage of RFs is however that they do not extrapolate outside their training domain at all and so are inherently limited in their application for a changing climate. They can also struggle to represent the diversity of complex data.

To combine conservation properties that are essential for a climate model, and the ability to extrapolate to some extent, Yuval et al. (2021) used MLPs to predict vertical fluxes instead of tendencies (the vertical convergence of the fluxes). More recently, they extended their work by including convective momentum transport in an idealized aquaplanet setting as well (Yuval and O'Gorman 2023). Wang et al. (2022b) used residual neural networks to emulate the physical tendencies resulting from a superparameterization of moist physics and radiation in a realistic setting with coupled simulations running stably over 10 years.

With this work we build on previous studies on data-driven convection parameterizations and ML-based schemes, targeting the ICON model (Grundner et al. 2022, 2024). We extend these approaches in several aspects. We use high-resolution data that explicitly resolve convection and employ a coarse-graining method to calculate and isolate the convective mesoscale flux that is subgrid for a coarse climate model, here ICON in a real-world setting. We benchmark a set of different machine learning methods trained on a realistic data set with orography (Dataset section). Although it can be argued to what extent explicit process separation is sensible (Randall et al. 2003), most parameterization schemes act independently (in parallel or sequentially) from each other for different subgrid processes (Giorgetta et al. 2018). For this reason, simplicity, and because the trained ML models should be easily interoperable with the GCM in a coupled mode we treat convection as a separated process. Furthermore, this enables us to use XAI methods to interpret the ML models with respect to our physical understanding of atmospheric convection. To focus on the effects of subgrid convection for coarse resolution simulations, where convection must be parameterized, we introduce a filtering technique to capture convective circulations as resolved in storm resolving simulations. Apart from making it possible to selectively replace only the conventional parameterization,

this approach allows to better interpret the physics of the learned ML model as it does not mix different processes such as convection and radiation. We propose a new way of computing the coarse-grained target quantities by not neglecting horizontal fluctuations (not applying the Boussinesq approximation) in the density as is typical for Reynolds-averaging. Additionally, we use an XAI technique to interpret the model predictions and relate the revealed connections to physical process understanding. Similarly to the spectral analysis tool by Brenowitz et al. (2020), this method builds trust in the retrieved models and can be used to evaluate the ML model, going beyond common metrics such as the root mean square error (RMSE) or the coefficient of determination.

In the end we will test the stability of the U-Net when coupled to the ICON model. Here we test the extrapolation capabilities of the ML models as they are trained on regional data and then applied on larger/global domains.

This chapter is structured as follows. First, results of the offline evaluation/benchmarking of different machine learning models are shown and their predictions interpreted using an XAI technique in Section 4.2. We will conclude Section 4.2 with an online stability test of the developed U-Net parameterizations. Finally, we discuss our results and give a conclusion of our work.

4.2. Results

This section will first introduce a model evaluation for all ML models used and then focus on a more detailed comparison of the highest performing (offline) deep and non-deep learning method in Section 4.2.1. Afterwards, in Section 4.2.2, we will investigate what the models have learned and find that, in fact, an ablated version of the U-Net (without precipitating tracers as input) learns physically explainable relations as opposed to the non-ablated version. This ablated model, in comparison with the non-ablated version, will also be tested in the online stability test section in the end of this chapter in Section 4.2.3.

The architecture of the best performing model, the U-Net, is first introduced in Section 4.2.1 and the ablation, improving online stability, is described in Section 4.2.2.

4.2.1. Machine Learning Model Benchmarking

First, we focus on the simple aggregated evaluation of the coefficient of determination (R^2) values for all examined model classes. The R^2 value is calculated as 1 minus the mean squared error of the predictions over the variance of the data. We compute the R^2 value across variables and levels, a more detailed (per variable/level) comparison is given later in Figure 4.4. All models have been hyperparameter-tuned according to the method described below. Briefly this HPO consisted of running a large ensemble of models with parameters sampled from predefined search spaces and their performance evaluated on a validation set (more details in section A.2).

Figure 4.1 displays the R^2 values for all models over all variables and levels. On the left hand side of the dashed green line the deep learning models are shown as opposed to the simpler models on the right hand side.

The R^2 value of the Random Forest is the lowest of the examined models. RFs have been used as data-driven convection parameterizations with some success (O'Gorman and Dwyer 2018; Yuval and O'Gorman 2020) in idealized settings before. Limitations in the application of RFs for realistic parameterization schemes have been observed before due to their computational inefficiency, memory requirements, and comparably low complexity (versus deep neural networks for instance), limiting their capacity to capture high dimensional features (Limon and Jablonowski 2023). The GBT model class has a strikingly high R^2 value, comparable to the ones of the deep learning methods. This suggests that these RF-based parameterization schemes could improve in performance if they were based on Gradient Boosted Trees (besides deep learning networks). The Extra Trees model has a similarly low performance as the RF. Considering that the ET model is structurally similar to RFs, including an additional element of randomness as explained above, this is not surprising. The linear models (Ridge, LinMLP, and Lasso) show relatively high performance compared to that of the RF/ET model with R^2 values of $0.68, 0.63, 0.62$. The L^2-regularization term seems to have a higher impact on the generalization capabilities of the linear model compared to the L^1-regularization in Lasso regression. The generally better performance of the linear models compared to the tree based models, RF and ET, is surprising and might be connected to the fact that linear models are able to extrapolate to unseen data points based on the linear relationships learned during training. These tree based methods, however, are limited to the range of the training data and cannot extrapolate beyond it because they predict based on averages of similar seen samples. As we are using high-dimensional data some degree of extrapolation is very probable (Balestriero et al. 2021). Another point is that in cases of high-dimensional data with many uninformative or noisy features, linear models, especially when combined with regularization techniques like Lasso, can perform better by effectively reducing the dimensionality and focusing on the most relevant features. Random Forests might not be as effective in ignoring these irrelevant features to that extent. Another option might be that the linear models are being too heavily tuned to the tropical convection problem. More on this in the discussion.

The deep learning models outperform the other methods but, e.g., for the GBT model only by a small amount. While the R^2 value for the GBT is almost as high as the value for the U-Net, the other nonlinear methods show a rapid decrease in performance when ordering by their respective R^2 value. Figure 4.1 shows that the performance difference between the various deep learning models measured by R^2 is negligible. One could suspect that the best performance of the U-Net could originate purely by chance. Therefore, we performed an extensive HPO with over 5000 ensemble members in total. The resulting median/upper/lower quartile profiles can be seen in Figure 4.2. We varied hyperparameters such as the learning rate, number of neurons/layers/blocks, or activation functions. More details on the HPO search spaces can be found in section A.2. A visualization of the HPO and the training and validation process in general is shown in Figure A.8. We notice that the U-Net has a consistently lower error than the

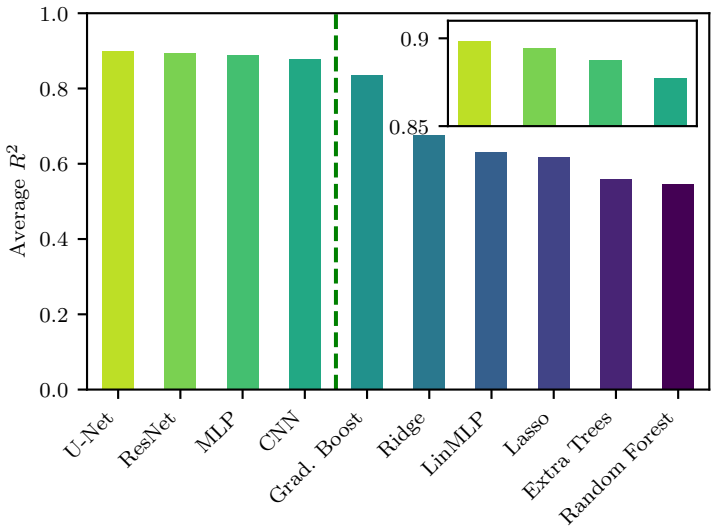

Figure 4.1.: Coefficient of determination (R^2) on a test set for different types of models. All models were hyperparameter-optimized, and the best models were then trained on the whole data set. The deep learning methods are displayed on the left of the green dashed line and the non-deep learning methods on the right of it. The inset in the top right shows a zoomed-in version of the R^2 for the deep learning models. Adapted with permission from Heuer et al. (2024).

other models, and the upper quartile of its distribution is on the same level as the lower quartile of the second best performing model, the ResNet. The difference between the other model classes is smaller, and the spread around each median profile is larger than for the U-Net.

Furthermore, the model complexity of the U-Net is comparatively low. As it can be seen in the number of parameters of our network configurations (Table A.1) and Figure A.6, the most complex (judging by number of parameters) deep learning model is the ResNet with more than four times the number of parameters of the U-Net. The MLP architecture has the lowest number of parameters, the U-Net has the second lowest number before the CNN and ResNet. Despite this, the U-Net shows the consistently lowest error on the validation/test set (see Figure 4.2) over a large set of hyperparameter configurations, presumably because of its multiscale architecture and the resulting ability to capture multiscale problems such as convection well.

Based on these results, we will focus on the respectively best performing deep and non-deep learning models from now on. These models are the U-Net and the GBT model as seen in Figure 4.1. We first compare the U-Net and GBT flux predictions with the true values for F_u^{sg}, F_v^{sg}, F_h^{sg}, $F_{q_c}^{\mathrm{sg}}$ over all levels. The results can be seen in Figure 4.3, and a corresponding plot showing the distribution for the remaining tracer subgrid fluxes can be seen in Figure A.1. The correlation is always higher for the U-Net predictions, and for both models the meridional momentum fluxes are the hardest to predict. This has been noted before e.g., for a data-driven gravity wave scheme (Espinosa et al. 2022). The diurnal cycle and its annual variability are typically more pronounced (Giglio et al. 2022) for the meridional wind and can be out of phase in the northern and southern hemisphere (Ueyama and Deser 2008). We assume that, therefore,

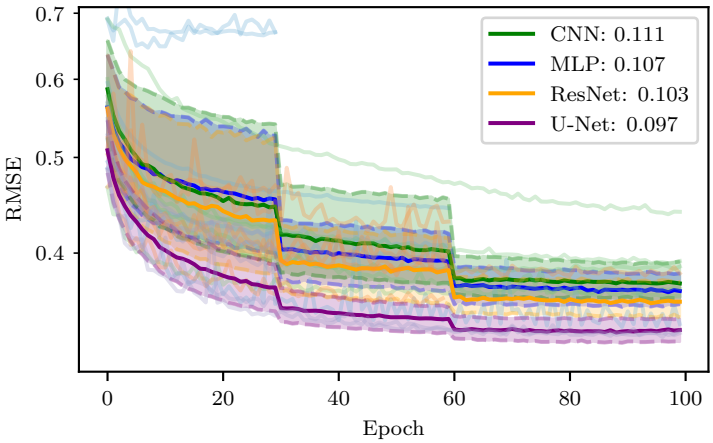

Figure 4.2.: Root mean squared error during HPO on the validation set of the four different deep learning methods. The straight thick lines correspond to the median of the HPO ensemble, the shaded areas are drawn in between the first and third quartile. Additionally, ten realizations for each DL method are shown in similar colors. The legend shows the minimum of the validation loss for each of the methods. The scheduler of the HPO filters badly performing runs after 30 and 60 epochs, causing the steps in the profiles. For this task we used the `AsyncHyperBandScheduler` (Li et al. 2018) of the `Ray Tune` library (Liaw et al. 2018). Adapted with permission from Heuer et al. (2024).

it is a challenge for the ML models to predict the meridional momentum flux receiving as input only the large-scale state, which my not adequately represent the nuances of meridional dynamics.

Especially for high values of the flux, both models tend to underestimate the true flux, which can be seen by the points below the diagonal in plot b). To a similar extent, this trend can also be seen for the fluxes F_u^{sg} and $F_{q_v}^{\text{sg}}$. The mentioned fluxes of the GBT show a slight corresponding overestimation for low flux values. In contrast to that, the U-Net data distribution is more symmetric about the main diagonal. This means that there is no or a very small systematic under- or over-prediction for these values by the U-Net. In general, the spread around the diagonal is bigger for the GBT than for the U-Net. This confirms the better performance of the U-Net seen in Figure 4.1 based on R^2 values.

After having examined the model performance aggregated over all levels we now look at the average R^2 values of the 3D variables on individual vertical levels. This is shown in Figure 4.4, again for the U-Net and GBT. Some vertical levels are not shown in the figure because the variation of the variables on these levels is close to zero. We determined the variables for which this is true by first finding the 99th percentile of their absolute values. Then, for each variable all levels in which the computed percentile is below 1 % of the maximum percentile for the variable were excluded from the plot.

This method filters all levels which show significantly less variation compared to all other levels. Looking at Figure 4.4 we filtered out the lower tropospheric values for the ice and snow tracers as well as the higher tropospheric values for cloud water and rain tracers. This is reasonable because we do not expect much snow/ice in the lower troposphere of the tropics, and similarly, the temperatures are too low for cloud water/rain to exist close to the tropopause.

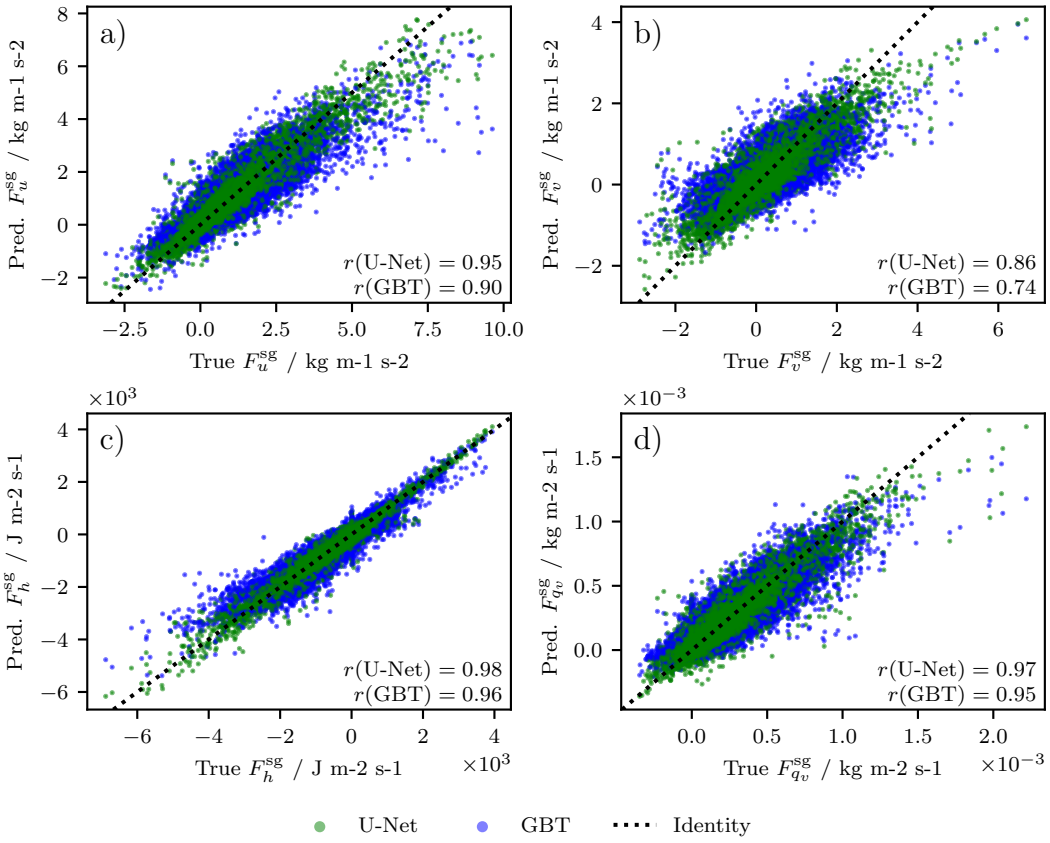

Figure 4.3.: Scatter plot for the subgrid fluxes of a) zonal, b) meridional momentum, c) liquid/ice water static energy, and d) specific humidity. Data for the U-Net is shown in green, for the GBT in blue, and the diagonal is marked by a dotted line. The Pearson correlation coefficient r between the true and the predicted subgrid flux is noted in the lower right corner of each plot for both U-Net and GBT. Adapted with permission from Heuer et al. (2024).

Comparing the plots in Figure 4.4, the two models show similar patterns as seen, e.g., for the F_v^{sg} curve, but the GBT curves are mostly shifted towards lower R^2 values compared to the U-Net. For most variables we find a clear advantage of the U-Net in the upper layers and around the height of the planetary convective boundary layer at ~ 1 km. Other than for tracer species on levels in which the corresponding concentration is typically very low, the models show difficulties to predict the subgrid momentum fluxes compared to other variables, as is particularly visible for F_v^{sg}. For subgrid momentum transport in general this has been noticed before in Yuval and O'Gorman (2023). This problem could arise from the fact that the sign of the subgrid convective momentum flux depends on the nature of convective organization (LeMone 1983; Yuval and O'Gorman 2023), which is not resolved in the coarse data. A few points are marked by red circles, which correspond to higher R^2 value for the GBT. Most of these are close to the R^2 U-Net value (within an R^2 relative deviation of 1.5 %) except for the low ice and snow tracer values. Here we assume that the GBT shows an increased performance due to the small number of training data and its lower model complexity. Using the U-Net

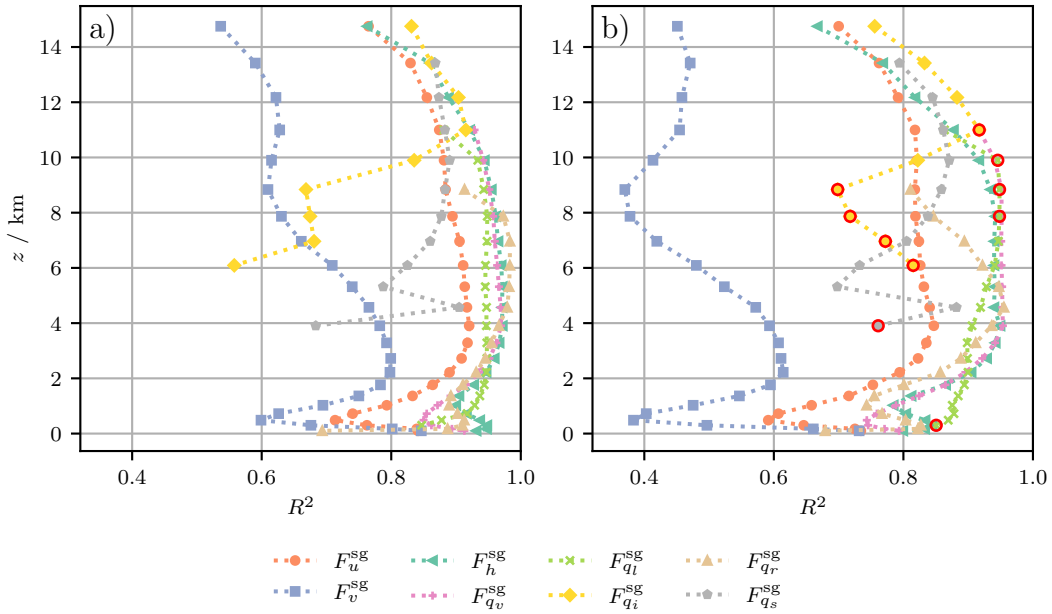

Figure 4.4.: Average R^2 profile for all subgrid flux variables for a) the U-Net and b) GBT model. Data points where the GBT model actually has a higher R^2 than the U-Net are additionally marked by a red circle. Adapted with permission from Heuer et al. (2024).

increases the mean R^2 value of all variables. The highest improvement by using the U-Net instead of the GBT can be seen for F_v^{sg} with an average R^2 improvement of 0.19 and the second highest for F_u^{sg} with a gain of 0.09. In the vertical, the highest average increase in skill is observed in the boundary layer. On these lower model levels, the dynamics are typically more complex/turbulent and therefore the higher model complexity of the U-Net is especially beneficial. This complexity in the planetary boundary layer arises from different mechanisms such as direct surface forcings, e.g., heat and moisture flux to/from the atmosphere as well as surface drag. Also, the dynamics are inherently more turbulent because of large wind velocity gradients and shear. Furthermore, diurnal variations and therefore general variability are much higher close to the surface layer than in the upper troposphere/atmosphere due to the direct surface interaction.

The 2D fields are also predicted more skillfully by the U-Net, the R^2 values for all five predicted 2D variables are higher for the U-Net than for the GBT. As an example, the true and predicted precipitation distribution is shown in Figure A.2. Even though the R^2 values for precipitation are similar (0.897 vs. 0.860), the U-Net predicts the extremes of the distribution much more accurately. For instance, the 95th percentile of the true distribution and the predicted distributions of U-Net and GBT are approximately 22.28 mm h^{-1}, 19.75 mm h^{-1}, and 16.83 mm h^{-1}. This shows that the U-Net captures the high precipitation cases much better than the GBT.

Looking at the spatial distribution of the normalized RMSE across all variables (see Figure A.3) we notice that both models have a lower error in the region of the ITCZ and an increase in

error towards higher latitudes. This reflects the difference in the abundance of training data as seen in Figure 3.2.

4.2.2. Explainability of U-Net and GBT

Having looked into the prediction results we now want to find out what the models actually have learned in order to predict the parameterization output. This will be based on the SHAP (Lundberg and Lee 2017) library which analyzes ML model predictions using a game theoretic approach. A SHAP value $shap(x = x_0, y)$ gives the deviation in an output variable y due to a specific value x_0 of the variable x from the average prediction of y over a given data (sub)set $\mathcal{X}_b$. We used the `DeepExplainer` class (Lundberg and Lee 2017) as an efficient explainer for deep neural networks, and the `TreeExplainer/KernelExplainer` class for decision tree-based models such as GBT.

Figure 4.5 a) shows the mean absolute values of the calculated SHAP values for the U-Net model. These correspond to feature importances and in this case show that the model mainly focuses on using the precipitating tracer species to predict the subgrid fluxes. The top plot shows that q_r dominates the importance attribution with over 50 % of all values. As second most influential feature we see q_s, another precipitating tracer species, even though it is only highly influential in the upper layers. Additionally, one notices that the standard deviation is relatively large for q_r/q_s, indicating the ambiguity of the learned relations. This is a first hint that the model learned non-causal relationships between convective precipitation and convective subgrid fluxes. When the model "sees" coarse-grained precipitation in the data it predicts that convective subgrid fluxes must be present. This behavior can also be observed in a more detailed analysis of the SHAP values (Figure A.4). Learning this connection is consistent as the link between convective precipitation and convective fluxes in the tropics is especially pronounced. Nevertheless, this represents a weakness and non-causal link as the ML parameterization would never/rarely encounter convective precipitation in a coupled setting if it would not predict the effect of convective fluxes before.

To prevent the model from learning these non-causal connections we trained another set of models with less input variables. We left out the precipitation input tracer species q_r and q_s. For this ablated model versions we performed a new HPO. These models will be discussed henceforth. The R^2 performance of both models (U-Net and GBT) on the test set decreases marginally, by ~ 0.03, by ablating the precipitating tracers as inputs. A third HPO was performed neglecting horizontal density fluctuations, with the result that the validation error increased for all model classes by about 4 %, and for the MLP only negligibly. This is a hint that the irreducible error of the models increases by neglecting density fluctuations.

The feature importances for the ablated U-Net are displayed in plot b) of Figure 4.5. A more spread-out feature importance assignment can be seen in this plot: the difference between highest and lowest valued feature is only 14 % which is much less than 50 % as before. This model now does not rely on spurious correlations between precipitation and convective subgrid fluxes and should generalize better outside the training domain. The general trend for most

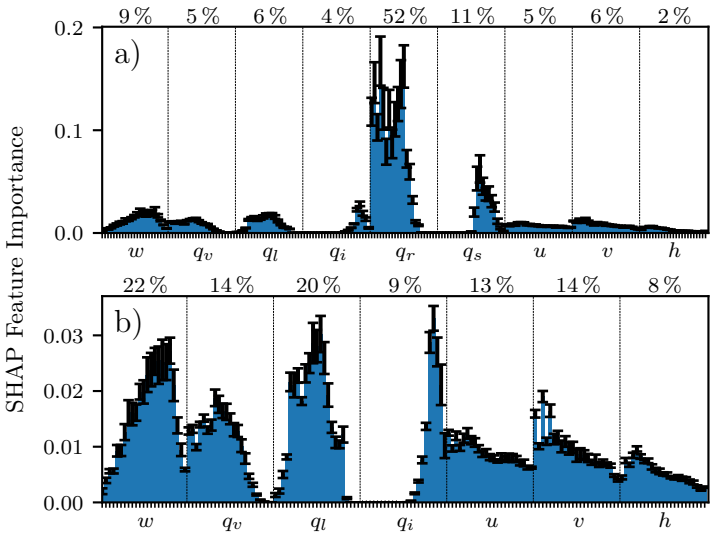

Figure 4.5.: Feature importances (i.e., the mean absolute values of the calculated SHAP values) of input variables for a) the full U-Net model, and b) the ablated (without qr,qs) U-Net model. The mean feature importance is visualized by the height of the bar, and the standard deviation over five different computations by the errorbars. The x-axis shows different height levels for each variable, increasing from left to right. Vertical lines separate the variables. The integrated fraction of feature importances over all vertical levels is written above each variable range. Adapted with permission from Heuer et al. (2024).

variables seen in Figure 4.5 indicates that the model focuses more on the lower model levels, and the importance is decreasing with height. For w, q_l, and q_i this is not the case, the feature importance peaks at higher model levels. The specific cloud ice content is only present at higher altitudes as already discussed. For the cloud water content we have very low concentrations at low model levels as clouds generally form in the boundary layer during daytime (Stull 1988), and the mean vertical velocity profile also shows higher values at greater altitudes, indicative of the importance of shear such as on mesoscale convective system organization (Rotunno et al. 1988).

We looked at the feature importance in Figure 4.5 but did not discuss the influence of an input on the various output variables. For this, we now first explain the method and then discuss the results. For ease of notation, we focus here on a single output model with output variable y as before, but this can easily be generalized to higher dimensional output. To get the average effect of an input variable x_i on the output variable y we first define the fluctuation of x_i for sample j as $x'_{ij} = x_{ij} - \langle x_i \rangle$, where the brackets $\langle \cdot \rangle$ denotes the average value over x_i in the set $\mathcal{X}$. The data set $\mathcal{X}$ is a random subset of the whole data set as to save computational costs. Now, we define the normalized fluctuation as

$$\hat{x}_{ij} = \frac{x'_{ij}}{max_k(|x'_{ik}|)}. \tag{4.1}$$

The weighted average effect of x_i on y can now be quantified in a similar way as in Beucler et al. (2024) in a vector $\mathbf{S}$, with

$$S_i = \langle \hat{x}_{ij} \cdot shap(x_{ij}, y) \rangle_j. \tag{4.2}$$

A positive S_i expresses an increasing/decreasing y for an increasing/decreasing x_i independently of other values, and for a negative S_i we have the opposite effect. For a multi-output model this vector $\mathbf{S}$ becomes a 2D matrix S_{ij} quantifying the influence of the ith input on the jth output. We will refer to the SHAP values obtained by this method as weighted SHAP values from now on.

Applying this method to the trained U-Net model gives the matrix visualized in Figure 4.6. We see many interpretable, vertically local influences (main diagonal patterns) in this figure, for example, controlling for q_l, there is a mainly negative influence of specific humidity q_v on $F^{\mathrm{sg}}_{q_v}/F^{\mathrm{sg}}_{q_l}$ visible. As previously observed by Beucler et al. (2018), this vertically local drying effect is plausibly related to the entrainment of water vapor into convective plumes and its subsequent downwards advection (Beucler et al. 2018). Moreover, an increase in water vapor also increases the moisture gradient to the environmental air and leads to the entrainment of drier air. The local drying effect is seen for levels in the lower to middle troposphere, approximately at 700 m to 5 km. Furthermore, we see a slightly positive impact and moistening flux of the lower model levels on higher levels. This is indicative of the decrease in air density for increased water vapor content and the decreased lapse rate for buoyant air parcels (and therefore higher convective instability). For cloud liquid water q_l the opposite effect can be observed on the convective subgrid fluxes of q_v/q_l. This learned correlation can be understood by looking at the condensation process of water vapour. When water condenses in an atmospheric grid cell, latent heat is released and the air becomes more buoyant. This in turn can lead to more condensation and therefore to moisture convergence in the area and cloud formation. Furthermore, more liquid water can lead to precipitation. The evaporation of falling raindrops can consequently lead to an increase in local humidity, especially if the layers below are far from saturation. Finally, hygroscopic effects could play a role as cloud droplets can act as condensation nuclei, attracting more water vapor and leading to cloud growth.

A direct comparison with the linearized response functions from Brenowitz and Bretherton (2019) and Kuang (2018) is difficult as we use different variables and a dataset from a non-idealized simulation (e.g., no aquaplanet configuration, active diurnal cycle, and spherical simulation domain). Nevertheless, for the influence of water vapor on the subgrid flux of water vapor and cloud liquid water we see similarities to the response of Q_2 (apparent moistening) to the total nonprecipitating water mixing ratio in Brenowitz and Bretherton (2019). For both analysis methods a vertically local negative influence is visible. In the study Kuang (2018) this response is similarly traced back to the impact of relative humidity on the specific humidity tendency. Furthermore, we also observe a positive convective heating response to an increase in moisture (influence of q_v on F^{sg}_h) as shown in both studies, although more local in this study as opposed to a non-local heating of higher layers in response to a moistening lower troposphere.

Apart from that, the main visible signatures are visualized in the insets of Figure 4.6. Inset a) shows the influence of w on F^{sg}_h. The main pattern is in the upper layers where we can

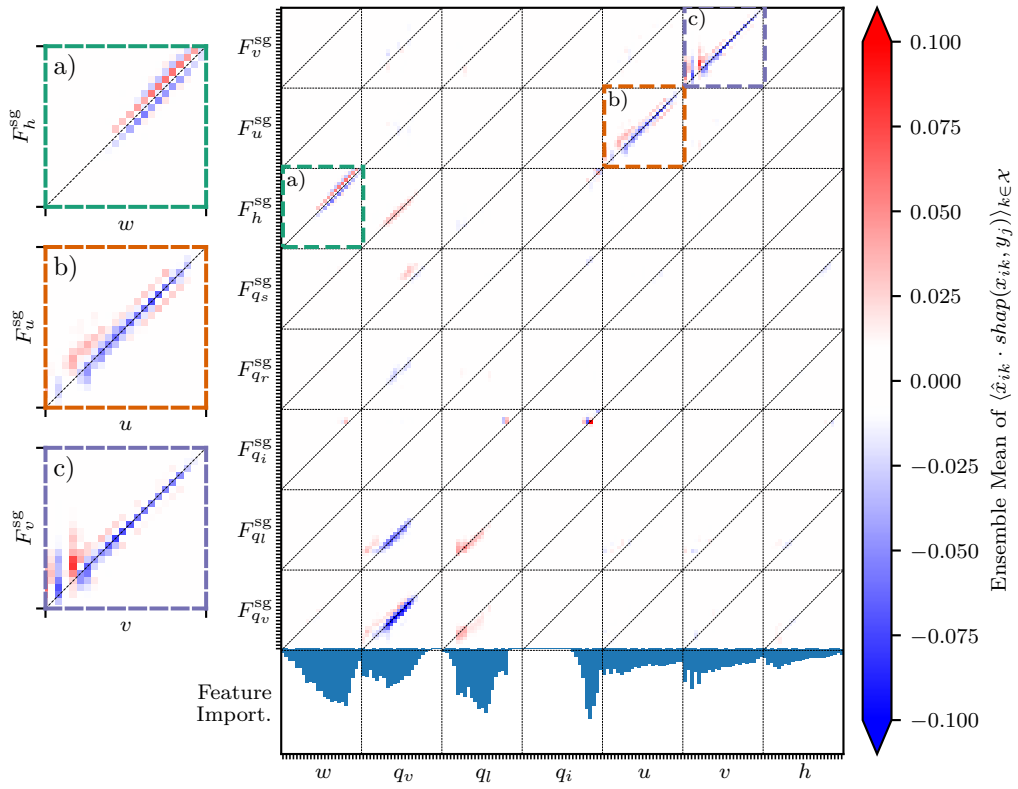

Figure 4.6.: Ensemble mean of weighted SHAP values aggregated according to Equation (4.2) for the U-Net. The variables q_r, q_s were ablated. The height level for each variable is increasing from left to right / from bottom to top. The feature importance depicted in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. Insets a), b), and c) show a zoom into the plot for three specific variable pairs, the colors indicate which inset corresponds to which part of the large plot. Adapted with permission from Heuer et al. (2024).

see primarily a positive super- and negative sub-diagonal (S_{ij} with $j = i - 1$ and $j = i + 1$, respectively). This means that cells with a high vertical velocity have a positive influence on the subgrid flux in the cell above them and a negative influence below them respectively. Considering that mesoscale convergence and large scale ascent can initiate/enforce convective cells (Kalthoff et al. 2009), this seems reasonable. Below the convective region, the atmospheric column becomes more stably stratified, explaining the negative sub-diagonal of the figure. In Inset b), a negative diagonal pattern with some positive signatures above can be observed. Consequently, high horizontal wind speeds imply a positive horizontal momentum flux to higher levels. This signifies that the U-Net has learned a downgradient diffusive momentum flux parameterization. We also see a positive pattern in the sub-diagonal for higher levels looking at subplot b). Vertical wind shear has been found to be an essential ingredient for long-lived and well-organized convective storm cells (Doswell and Evans 2003; Roca et al. 2017; Rotunno et al. 1988). A very similar pattern can be observed in Inset c), the main difference is that for lower levels there are a few vertically non-local transport signatures. These patterns

are consistent (with relative standard deviations of max $\sim 10\,\%$) over different realizations of $\mathcal{X}$ so that the result here seems not to be dependent on the set $\mathcal{X}$.

As a comparison, the corresponding weighted SHAP values for the GBT are displayed in Figure 4.7. First, the GBT feature importances have a much less regular pattern and look more "randomly" distributed. These patterns show a less coherent picture and are not so easily interpretable. Looking at the aggregated feature importance, both models weigh the liquid/ice water static energy the least. The GBT model weighs the specific humidity higher in its predictions with an aggregated importance of 29 % compared to the U-Net with 14 %. As most important features for the U-Net, on the other hand, we have the vertical velocity w and cloud water content q_l. These two variables are also part of the condition formulated in Equation (3.4) for convective conditions in a grid cell. Therefore, it is reasonable that the network learns to pay attention to these inputs.

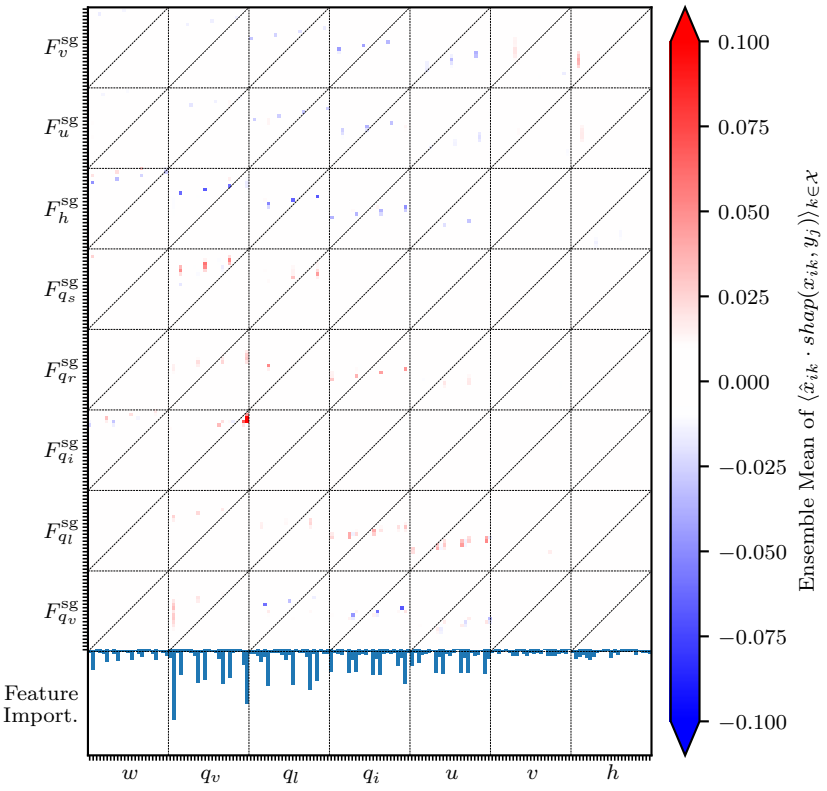

Figure 4.7.: Ensemble mean of weighted SHAP values aggregated according to Equation (4.2) for the GBT model. The variables q_r, q_s were ablated. The feature importance depicted in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. Adapted with permission from Heuer et al. (2024).

Since the weighted SHAP values displayed in Figure 4.6 consistently show vastly different patterns than in Figure 4.7, we used the same method for the RF as well and got a similar picture to what is displayed here for the GBT. In order to rule out a dependence of the obtained results on the Shapley value approximation method, we also used the `KernelExplainer` (Lundberg and Lee 2017) as an alternative to the `TreeExplainer`. The resulting weighted SHAP values

have almost the same form as for the `TreeExplainer` class, emphasizing that our results are independent of the explanation method. We also looked at the standard deviations of all weighted SHAP value plots and observed that the uncertainty is very low compared to the mean values shown (The maximum deviation is 0.02, and 99 % of the standard deviation values are below 0.002), further demonstrating that those interpretation statistics are stable across samples.

For the data in Figure 4.6, these values are 0.02 and 0.004, respectively. Looking at the scales in both figures, these uncertainties are very small.

Overall, this indicates that although the predictive performance of the GBT is comparable to that of the U-Net, it relies on very different statistical patterns in the data. These patterns are more non-local and mostly unphysical so that the resulting model is expected to have less skill in extrapolating outside its training domain.

4.2.3. Online Stability Tests

In this section we will test the U-Nets ability to run stable in a coupled setting and consequently test their (global) extrapolation capabilities. We do not perform an offline extrapolation test with another data set since the hypothesized non-causality of the full U-Net would not show any negative impact in this test. For this reason we decided to couple the developed parameterizations back to the host (ICON) model and thus have a stronger generalization test. We first couple both the ablated (without precipitation tracer inputs) and the full U-Net to the ICON model and observe that the ablated U-Net shows improved stability compared to the full U-Net, when coupled globally. We also find that the ablated U-Net gives improved extreme precipitation predictions as opposed to the full U-Net, which fails to predict the precipitation distribution accurately.

Coupling data-driven parameterizations to GCMs is typically intricate and the stability of the developed schemes is very sensitive to e.g., changes in the training data set (Rasp 2020) or the inclusion of variables on specific levels and the choice of the loss function (Brenowitz and Bretherton 2018, 2019). Trial and error is often used to find stable schemes among the offline trained parameterizations (Wang et al. 2022b). Stability issues of coupled models have been observed, even for idealized setups such as aquaplanet simulations (Brenowitz et al. 2020; Gentine et al. 2018; Rasp et al. 2018; Yuval and O'Gorman 2020). Other studies, in which coupled ML schemes have used more realistic setups, were trained and coupled with superparameterized GCMs (Han et al. 2023; Iglesias-Suarez et al. 2024; Wang et al. 2022b). A technical advantage of training on these datasets is that a clear scale separation is artificially introduced and therefore the training targets for the ML algorithms are well defined. On the other hand, this scale separation influences the emergent dynamics and the embedded SRMs are themselves idealized as they are 2D models with a limited extent (Brenowitz et al. 2020; Pritchard et al. 2014).

Introducing a new parameterization into a GCM typically requires a retuning of the host model to e.g., adjust for current compensating biases in the interplay of various parameterization

schemes (Grundner et al. 2024). There are potentially many feedbacks when coupling a new scheme to the GCM which can quickly lead to unstable configurations or incorrect results. Furthermore, because there are considerable design differences between storm-resolving and coarse-resolution global climate models (Satoh et al. 2019), there could be distributional shifts between both types of model classes. Substantial distributional shifts have already been observed within the class of storm-resolving models (Mooers et al. 2023), so that ML parameterizations trained on data from a different storm-resolving model cannot be expected to learn the same relations. Also, by coarse-graining high-resolution fields, disturbances which can be represented on the coarse grid but not accurately advected by the coarse model can be introduced as noted by Watt-Meyer et al. (2024). To tackle this problem and to keep the coarse dynamics close to the coarsened high-resolution state, they nudged the coarse simulation to a coarse-grained high-resolution reference state continuously and achieved stable coupled runs (with ML-predicted tendencies for heat and moisture) for about 35 days with realistic boundary conditions.

Because of these issues and limitations we do not expect our models to show accurate online performance without some further modifications. Nevertheless, we tried to couple the U-Net models to the ICON model to test their stability and therefore our hypothesis about the extrapolation capabilities of the full and ablated U-Net. For this coupling we used the FTorch library (Cambridge-ICCS 2024) to load our models within ICON and to run them in inference mode during the time integration. Before the actual coupling we added a preprocess/postprocess layer to both NNs which normalize all the input variables to zero mean and unit variance and apply a corresponding inverse transformation for the output variables.

To test the stability of our developed ML parameterizations we created four different ICON configurations:

1. Ablated U-Net applied for all longitudes and latitudes
2. Full U-Net applied for all longitudes and latitudes
3. Ablated U-Net applied for all longitudes and only tropical latitudes
4. Full U-Net applied for all longitudes and only tropical latitudes

For configuration 1 and 2 the convection schemes have to extrapolate substantially as e.g., temperature, humidity, and also wind patterns differ considerably in the extratropics. Configurations 3 and 4 are applied closer to their training data set domain, i.e. the tropics. We apply the U-Nets between the Tropic of Capricorn (23.436 16°S) and the Tropic of Cancer (23.436 16°N) while the training domain is approximately defined between 10°S and 20°N as shown in Figure 3.2. Outside of the tropics the conventional mass-flux convection scheme is applied for these two configurations (3/4). For all coupled simulations (and the reference simulations), we use ICON in its version 2.6.4, with an `R2B5` ($\Delta x \approx 80$ km) horizontal grid and 47 vertical layers. Parameterized processes include radiation, cloud microphysics, orographic and non-orographic gravity wave drag, turbulence, and (ML-based) convection.

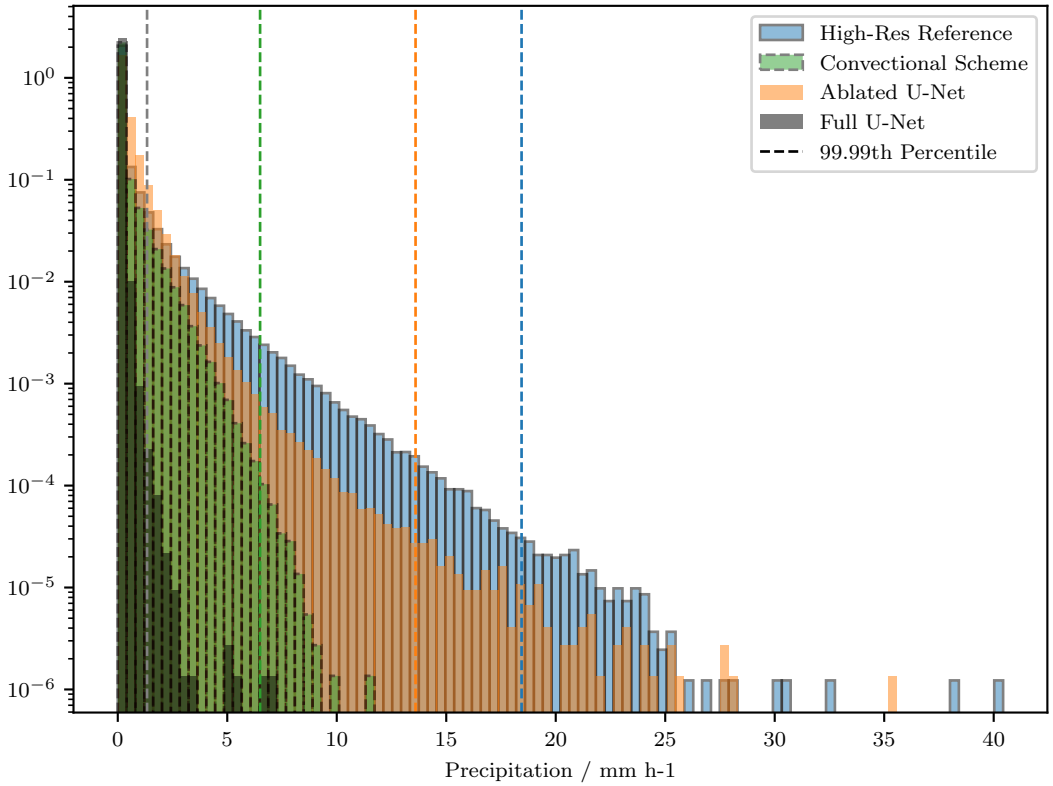

Figure 4.8.: The precipitation distributions of the first two weeks over the tropics for the three simulations starting on 01.02.1979 for the full U-Net (configuration 4) in grey/dark green, the conventional cumulus scheme in green, and the ablated U-Net (configuration 3) in orange. Also, the precipitation distribution for the high-resolution data set (NARVAL) is displayed in blue. The 99.99th percentiles of each data set are marked by dashed lines in the corresponding color. Adapted with permission from Heuer et al. (2024).

We initialized a simulation from interpolated IFS (Integrated Forecasting System) analysis data for the 01.01.1979 and ran the ICON model for one month. After this initialization phase we wrote out initial conditions for each day at 0000 UTC. With these initial conditions we started ten new runs with a length of half a year for each model configuration (from the 01.02.1979, 02.02.1979, . . . , 10.02.1979) to test the stability of the ML schemes. For columns with the ML scheme activated we applied the tendencies for heat, moisture, zonal and meridional wind which are derived by taking divergence of the ML-predicted fluxes instead of the ones derived by the conventional mass-flux scheme. Everywhere else, only the conventional convection parameterization of the ICON model was applied. No switch condition for the activation of our ML scheme was needed as we chose to add 10 % of non-convective columns to the training data set, as explained in Section 3.3, so that the U-Net learned when not to predict any convective fluxes. An alternative option would be to use the trigger condition from the conventional cumulus scheme, where convection is triggered for columns with moisture convergence, and some thresholds regarding humidity and buoyancy must be met (Möbis and Stevens 2012). We decided to not use such condition here but we could explore such methods in future work.

A first result of the online simulations is shown for the probability density function of precipitation in Figure 4.8 and for the spatial distribution of mean precipitation in Figure A.7. In Figure 4.8, the distribution of precipitation over the first two weeks of simulation over the tropics is displayed for the setup with the conventional cumulus scheme, the ablated U-Net (configuration 3), and the full U-Net (configuration 4). For both, configuration 3 and configuration 4, we set values of negative precipitation to zero. In future work this could be avoided by using an activation function with a non-negative codomain, like the `relu`-function, for precipitation. For the simulations shown here we set the large-scale precipitation to zero as said in Section 3.3.

The spatial distribution (monthly means) of precipitation over the region where we have a high-resolution reference can be seen in Figure A.7 in the supplementary information. The spatial mean precipitation patterns show that the coupled ablated U-Net results in a much more reasonable spatial distribution of precipitation than the full U-Net which heavily underestimates the mean precipitation. Compared to the high-resolution reference, the ablated U-Net produces a spatially more uniform precipitation distribution and has regions with too high mean precipitation. The conventional scheme shows a heavy land bias for the mean precipitation and shows too low precipitation values.

Figure 4.8 demonstrates the potential and added value of ML parameterizations as the precipitation distribution for the coarse model coupled with the ablated U-Net is much closer to the high-resolution (NARVAL) distribution than the reference simulation. For the full U-Net (configuration 4) we see an opposite effect: the distribution does show even less extreme values than the simulation with the conventional cumulus convection parameterization. This shows that the full U-Net, which heavily relies on the precipitation tracers (see Figures 4.5 and A.4), struggles to show good online performance. The reason lies in the hypothesized non-causal relations to the mentioned precipitation tracers. In coarse-grained (offline) data, precipitation is highly informative about convective events and further precipitation due to convective memory but as soon as the parameterization is coupled, the scheme struggles as the ML model itself has to predict some convective fluxes and precipitation in the first place.

The values for the 99.99th percentile further show the increased ability of the ablated U-Net to predict precipitation extremes more accurately and therefore the potential to reduce the common problem of GCMs to predict these extremes accurately (Stephens et al. 2010). These percentile values are 18.44 mm h^{-1} for the NARVAL data, 13.07 mm h^{-1} for the ablated U-Net, 6.67 mm h^{-1} for the reference simulation, and only 1.34 mm h^{-1} for the full U-Net.

Looking at the stability of the coupled simulations, Figure 4.9 displays the global mean surface temperature of all simulations of configuration 3 and 4 for 180 days. We can see that all simulations of configuration 3/4 are stable for the displayed period while the simulations with the full U-Net applied globally (config 2) very quickly become unstable, after about 6 to 18 hours. Configuration 1 (ablated U-Net coupled globally) is stable for the first day and simulations diverge only over the course of half a year as it can be seen for the orange lines in Figure 4.9. The simulations are stable for about 115 days on average with two simulations from these configurations staying stable for all 180 days. For the fully stable simulations, the

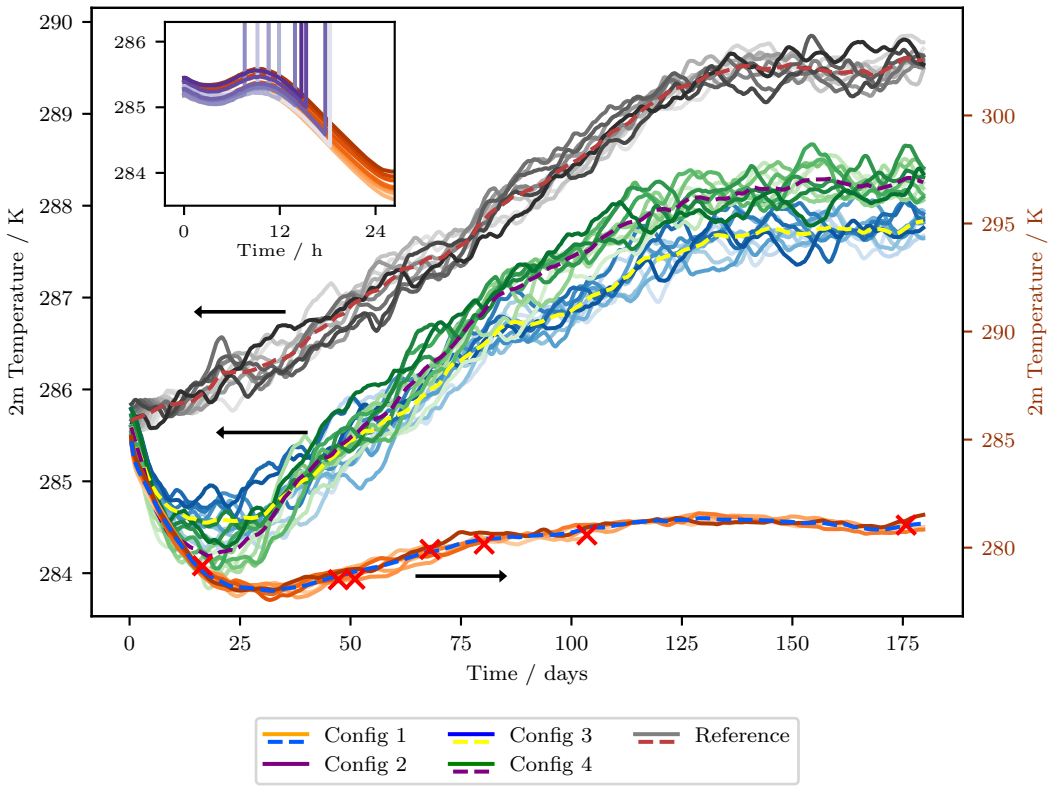

Figure 4.9.: The stability of the ablated vs. the full U-Net in form of a time series of the global mean air temperature on 2 m height over 180 days. For each defined configuration, the ten realizations are drawn in orange, purple, blue and green colors, respectively. Solely for the full U-Net coupled globally (Config 1), a second y-axis (also in orange) on the right side of the plot is introduced as this simulations shows a much higher reduction in 2m Temperature. To make this clearer, arrows are indicating the corresponding y-axis for each ensemble. An inset provides a close-up of the first 24 hours of the dynamics of configurations 1 and 2: the simulations with the full U-Net quickly become unstable. The data displayed in the inset has been saved with an output frequency of six minutes as opposed to the more stable simulations with an output frequency of six hours in the main plot. For all of the data except the inset, a rolling mean over 24 h was applied. Additionally, multi-model means over configurations 1, 3, 4, and the reference ensemble, respectively, are drawn as light green/yellow/violet/red-brown dashed lines. These colors are chosen as the complementary colors of the respective ensemble members and are marked in the legend as the second lower dashed line for each ensemble. For configuration 1, model blow-ups are marked by red crosses as to not obscure the other lines. Adapted with permission from Heuer et al. (2024).

surface temperature initially drops by about 1 K and then increases again to a higher value than the initial temperature. By looking at the full 180 days of time integration, the temperature for configuration 3/4 seems to equilibrate at about 287.8 K/288.2 K (~14.7 °C/15 °C) as seen in the figure. This is not unrealistic but the main point of this figure is to show the coupled stability for multiple months which is already three times the length of the training data set (two months). The global mean temperature of configuration 4 shows a similar trend compared to configuration 3 but becomes stable at slightly higher temperatures. The reason could lie in the fact that convection is much more infrequent for the full U-Net configuration as it can be

seen in Figure 4.8 and heat is therefore transported less efficiently to higher levels. Comparing these two fully stable simulations to the reference simulation mean in grey, we can see that there is an initialization shock (the mentioned initial temperature drop) (Bretherton et al. 2022). After this shock, the seasonal variation appears very similar in both magnitude and phase to the reference simulations. The initial shock indicates the off-set, that would have to be addressed by tuning, or nudging as in Watt-Meyer et al. (2024), as described earlier.

As all ensemble members of configuration 2 quickly diverge (as opposed to configuration 1, which is stable for minimally 16.5 days) we conclude that our hypothesis, that the full U-Net learned non-causal relationships, gains more support. However, the ablated U-Net configuration, does not guarantee stability when coupled globally.

To have a closer look at the dynamics we show the vertically integrated water vapor in Figure 4.10. A reference simulation with the conventional cumulus convection scheme is shown in the top row and the other rows are marked by their configuration number as defined above.

For the coupled full U-Net applied at all latitudes/longitudes we can only see one snapshot after six hours in Figure 4.10 because for the other dates the simulation has already diverged. For the snapshots after 4 days of simulation the structures with the ML parameterizations still look close to the reference simulation but there can already be seen some blurring effects in the tropics, especially over the ocean (e.g., over the Pacific). After a month of simulation configurations 1, 3, and 4 lost most of the structure in the tropics and instead there is a homogeneous high water vapor accumulation over these latitudes. This blurring effect is also displayed in the zonal mean and standard deviation plots in the last column. Especially for the ablated U-Net coupled globally (configuration 1), the standard deviation in the extratropics is very low. Furthermore, it is visible that for the ML coupled simulations, the mean water vapor path has a flatter peak compared to the reference and for the coupled full U-Net, the water vapor path has, additionally, a smaller magnitude in general. Note that for the ablated U-Net coupled globally (configuration 1), Figure 4.10 shows that the zonal mean water vapor path is less than zero for very high latitudes. This demonstrates the ML models failure to extrapolate to these latitudes, although, as most of the extratropical values still look reasonable and this configuration is stable compared to the full U-Net, this degree of extrapolation could also be considered unanticipated.

The blurring problem is a very common one for data-driven atmospheric models and can be related to the fact that ML models minimize a deterministic error and tend to predict some mean state rather than, possibly a more realistic, extreme state (Rasp et al. 2023). While this explanation cannot directly be transferred for the smoothing we see here, as we did not develop a fully data-driven atmospheric model, the used ML models are also incentivized to predict mean fluxes due to the used deterministic RMSE.

A similar effect has been observed by Kwa et al. (2023), by applying ML corrections to their coarse GCM they observed a reduction in tropical variability of precipitation. Alternatively, the existence of the observed blurring could be caused by the comparably low accuracy of U-Net at lower levels (see Figure 4.4) or the U-Net's failure to represent convection over steep

orography. Outside of the tropics, where the ML parameterization is not applied, there are still some structures, e.g., atmospheric rivers, visible in the configurations 3 and 4.

As we said before, there are many challenges to coupling an offline trained parameterization to a GCM and the results in Figure 4.10 show that, although many simulations run stably for a long time, there is still much room to improve the ML algorithms. Nevertheless, we were able to test the stability of our developed data-driven schemes. Both the ablated and the full U-Net support stable simulations when coupled only inside tropical latitudes. However, coupling the full U-Net, for which we hypothesized non-causal relations (see Figure A.4), leads to model blow-ups rather quickly when coupled globally, outside the training domain.

4.3. Conclusions and Discussion

In order to develop an ML-based parameterization for convection we first filtered, processed, and coarse-grained data from high-resolution simulations with explicit convection. To separate convection from other processes, we used a filtering method for convective conditions. That ensures that the ML models learn mostly convective fluxes. We then coarse-grained the high-resolution data to the target resolution and calculated the subgrid fluxes of the needed output quantities. The coarse-graining was performed without neglecting horizontal density fluctuations since we used data from a model with terrain following coordinates and the irreducible error increases if the model does not have the necessary input information. For the vertical coarse-graining we had to neglect some columns from the data set with especially steep orography. However, there are still many columns over heterogeneous terrain available and most trained models are able to run stable online. Nevertheless, future work could target including also these column and therefore profit from a orographically more diverse data set.

We found that the U-Net architecture is a very suitable machine learning model to parameterize convective subgrid fluxes, which is naturally a multiscale process. The U-Net outperformed other deep learning models by only a small margin judging by the R^2 metric. However, comparing the offline performance over a broad range of parameters, the error of the U-Net was consistently lower than the error of MLP, CNN, and ResNet architectures (Figure 4.2), this showed the structural advantage of the U-Net compared to the other models. A comparatively lower R^2 is achieved by most non-deep-learning models except for the Gradient Boosting Trees model. The linear models show a higher performance compared to the random forest and extra tree regression model. This could be related to the missing extrapolation capability of these tree based models, the effective feature selection of these regularized linear models, or, possibly, due to too heavy tuning to tropical convection. We will have to conduct more experiments in future research to train and test these models globally. Based on our offline evaluation we cannot claim that the tree-based models are not able to perform well online, therefore we plan to explore the online performance of the tree-based models by coupling them to ICON as well. The coupling of tree-based models to a GCM has been done successfully before by, e.g., Yuval and O'Gorman (2020, 2023) (although, in idealized aquaplanet settings). The GBT

model had a coefficient of determination of $R^2 \approx 0.84$ compared to the U-Net with $R^2 \approx 0.90$. Nonetheless, in a direct comparison between GBT and U-Net, the best performing non-deep learning and deep learning model, the U-Net had an advantage in almost all aspects. An exception to this is shown in Figure 3.4 by the R^2 value for a few levels for ice, snow, and cloud water tracers. For snow and ice these exceptions occurred in the lower levels and for cloud liquid water mainly in the higher ones, where the respective tracer species are rarely observed / have a very low concentration. This demonstrates the advantage of the lower complexity tree-based method for sparse data or rather for regions where an interpolation based on few relevant samples is needed. For the other levels and also for the predicted 2D fields, such as convective precipitation, we noticed a clear benefit of using the U-Net architecture. We do not claim exhaustiveness in the choice of ML models/NN architectures, the parameterization could profit from the combination of specific architectures benchmarked here, such as ResNets and CNNs, or other more advanced model such as recurrent NNs or Transformers (with the height as time/sequence dimension).

While the U-Net shows a high skill in parameterizing multiscale convection, we did not empirically test the multiscale representation of the NNs. Future research could target testing these multiscale properties by e.g., ablating the most compressed layers and looking at the decrease in accuracy for deep convection or testing the ability of the model to work on scaled in/outputs. Furthermore, other modifications, like dilated convolutions (Yu and Koltun 2015), could be tried to enhance the multiscale processing of the U-Net.

To get some insight into what the model exactly learned during training we applied the SHAP framework and first calculated feature importances. These revealed that the U-Net model focuses strongly on the precipitating tracer species rain and snow as input variables. Here, the SHAP values exposed that the model learned non-causal relations between convective subgrid fluxes and convective precipitation. This was also seen in the figure showing the weighted SHAP values (Figure A.4), as particularly the rain tracers showed heavy non-local influences on subgrid fluxes for liquid/ice water static energy, rain, cloud liquid, and water vapor tracers. For comparison, the weighted SHAP values for the MLP model can be seen in Figure A.5. Similar non-causal connections to precipitating tracer species can be observed in that figure and, in fact, we found that for all deep learning models with a full input, the precipitating tracer species show the highest (shap value-based) feature importance assignment. As a result we performed the same analysis on an ablated model without water species. A potential solution to be investigated in a future study would be to restrict the model to learn causal relationships as in Iglesias-Suarez et al. (2024). Another approach to improve the predictions of subgrid momentum fluxes specifically would be to model the degree of small scale convective organization (Shamekh et al. 2023). For higher stability in coupled simulations of the developed ML-based multi scale parameterization to a GCM it will be advantageous to use a global training data set. Convectively active regions in the extratropics would be especially important to include, e.g., regions where frontal systems and extratropical cyclones are common, extratropical monsoon regions, or locations with marine stratocumulus clouds.

Furthermore, it would be important that ML models learn the distributions corresponding to, e.g., the arctic climates so that out-of-distribution predictions can be avoided in high latitudes.

By looking at the weighted SHAP values we found that the ablated version of the U-Net was more physical and learned physically explainable connections between coarse-scale variables and subgrid fluxes. For example, there were patterns indicating local upwards transport of horizontal momentum and energy, moisture convergence, and the interaction between wind shear and mesoscale convective systems. This strengthens trust in the model as it can be expected to extrapolate better to data outside its training domain. However, many interpretations of the weighted SHAP value matrices, besides some objective features like locality, are rather subjective (e.g., mesocale convergence) and should be generally regarded as one out of many tools to build up trust in the models. The weighted SHAP values for the GBT model were not physically interpretable as they showed very scattered results and close to no coherent patterns. We applied a different explainer class to test the robustness of this outcome and saw consistent results. To investigate this further, we did the same analysis for the Random Forest as this model has been used in other studies before. Here, the weighted SHAP values were similarly scattered as for the GBT model. This result shows that seemingly well performing models (judging by e.g., R^2) can in fact rely on non-causal correlations in the data, achieving good results for the "wrong reasons". Therefore, these models are most likely not suited for the coupling to a GCM. The emergence of these non-causal relationships and possible methods of prevention, besides ablation, should be investigated further in future research.

In the section on online stability tests we coupled the ablated and full U-Net to the ICON model and showed that, when coupled globally, the hypothesized non-causal connections indeed lead to instability within a day for the full U-Net; as opposed to the ablated U-Net which support stable simulations for minimally 16 days (and on average, 115 days). For the ablated U-Net (and both U-Net parameterizations applied only in the tropics) we found stable simulations for at least 180 days. By coupling the ablated U-Net to the ICON model, we could show that the ML model is able to predict precipitation extremes more accurately online (see Figure 4.8) in contrast to the conventional parameterization and the full U-Net. The stable simulations are showing e.g., smoothing biases already after some weeks. Tracing back the specific output variables responsible for this smoothing bias would be significant to understanding and minimizing this effect in future research. An approach using a stochastic ML parameterization could mitigate the smoothing bias, possibly, related to the usage of the RMSE mentioned in Section 4.2.3. However, we did not expect perfect results because of distributional shifts between the training data set and the variable states of the coarse simulation. Furthermore, as our process separation is not perfect and at least some momentum fluxes from gravity waves will have an impact on the dynamics, we will do some further tests in the future e.g., without a parameterization for non-orographic gravity wave drag. Another possible approach for future research would be to build a combined parameterization for convection and microphysics to more accurately represent their interaction and the influence of convective updrafts on microphysics. For further improvement of the coupled model results

it might be necessary to train the models on a global data set, use climate-invariant variables (Beucler et al. 2024), or work on more physically constrained architectures (Beucler et al. 2023). With more physically constrained and robust ML parameterizations, an extensive validation against a range of climatic conditions to ensure that any improvements in parameterization translate to more accurate climate representations would be necessary.

Our study leads to the conclusion that interpretability/explainability of ML algorithms is important to investigate potentially non-physical mechanisms. Furthermore, we conclude that the U-Net is the best choice of the examined model classes as it is very accurate, not too complex, and its predictions can be explained physically after domain knowledge was applied to ablate spurious correlations. This advantage over other ML-model classes likely comes from the ability of the U-Net to capture multiscale phenomena like convection. In the future, we will expand our work by training ML models on global high-resolution data for which we ensure that input variables and fluxes are output after the dynamical core or respectively, after parameterizations for processes which are neither resolved for the high-resolution simulation nor the coarse scale, e.g., radiation. By doing this, we will avoid distributional shifts between the coarse-grained data set and the coarse simulations.

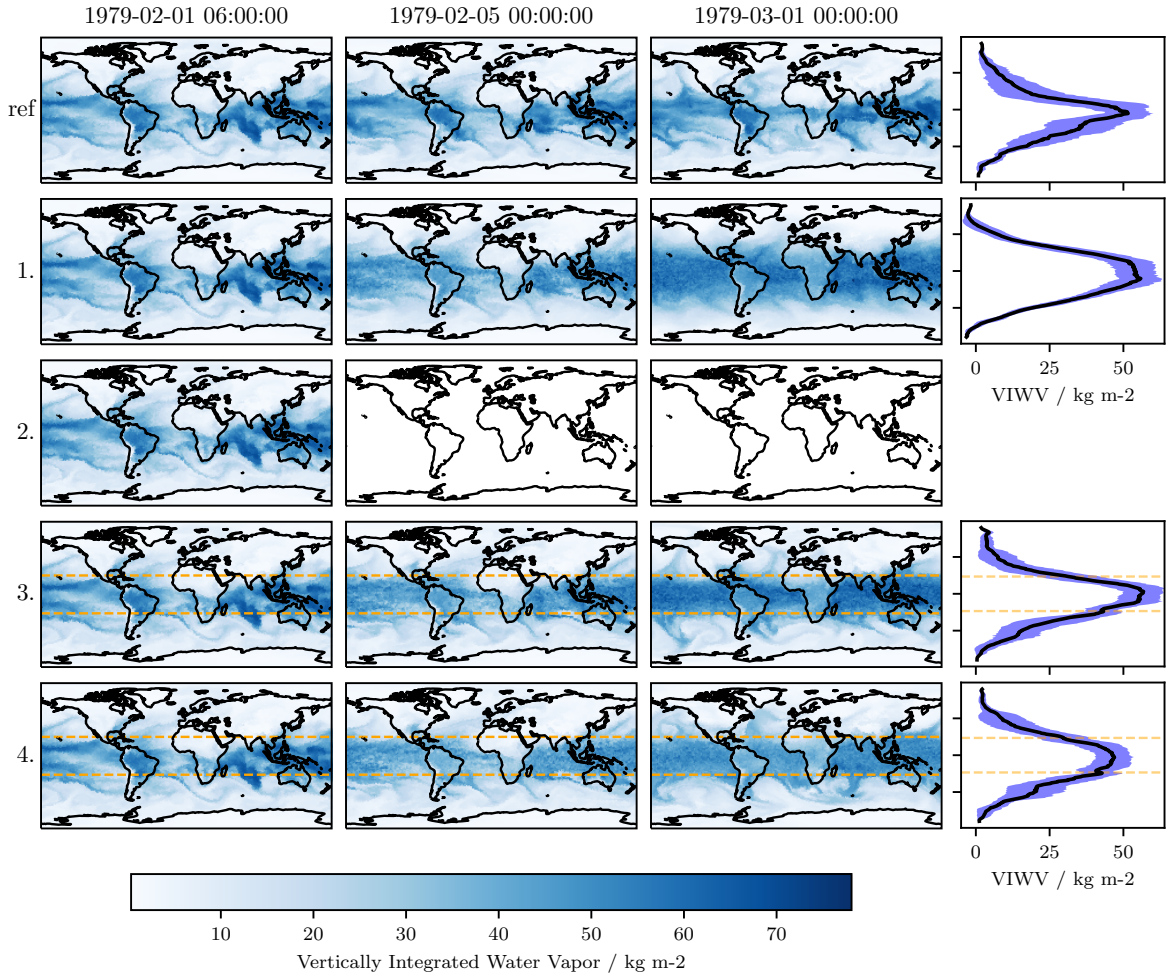

Figure 4.10.: The vertically integrated water vapor for three simulation snapshot with convection parameterized with the conventional physical cumulus convection scheme of the ICON model as a reference (ref), 1.) by the ablated U-Net, 2.) by the full U-Net, 3.) by the ablated U-Net applied only in the tropics, 4.) by the full U-Net applied only in the tropics. For row 3.) and 4.) the domain where the ML schemes are applied are marked by orange dashed lines. The last column shows the zonal mean and standard deviation of the vertically integrated Water Vapor (VIWV) for the last shown date (1979-03-01) of every configuration except the unstable one. The y-axis corresponds here to the latitudes of the corresponding row. Adapted with permission from Heuer et al. (2024).

5. Beyond the Training Data: Confidence-Guided Mixing of Parameterizations in a Hybrid AI-Climate Model

5.1. Introduction

This chapter was already published as a preprint in Heuer et al. (2025). As indicated in Section 1.2, the author of this thesis created all the content, including text, figures, and tables, that is presented from this publication and implemented the code[1] to reproduce this study. All three key science questions (see Section 1.1) are addressed by this chapter.

Mass-flux parameterization schemes, which represent the vertical transport of energy, water, and momentum in convective up- and downdrafts as a function of environmental conditions (Arakawa and Schubert 1974; Tiedtke 1989), remain the de facto standard for parameterizing deep convection in modern ESMs. Such parameterizations can however introduce substantial biases into climate projections (Christopoulos and Schneider 2021; Judt 2018; Lee et al. 2021; Stevens et al. 2019b) because they are often based on empirical relationships and simplifying assumptions.

Recent years have seen a surge of ML-based parameterizations for deep convection and cloud physics (Gentine et al. 2018; Heuer et al. 2024; Yuval and O'Gorman 2020, 2023). Training ML-based schemes on coarse-grained high-resolution data and implementing them in conventional ESMs promises to reduce long-standing biases in coarse-scale global simulations. To design a suitable training dataset, the choice of the coarse-graining and filtering operator is however critical and not uniquely defined (Brenowitz et al. 2020; Ross et al. 2023). Furthermore, coarse-graining storm-resolving ICON data does not yield a clean separation of convective versus other subgrid processes (Heuer et al. 2024). Additionally, generating global storm-resolving training data is extremely expensive (Satoh et al. 2019) and most available storm-resolving ICON datasets are not ideal as a ground truth because they do not offer an appropriate temporal output frequency (sub-hourly), global coverage, or do not include the needed variables for training: DYAMOND (Stevens et al. 2019b) or NextGEMS (Koldunov et al. 2023) provide only 3-hourly 3D fields and at best 15-min 2D surface variables; NARVAL (Klocke et al. 2017; Stevens

[1] published under https://github.com/EyringMLClimateGroup/heuer25james_ml_convection_climsim (last access: 14.10.2025) and preserved (helgehr 2025)

et al. 2019a) is confined to the tropical Atlantic with hourly output. Challenges related to using complex high-resolution training data are illustrated in our previous work, Heuer et al. (2024), where an ML model for deep convection was trained on coarse-grained and filtered two-month-long high-resolution tropical data. This yielded promising online results such as an improved representation of precipitation extremes, but also introduced heavy blurring and biases in variables such as column water vapor or temperature.

Furthermore, whereas previously developed ML parameterizations have shown success in modeling the subgrid convective fluxes and convective precipitation, stability issues remain very common, even in idealized aquaplanet setups (Brenowitz and Bretherton 2018, 2019; Brenowitz et al. 2020; Gentine et al. 2018; Lin et al. 2025; Rasp et al. 2018; Yuval and O'Gorman 2020). Hybrid ML–physics climate models have yet to demonstrate stable, accurate simulations suitable for operational use; emerging real-geography runs are still too short (Watt-Meyer et al. 2024) or too coarse (Hu et al. 2025).

In an attempt to mitigate the discussed challenges of training ML models on global storm-resolving data directly, we use the ClimSim dataset (Yu et al. 2025), generated with the E3SM-MMF (E3SM Project 2018). In this superparameterized setup (Hannah et al. 2020), 2D SRMs with periodic boundaries are embedded in each coarse atmospheric column, replacing conventional subgrid parameterizations. ClimSim pairs coarse-scale atmospheric states (inputs) with tendencies derived from the embedded SRMs (targets), providing a well-defined scale separation between resolved coarse dynamics and unresolved physics. This reduces ad hoc choices in coarse-graining and process separation when training ML models. In addition, a 2024 challenge on Kaggle, an open ML competition platform, built around ClimSim, attracted over 690 final submissions (Lin et al. 2024), yielding strong baselines and architectures we leverage here.

In this proof-of-concept, we leverage these developments to create a new ML-based parameterization of convection for the ICON model (Giorgetta et al. 2018; Zängl et al. 2015) with a horizontal resolution of ~160 km × ~160 km, trained on the ClimSim dataset. Our ML approach draws inspiration from models developed in the Kaggle competition, in which vertically recurrent NNs (Ukkonen and Chantry 2025), such as BiLSTM architectures, emerged as competitive contenders for predicting subgrid-scale tendencies from large-scale inputs. We additionally implement a physically informed loss function encouraging the trained networks to adhere to conservation laws and to discourage non-conservative sources and sinks in single-column predictions. A key modification, inspired by the first-place winner "greySnow" of the Kaggle competition, is the incorporation of a CL. This adds a second prediction head that estimates the loss for all targets, effectively quantifying model uncertainty. Using this confidence metric during online inference, we mix ML predictions with the conventional convection scheme when the ML scheme is uncertain, thereby improving overall performance. The approach is similar to the novelty-detection method of Sanford et al. (2023) or the "compound parameterization" proposed by Krasnopolsky et al. (2008) and used in Song et al. (2021), identifying and responding to out-of-distribution or uncertain conditions during inference. Rather than applying ML corrections unconditionally, we use the confidence metric as a proxy

for uncertainty to detect potential extrapolation beyond the training domain. By avoiding extrapolation and applying ML corrections only in specific regions of input space, this method prevents unphysical or biased outputs and enhances stability and reliability. With this work, we build upon previous studies demonstrating ML-based parameterizations in ICON (Grundner et al. 2022, 2024; Heuer et al. 2024; Sarauer et al. 2025).

This chapter is organized as follows: In Section 5.2, we evaluate one-year-long coupled simulations, analyzing climate statistics and the physical behavior of the ML parameterization to gain process-level insights. As a comprehensive validation, we conduct historical AMIP-type simulations with prescribed SSTs, sea-ice concentrations, and greenhouse gas concentrations. Finally, Section 5.3 discusses the key findings and concludes the study.

5.2. Results

This section first compares ICON simulations coupled to the various ML schemes developed in this study and the conventional Tiedtke scheme with observations. These comparisons use ESMValTool (Andela et al. 2025; Righi et al. 2020) to calculate evaluation metrics. We then examine the conservation properties of the developed models and investigate under which conditions they exhibit higher or lower confidence. Additionally, we explore why the mixed model demonstrates better skill than both the Tiedtke and pure ML models to ensure the improvements to convective physics are interpretable. Finally, this section concludes with an application of the developed schemes in 20-year-long AMIP-style simulations.

5.2.1. Benchmarking with Observations

To evaluate the online performance of various ML models, we systematically varied the weight of the physics-informed loss term, α, during training, with $\alpha \in \{0, 0.01, 0.1, 0.5, 0.9\}$. The offline coefficients of determination on the test set for the models with $\alpha \leq 0.5$ are approximately $R^2 \approx 0.89$ and $R^2 = 0.631$ for $\alpha = 0.9$ as seen in Table B.2. Furthermore, we explored the impact of adjustments to the percentile parameters p_0 and p_1, which generated diverse ML weight configurations, λ. Specifically, we tested p_1 values within the range of 20 % to 90 %, while p_0 was varied between 10 % and p_1. Additionally, we evaluated a model without the proposed mixing mechanism and no physics-informed loss terms ($\alpha = 0$), referred to as the "pure ML" model, to establish a further baseline for comparison. The simulations were run in an AMIP-style setup over an entire year starting on January 1st 2010. First, we will evaluate the performance of the ML-based schemes on some key climate metrics mainly related to water vapor and precipitation as the representation of water in the atmosphere is crucial for improving current climate models (Stevens and Bony 2013).

Figure 5.1 shows the performance of various model configurations evaluated by four different online metrics using ESMValTool. The conventional Tiedtke scheme is located near the Pareto front in panel (a) and on the Pareto front for (b). This is not surprising, as the ICON model has been tuned to perform well when used with the default Tiedtke convection scheme.

Nevertheless, many coupled ML schemes lie along the Pareto front and we could expect even better results if ICON was calibrated with these schemes, which is not feasible for all of them. In panel (a), we find a model with an α parameter of 0.5, showing an increase of $\Delta R^2 \approx 0.015$ relative to the Tiedtke scheme in both metrics. Interestingly, some schemes outperform the Tiedtke scheme by a large margin with respect to one metric but have a lower skill in another metric. For example, there is a model with $\alpha = 0$ having a precipitation R^2 increase of over ~ 0.12 compared to Tiedtke and a scheme with $\alpha = 0.9$ showing a column water vapor (CWV) R^2 increase of ~ 0.25. On panel (b), a clearer ordering of the α parameter with respect to the two metrics is observed. Furthermore, panel (b) demonstrates that there exist ML schemes outperforming the Tiedtke scheme by ~ 0.075 in near-surface (2 m) air temperature R^2 and $\sim 0.12\,\mathrm{mm\,d^{-1}}$ RMSE of the zonal mean precipitation.

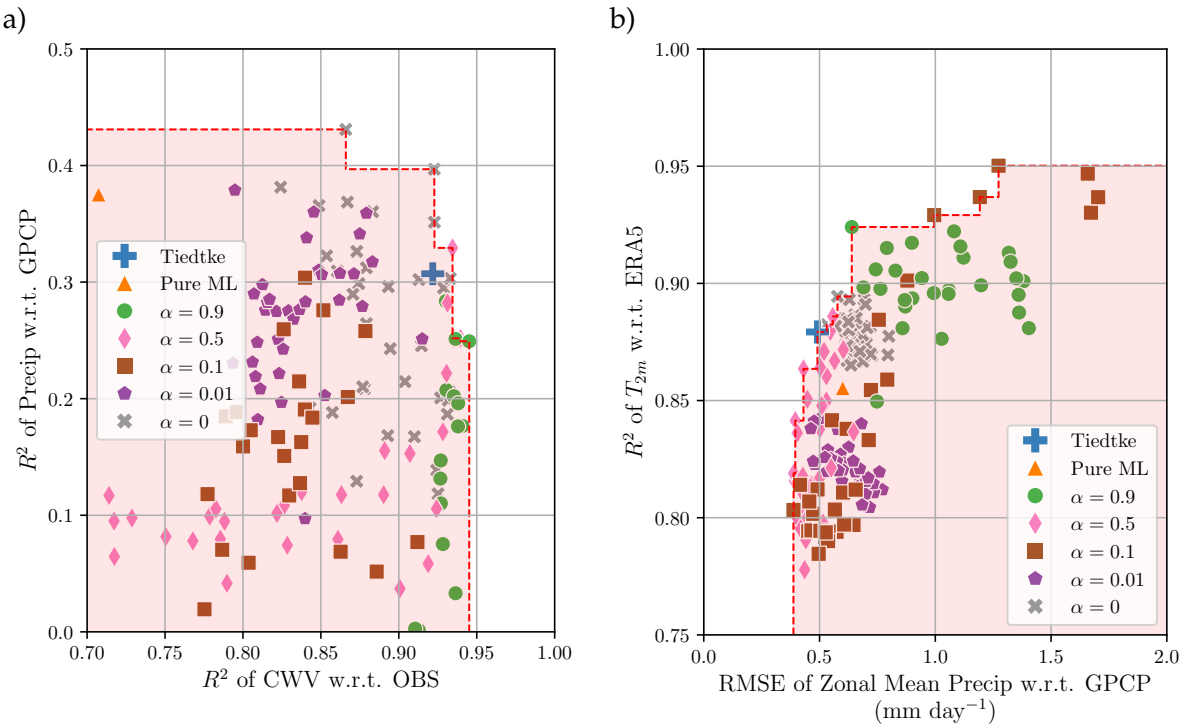

Figure 5.1.: Evaluation scores for coupled ICON runs, each dot represents a one-year long coupled ICON run at a horizontal resolution of 158 km × 158 km. The runs are colored according to their physics-informed loss weight α for the coupled ML schemes and the conventional Tiedtke scheme is colored in blue. Within each coloring group, the models have different values for p_0 and p_1. Panel (a) shows the spatial R^2 score of precipitation with respect to the observational dataset GPCP versus the R^2 score of CWV with respect to the mean of multiple observation sets as explained in Section 3.5.3. Panel (b) displays the R^2 score of near-surface (2 m) air temperature with respect to ERA5 versus the RMSE of zonal mean precipitation with respect to GPCP. In both panels, the Pareto front between the two skill metrics is marked with a dashed red line. Adapted with permission from Heuer et al. (2025).

The mixed models are named in the format "Mixed:p_0-p_1_$x\alpha$", with x indicating the value of the physics-informed loss weight α. Models with $\alpha = 0.1$ show the least error with respect to

zonal precipitation of the observations and are used for further analysis. For ease of notation, we will therefore leave out the α parameter in the naming of the model whenever $\alpha = 0.1$.

We next analyze the representation of precipitation in the various models by looking at zonal means of annual surface precipitation (Figure 5.2). The Tiedtke scheme significantly underestimates the peak in mean precipitation (Figure 5.2 (a)). The pure ML scheme exhibits a stronger peak, although it remains lower than the GPCP reference. The mixed scheme yields values slightly below the pure ML scheme, yet it outperforms the Tiedtke model. The displayed Mixed:10-60 scheme represents a model "tuned" to observations as it shows the least RMSE of the tested model with respect to zonal mean precipitation of GPCP. Notably, both the Tiedtke and pure ML schemes clearly display a signature of a double ITCZ in the sense that they show a pronounced second precipitation peak in the Southern Hemisphere. The double ITCZ is however substantially less pronounced in the mixed scheme and more closely resembles the observational reference. In the high latitudes all schemes exhibit a similar behavior. Overall, the mean absolute error with respect to the GPCP zonal mean precipitation is ~0.39 mm d^{-1} for the Tiedtke scheme, ~0.45 mm d^{-1} for the pure ML scheme, and ~0.3 mm d^{-1} for the mixed scheme.

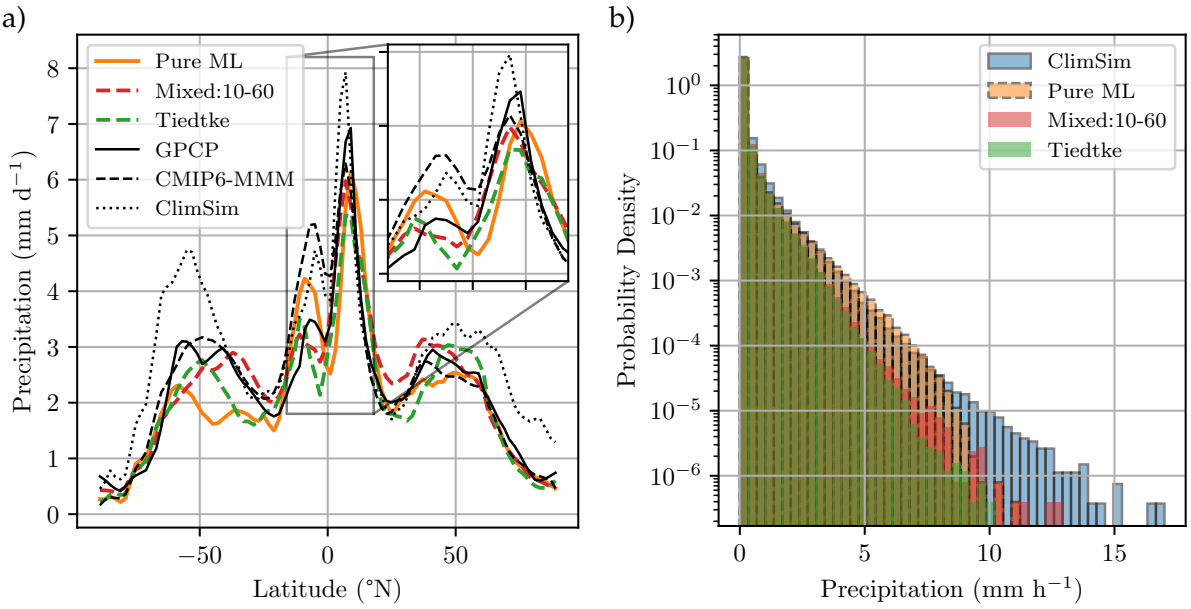

Figure 5.2.: Zonal mean precipitation in one-year-long runs (a) and precipitation distribution (b) for the pure ML scheme, the Tiedtke scheme, a mixed scheme (Mixed:10-60), and references; GPCP observations, CMIP6 multi-model mean (MMM), and ClimSim for the mean precipitation, and ClimSim for the precipitation extremes. Adapted with permission from Heuer et al. (2025).

To investigate the double ITCZ bias more quantitatively, we use the tropical precipitation asymmetry index A_P (Hwang and Frierson 2013) and the equatorial precipitation index E_P (Adam et al. 2016). The tropical precipitation asymmetry index quantifies the asymmetry of tropical precipitation, with positive values indicating higher precipitation in the northern (0° – 20 °N) tropical hemisphere $\bar{P}_{0-20\,\mathrm{N}}$ vs. the southern (20 °S – 0°) tropical hemisphere $\bar{P}_{20\,\mathrm{S}-0}$ (and vice versa for negative values):

$$A_P = \frac{\bar{P}_{0\text{–}20\,\mathrm{N}} - \bar{P}_{20\,\mathrm{S}\text{–}0}}{\bar{P}_{20\,\mathrm{S}\text{–}20\,\mathrm{N}}}. \tag{5.1}$$

The equatorial precipitation index represents the symmetric component of tropical precipitation by relating the mean precipitation within 2 °S - 2 °N, $\bar{P}_{2\,\mathrm{S}\text{–}2\,\mathrm{N}}$, to the mean precipitation estimated between the tropics, $\bar{P}_{20\,\mathrm{S}\text{–}20\,\mathrm{N}}$:

$$E_P = \frac{\bar{P}_{2\,\mathrm{S}\text{–}2\,\mathrm{N}}}{\bar{P}_{20\,\mathrm{S}\text{–}20\,\mathrm{N}}}. \tag{5.2}$$

The respective biases are defined as the index for a model run minus the index evaluated for the observations.

Data	A_P	E_P	A_P Bias	E_P Bias	RMSE (mm/d)
GPCP	0.454	0.920	-	-	-
Tiedtke	0.417	0.848	-0.037	-0.072	0.491
Pure ML	0.253	0.716	-0.201	-0.204	0.600
Mixed:10-60	0.451	0.911	**-0.003**	**-0.009**	**0.387**
ClimSim	0.268	0.973	-0.186	0.053	0.884
CMIP6-MMM	0.060	1.037	-0.394	0.117	0.525

Table 5.1.: The tropical precipitation asymmetry index A_P and the equatorial precipitation index E_P, and their biases, as well as the RMSE, with respect to GPCP for the data shown in Figure 5.2 (a). Adapted with permission from Heuer et al. (2025).

As Table 5.1 shows, the double ITCZ bias is lowest for the mixed model while the Tiedtke and pure ML models have significantly higher biases. The informative value of these indices is rather limited due to their simplicity, but they give a further indication that the mixed model captures the zonal precipitation distribution well. The mixed model also displays the lowest error as indicated by the RMSE of the curve of zonal mean precipitation with respect to the GPCP curve (Table 5.1).

The distributions of daily precipitation values (Panel (b) of Figure 5.2) reveal notable differences between the various datasets. The ClimSim dataset stands out with the highest extreme precipitation values, which is expected given that it is based on the MMF data. In contrast, the Tiedtke scheme underestimates precipitation extremes compared to ClimSim and exhibits an overabundance of minor precipitation events, a phenomenon commonly known as the "drizzle problem" (Stephens et al. 2010; Wang et al. 2016). The ML scheme presents a distribution more akin to ClimSim but appears to slightly overemphasize mid-level precipitation events, specifically those ranging from 2 mm h^{-1} to 9 mm h^{-1}. Meanwhile, the mixed scheme offers a balance between low and high precipitation events, showcasing slightly more heavy precipitation events than the Tiedtke scheme, although still falling short of replicating the reference data provided by ClimSim.

As a comparison to Figure 14 of Heuer et al. (2024), we also visualize three snapshots of the column water vapor for some of the tested configurations. This is shown in Appendix B.3 (Figure B.2). In Heuer et al. (2024) a significant smoothing for the stable simulations was visible

after 4 days and after one month there were no structures visible in the troposphere anymore. Figure B.2 clearly shows that this is improved substantially as clear structures are still visible for all configurations after a month and even a year of integration.

5.2.2. Advantages of Physics-Informed Loss via Conservation Laws

To assess the fidelity of the learned physics, we monitor the mean absolute enthalpy residual, i.e., the mean absolute value of Equation (3.9), throughout the simulations, alongside the global mean ML weight, $\langle\lambda\rangle$ (Figure 5.3). As expected, the conventional Tiedtke scheme demonstrates perfect enthalpy conservation. Conversely, the pure ML scheme exhibits the largest residuals as it has learned no physical conservation laws during training and also does not mix in any conservative Tiedtke output profiles. Notably, the NNs enforcing soft constraints on enthalpy, mass, and momentum conservation, exhibit intermediate behavior. This demonstrates that the proposed hybrid approach effectively constrains the ML predictions, resulting in improved physical consistency compared to a purely data-driven model, which is particularly relevant for long-term integrations.

5.2.3. Process understanding: Why is the mixed model better than both the Tiedtke and pure ML model?

In this section, we analyze the mixed scheme across environmental regimes defined by geography (latitude), CWV, and lower tropospheric stability (LTS). Our goals are to (i) explain why the mixed scheme outperforms both Tiedtke and pure ML, (ii) identify regimes of high/low model confidence and its spatial structure, and (iii) characterize conditional mean heating and moistening profiles as functions of CWV and LTS. These analyses provide process-level insight into the hybrid model's strengths, demonstrate improved precipitation skill, and clarify how convective processes interact with the large-scale climate as constrained by observational products.

First, we investigate the spatial distribution of the average weight, $\langle\lambda\rangle$, for the Mixed:10-60 model with $\alpha = 0.1$ (Figure 5.4). The average ML weight is generally higher over land than over oceans, reflecting greater confidence in ML predictions in continental environments. Furthermore, the model exhibits increased confidence in high-latitude regions compared to the tropics. In the tropics, where convective activity is abundant, the model's confidence is reduced, likely due to inherent variability in this region. This fits the observation in Figure 5.5 that the ML models' confidence decreases with the magnitude of the column water vapor in the column as higher magnitudes of water vapor are expected in the tropics. Importantly, regions with complex orography – including the Himalayas, Andes, Ethiopian Highlands, and Rocky Mountains – tend to exhibit lower model confidence, even without explicitly providing orographic information to the ML models.

For comparison, the spatial distribution of the average ML weight is shown for two more models in Figure B.3. The patterns are very similar, but the overall ML weight increases with higher p_1 values as expected.

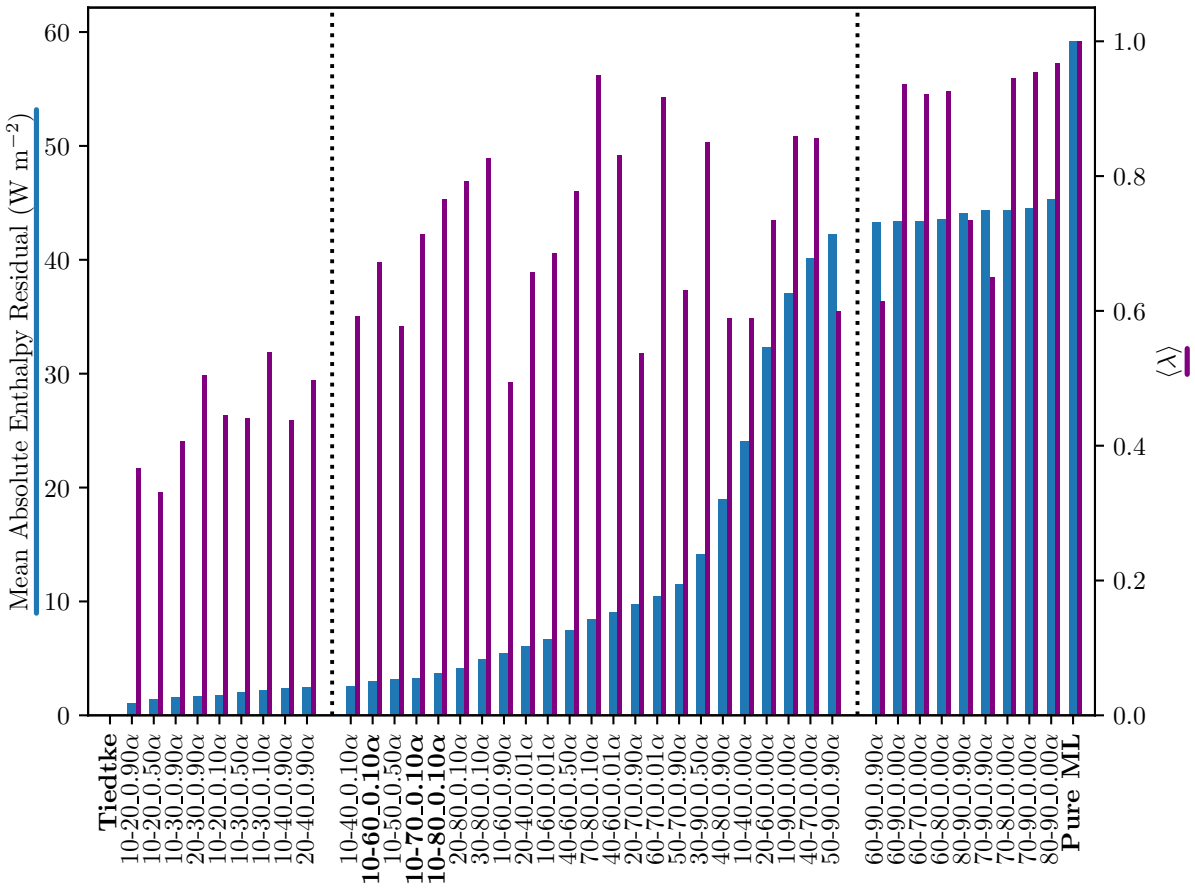

Figure 5.3.: Mean absolute enthalpy residual (blue, left axis) and average ML weight λ during the one-year long online integration (purple, right axis) for a selection of tested models. The ten most-conserving (left in the plot) and least-conserving (right) models in terms of enthalpy conservation are displayed. In between the black dotted lines every 8$^{\text{th}}$ model is displayed so that the figure is still readable. Additionally, models which are used for a deeper analysis in this section are marked by bold labels. Adapted with permission from Heuer et al. (2025).

To understand under which conditions the ML-based schemes predict convective precipitation, Figure 5.5 shows the conditionally averaged convective precipitation and average ML weights $\langle\lambda\rangle$ predicted by different schemes as a function of cumulative CWV and LTS, defined as

$$\text{LTS} = \theta_{\sim 700\,\text{hPa}} - T_{\text{sfc}}, \tag{5.3}$$

with the potential temperature θ at approximately 700 hPa and the surface temperature T_{sfc}. Low values of LTS indicate potential for deep convection due to conditionally unstable conditions in the lower troposphere (Brenowitz et al. 2020).

Panel (a) of Figure 5.5 reveals that the curves show comparable behaviors, especially among all mixed models, similarly to panels (b) and (c). Notably, the mixed models and the Tiedtke show a sharp pickup of precipitation around 50 mm to 60 mm globally, similar to the critical value of 66 mm reported for tropical environments in Holloway and Neelin (2009). The Tiedtke

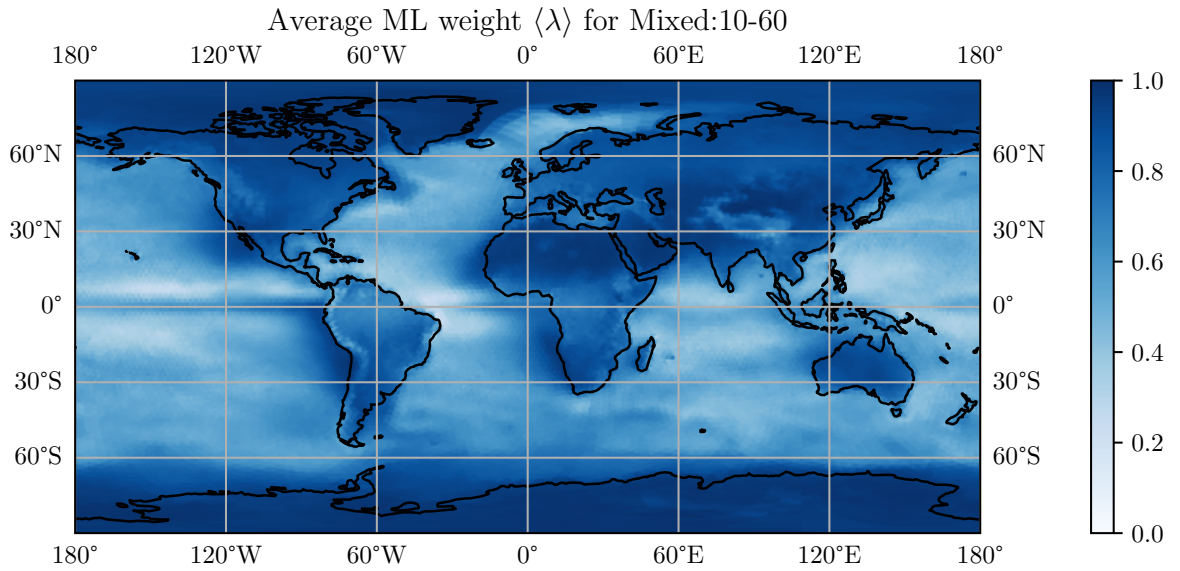

Figure 5.4.: The spatial distribution of the temporally-averaged ML weight $\langle\lambda\rangle$ over one year of simulation for the Mixed:10-60 model with a physics-informed weight $\alpha = 0.1$. The overall time averaged ML weight was $\langle\lambda\rangle \approx 0.67$ for the coupled run. Adapted with permission from Heuer et al. (2025).

scheme robustly shows the lowest precipitation values for all CWV conditions, consistent with Figure 5.2 (b). In contrast, the pure ML model exhibits relatively low precipitation for low CWV but high precipitation for mid-level CWV values. For very high CWV values, the schemes show slightly different behavior, although it is notable that this region contains very few samples. The decreasing ML model confidence (hence increasing λ) observed as CWV is increased therefore results from both the scarcity of training samples and the large inherent variability associated with convective processes in this region of the CWV space (Bretherton et al. 2004; Jones et al. 2004; Sukovich et al. 2014).

In panel (b) of Figure 5.5, the mixed models vary more smoothly with LTS than either Tiedtke or the pure ML models, which show discontinuities. Tiedtke also shuts down convection quickly at high LTS, likely missing cases where large-scale forcing (e.g., mesoscale convective systems or at higher latitudes) can trigger convection under relatively stable conditions. This helps explain why mixed schemes that place more weight on the ML component at high latitudes perform best, e.g., the 10–60 mixed model depicted in Figure 5.4. As expected, convective precipitation generally increases with decreasing stability (decreasing LTS). The Tiedtke scheme shows a sudden decrease in precipitation for very low LTS values, although it is worth noting that this region contains very few samples. The ML weight, i.e. $\langle\lambda\rangle$, of the models initially exhibits a modest increase (or even a slight decrease) as stability increases, but then rises more sharply until an LTS of 25 K is reached, after which it levels off and remains almost constant close to 1 under more stable conditions. This trend is reasonable because convective precipitation is expected to be low under very stable atmospheric conditions and more intense and difficult to predict for unstable environments.

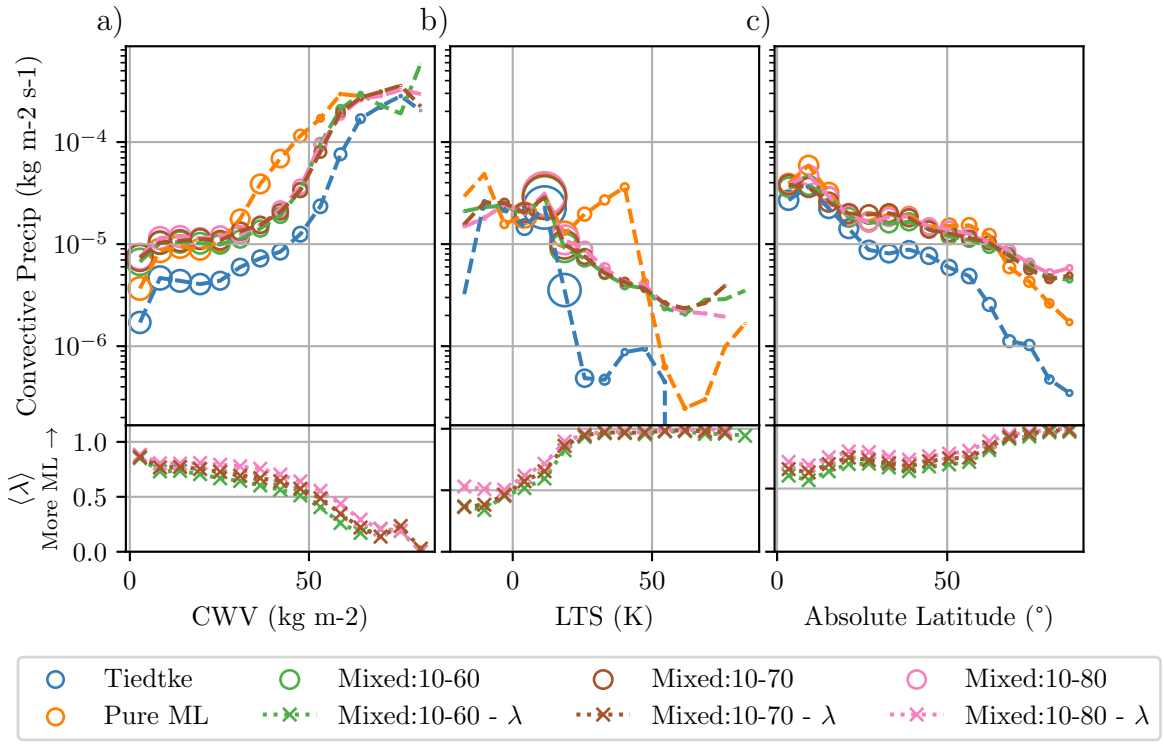

Figure 5.5.: Conditionally averaged convective precipitation (top row) and average ML weight $\langle\lambda\rangle$ (lower row) as a function of CWV (a), LTS (b), and absolute latitude (c). Circles represent the convective precipitation (circle sizes indicate the number of samples in the respective region) and crosses the average ML weight $\langle\lambda\rangle$. All plots within one row share the same y-axis scale. Adapted with permission from Heuer et al. (2025).

The convective precipitation decreases with increasing latitude (Panel (c)), as expected. In contrast, the ML weight increases with absolute latitude, reaching values close to 1 for latitudes exceeding 80°, consistent with the patterns observed in Figure 5.4. Similarly to Panel (a), the Tiedtke scheme demonstrates the lowest convective precipitation for almost all data points while the pure ML model and also the mixed models, predict relatively high values overall.

Taken together, Figure 5.5 illustrates that when the mixed model parameterizations are observationally informed, the resulting schemes predominantly converge toward the behavior of purely data-driven approaches across a wide range of atmospheric conditions. However, under moist and unstable conditions, the mixed schemes exhibit a modest shift toward the conventional Tiedtke scheme. This calibration enables a more robust interpretation of convective processes by constraining the inverse problem of mapping convective tendencies as a function of column water vapor, lower-tropospheric stability, and geographic context. The resulting parameterizations yield physically interpretable regime behavior while mitigating the risk of extrapolation in regions of low confidence.

As illustrated in Figure 5.4, the ML weight exhibits a dependence on both latitude and topography. To further investigate this relationship, Figure B.4 presents the convective precipitation and ML weight as functions of the surface height. The convective precipitation

displays a non-monotonic relationship with surface height, characterized by an initial decrease followed by a sharp increase at high elevations (above 3 km to 4 km). The relatively low (but still over 60 %) ML weights obtained for sea surface heights are consistent with the challenges associated with predicting convection within the tropics and ITCZ. Furthermore, the ML weight decreases moderately at high surface heights, indicating a subtle dependence on topography in these regions.

To investigate how the 3D outputs of the ML/mixed scheme behave we now turn our attention to profiles of the convective temperature and humidity tendencies as well as the corresponding enthalpy changes conditionally averaged on CWV for the Mixed:10-60 model with $\alpha = 0.1$. These profiles are displayed in Figure 5.6 for different values of CWV. These correspond to the transects visualized as dashed red lines in Figure B.5. Similar profiles conditionally averaged on LTS are shown in Figure B.6.

A comparison between the ML/mixed schemes and the Tiedtke scheme reveals similarities in the heating rate behavior, as evident in panels (a,c,e). The mixed scheme exhibits slightly higher tropospheric heating rates and correspondingly lower surface heating rates than the Tiedtke scheme. In contrast, the pure ML scheme displays a similar overall magnitude, but with smoother profiles as a function of height. Notably, the ML scheme lacks the mid-tropospheric decrease in heating rates observed at higher humidity values, distinguishing it from the other two schemes. The analysis may exhibit a slight bias towards higher CWV values and a relatively low ML weight, correspondingly (Figure 5.5), due to the x-axis scale. However, by zooming in, the mixed scheme and the Tiedtke schemes still show a high level of similarity.

The moistening rates depicted in panels (b,d,f) show that the mixed scheme closely resembles the Tiedtke scheme, despite the ML weight being approximately ~67 % on average. This suggests that the mixing approach effectively retains the simulation's proximity to the conventional ICON model's distribution, while incorporating ML predictions to enhance agreement with observational data, as evident in Figures 5.1 and 5.2. In contrast, the pure ML model yields smoother predictions that lack some features, such as the moistening peak at around 900 hPa, highlighting the importance of combining ML predictions with conventional approaches.

It is worth noting that for the shown profiles, the mixed model predicts heating, moistening, and precipitation in a manner that nearly conserves enthalpy, whereas the pure ML model exhibits net fluxes into the column of up to $50\,\mathrm{W/m^2}$, indicating a notable deviation from enthalpy conservation as already seen in Figure 5.3. The mean absolute enthalpy residuals are $0.003\,\mathrm{W/m^2}/1.024\,\mathrm{W/m^2}/26.037\,\mathrm{W/m^2}$ for the Tiedtke/Mixed:10-60/pure ML scheme, respectively. The residual of the pure ML model is therefore higher than for the Mixed:10-60 model by factor of over 25. Looking at the ML weight $\langle\lambda\rangle$, conditionally averaged for the same conditions, we find that the weight has a magnitude of $\langle\lambda\rangle \approx 0.65$. Therefore, the ML model is called in $\sim 65\,\%$ of the cases, showcasing that the reduced enthalpy residual is not only due to mixing with the Tiedtke scheme but also to introducing the physics-informed loss terms (see Equations (3.9)–(3.12)) during training.

For the tendencies and enthalpy changes for varying LTS and fixed $19.6\,\mathrm{kg/m^2}$ displayed in Figure B.6, the profile comparison is less clear since the Tiedtke scheme shows a high variability,

especially for lower layers. In general, the mixed model exhibits the smoothest profiles with, e.g., upward moisture transport being more visible than for the Tiedtke scheme. The net column enthalpy flux reveals the same behavior as the pure ML scheme is far from conserving enthalpy, while the mixed scheme is much closer to conservation.

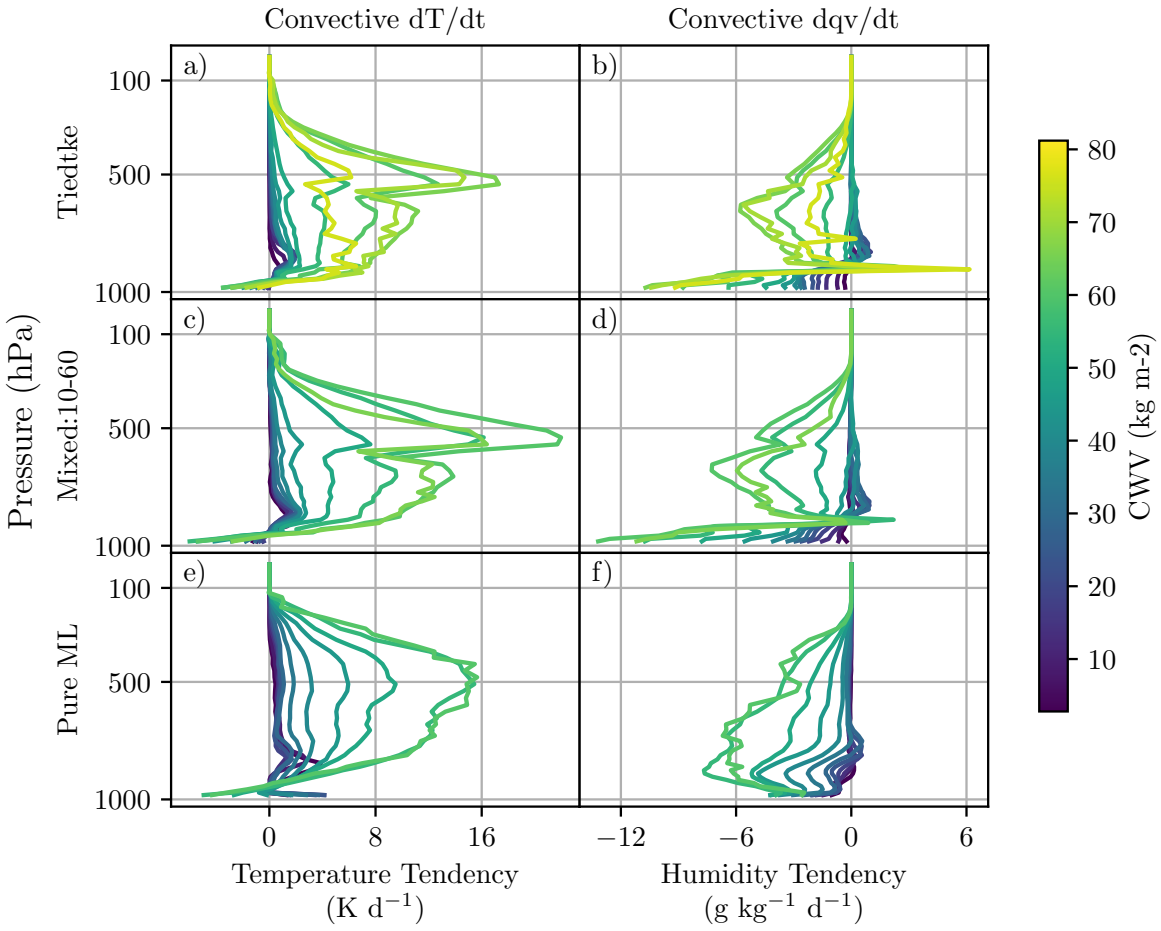

Figure 5.6.: Conditional averages of convective heating rates (first column) and moistening rates (second column) as a function of height. The conditioning is based on CWV while we keep the value for the LTS fixed to LTS = 11.4 K. Each row corresponds to a different coupled scheme: (a,b) for Tiedtke, (c,d) for Mixed:10-60, and (e,f) for the pure ML scheme. Conditional averaged curves are only computed for CWV conditions having at least ten samples. Adapted with permission from Heuer et al. (2025).

5.2.4. Twenty-year AMIP run

In this section, we evaluate AMIP-style simulation runs for 20 years (1979-1998) with the presented ML and mixed schemes. We have already demonstrated the stability and skill of the method for one year long simulations, but longer simulations remain to be investigated.

Online runs with the originally developed schemes diverged after 1.5 - 3 years. As the schemes are trained on the ClimSim dataset and even under the assumption that they are unbiased estimators of the true subgrid tendencies on this dataset, the transfer to the new

domain (ICON) can transform them into biased estimators. Therefore, small errors can add up over time and finally lead to the coupled model diverging.

Using the method introduced in Section 3.6.6, we therefore made the schemes more robust by dynamically adjusting the noise variance such that the model maximally loses ΔR^2 of its predictive skill while increasing its robustness through the addition of noise. We applied this method to the pure ML model and the ML model with a physics-informed weight $\alpha = 0.1$ with $\Delta R^2 = 0.2$.

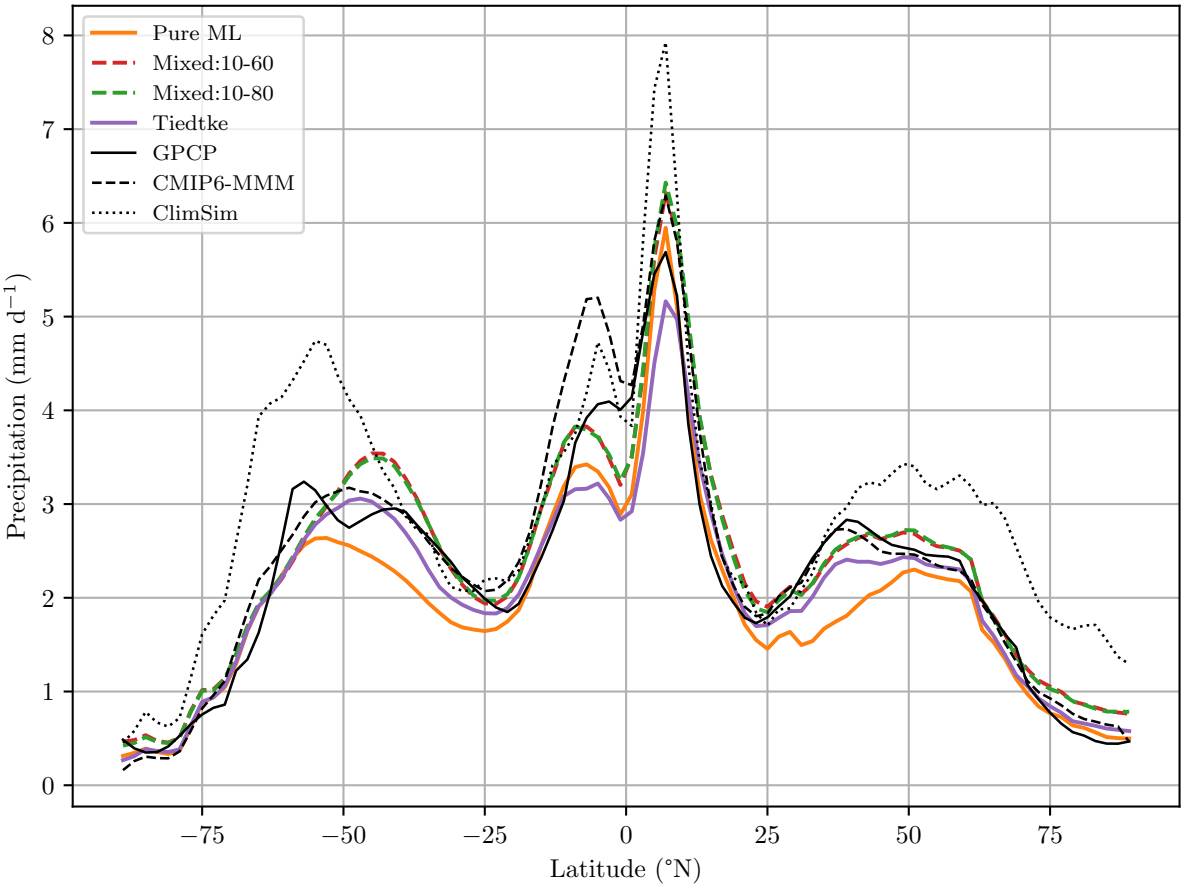

Figure 5.7.: Zonal mean precipitation evaluated over twenty years for the observational dataset (GPCP), the Tiedtke scheme, the pure ML scheme, the Mixed:10-60 scheme, the Mixed:10-80 scheme, the CMIP6 MMM, and the ClimSim dataset. For ClimSim, the zonal mean precipitation is evaluated over its available 10-year simulation period. Adapted with permission from Heuer et al. (2025).

Although ClimSim shows the smallest deviations in the A_P and E_P metrics, Figure 5.7 reveals that its zonal distribution deviates substantially from observations. Notably, precipitation is overestimated in the extratropics, along the ITCZ, and at high latitudes in the Northern Hemisphere. This larger mean bias is also reflected in the RMSE score for ClimSim in Table 5.2, which is approximately twice as high as that of the second-worst model, indicating a significant overall bias.

The CMIP6 MMM shows a reasonable zonal mean precipitation distribution in general but has a substantial double ITCZ bias which is also reflected in the highest overall bias of the

Data	A_P	E_P	A_P Bias	E_P Bias	RMSE (mm/d)
GPCP	0.189	1.163	-	-	-
Tiedtke	0.247	0.909	**0.058**	-0.254	0.382
Pure ML	0.257	0.924	0.068	**-0.239**	0.459
Mixed:10-60	0.275	0.901	0.086	-0.262	0.380
Mixed:10-80	0.279	0.902	0.090	-0.261	**0.375**
ClimSim	0.268	0.973	0.079	-0.190	0.904
CMIP6-MMM	0.060	1.037	-0.129	-0.126	0.394

Table 5.2.: The tropical precipitation asymmetry index A_P and the equatorial precipitation index E_P, and their biases with respect to GPCP for the data shown in Figure 5.7. Adapted with permission from Heuer et al. (2025).

tropical precipitation asymmetry index A_P of -0.129 as reflected in Table 5.2. The MMM also has a relatively low RMSE with $0.394\,\mathrm{mm\,d^{-1}}$ but is outperformed by, e.g., the Mixed:10-80 model with a RMSE of $0.375\,\mathrm{mm\,d^{-1}}$.

The zonal mean precipitation shown in Figure 5.7 and the corresponding biases summarized in Table 5.2 indicate that all models produce reasonably realistic distributions over the 20-year simulation period. Among these models, the Tiedtke model exhibits the smallest bias in the asymmetric precipitation component, while the pure ML model performs best in capturing the symmetric component among the schemes. The Mixed:10-80 model achieves the lowest RMSE when compared to observational data, despite the relatively high RMSE of the ClimSim distribution. The RMSE, which is arguably the more meaningful metric as mentioned in Section 5.2.1, is slightly better for the mixed mode compared to the Tiedtke model with a difference of $0.007\,\mathrm{mm\,d^{-1}}$.

For the spatial distribution of the mean precipitation shown in Figure 5.8, the Tiedtke and pure ML models show negative biases of $-0.21\,\mathrm{mm\,d^{-1}}$ and $-0.34\,\mathrm{mm\,d^{-1}}$, respectively, whereas the Mixed:10-60 scheme yields a slightly smaller positive bias of $0.14\,\mathrm{mm\,d^{-1}}$. Spatially, the mean biases shown in Figure B.7 show a very similar distribution with a general slight overestimation of mean precipitation and underestimation patterns mainly seen over low-latitude, continental regions. Similarly, in terms of near-surface temperature T_{2m} (Figure B.7), the Mixed:10–60 model exhibits the smallest mean bias ($-0.26\,\mathrm{K}$) over the 20-year period, compared to the Tiedtke scheme ($0.5\,\mathrm{K}$) and the pure ML model ($1.03\,\mathrm{K}$). When looking at the timeseries of the global mean near-surface temperature no considerably drift could be observed for all the shown simulations here (not shown). These results, summarized in Table B.3, further highlight the potential of the confidence-guided-mixing approach to enhance the accuracy of climate simulations, particularly in long-term integrations.

5.3. Summary

Through our proposed confidence-guided mixing, we developed robust parameterizations that yielded successful decade-long runs. Impressively, this is true despite our parameterizations

being trained on a dataset generated by another GCM, enabling the ICON-A model to benefit from the advantages of a superparamerized GCM. This study provides a proof-of-concept demonstrating that, through careful data preprocessing and deliberate model design choices — including confidence-guided mixing, loss function design, physics-informed training, and additive noise injection — it is possible to transfer ML convection schemes from one GCM to another without compromising stability and accuracy. The mean weight given to the ML-transferred parameterization is ≈ 0.67, confirming a fundamental change in the convective parameterization's behavior rather than a simple bias correction of the Tiedtke scheme.

When training on the ClimSim dataset, we first separated the radiative from the convective heating tendencies using the RTE+RRTMGP radiation scheme. To achieve this, we modified the scheme slightly to match the version used in ICON and to allow us to input full columns from the ClimSim data as explained in Section 3.5.2. We note that this separation represents an approximation of the true radiative tendencies employed by the E3SM-MMF model, as the radiation scheme was run for multiple radiation columns in each grid cell of the multiscale modeling framework, and we only have access to the coarse-grained state in ClimSim. Future versions of ClimSim would benefit from outputting radiative tendencies explicitly, enabling process-based training rather than emulating all subgrid physics. Likewise, the SRMs in E3SM still parameterize sub-SRM processes (e.g., turbulence, microphysics), which contribute to ClimSim tendencies; outputting those terms separately would further facilitate process-based schemes. The training would also benefit from a more accurate representation of precipitation in the ClimSim data (see Figure 5.7).

After generating the training data and designing the model and loss function, we performed a thorough hyperparameter search, an essential step for finding a good trade-off between accuracy and computational efficiency, with the number of multiply-accumulate operations proving to be a well-suited measure of computational complexity (for CPU inference). Our results revealed 181 candidate schemes along the Pareto front when comparing different metrics. Some of these models were found to perform even better than the conventional Tiedtke parameterization used in ICON, a promising outcome considering that ICON has been calibrated to behave optimally with the Tiedtke scheme. In particular, the representation of precipitation, water vapor, and near-surface temperature potentially benefits from the confidence-guided mixing approach as demonstrated in Figure 5.1.

The inclusion of physics-informed terms in the loss function improves model performance across various metrics. Specifically, adding the residuals of conserved quantities to the loss function led to improved conservation online, as evident in Figure 5.3. However, it is likely that using a training dataset where conservation laws can be strictly enforced without any net in- or out-fluxes into the columns would further improve the method. Creating such a dataset would be a crucial next step in further improving the here shown proof-of-concept method.

Investigating the conditions under which the ML/mixed schemes produce convective precipitation revealed a reasonable behavior, with precipitation generally increasing with higher column water vapor and decreasing with higher atmospheric stability as shown in Figure 5.5. Notably, the mixed scheme does not fully shut down convection under high-

stability conditions, which may help when convection is forced by, e.g., large-scale horizontal advection or orographic forcing. Moreover, we observed that the confidence of the mixed schemes decreased in regimes with few training samples as well as in regions characterized by high variability of precipitation. Conditionally averaged heating and moistening profiles in Figure 5.6 show substantial differences between the pure-ML, mixed, and Tiedtke schemes. Despite an average ML contribution of approximately ~67 %, the mixed scheme resembles the conventional ICON model in its physical behavior more closely than the pure ML model, maintaining dynamical consistency and avoiding out-of-distribution predictions while still leveraging the ML component's learned physical relationships. Additionally, our analysis of the enthalpy profiles demonstrated again that the mixed scheme learned with a physics-informed weight of only 0.1 substantially improved conservation of enthalpy. These results were based on one-year-long simulations and cannot really be expected to be robust, but as a proof-of-concept, it shows that the schemes could be adjusted to work well, even outperforming Tiedtke for some metrics. Additionally, we showed that they potentially can be tuned to observations and learned from due to analyzing their emergent precipitation statistics as shown later. Performing 20-year-long simulations for all 181 candidate schemes would have been computationally infeasible.

Finally, as demonstrated in Section 5.2.4, we achieved long-term stability using an engression-like technique, which provided data-driven extrapolation by effectively forcing the ML model to behave smoothly for small input perturbations. This result could potentially help many more ML-based parameterization schemes which very commonly struggle with long-term stability when coupled to GCMs. The results regarding precipitation and temperature patterns shown in Sections 5.2.1 and 5.2.4 indicate that the pure ML and mixed schemes are capable of generating realistic patterns, which for near-surface temperature even outperform the Tiedtke baseline with respect to observational references by having a mean bias about half as large as for the Tiedtke model for the 20-year evaluation as shown in Table B.3. However, calibration against observational data may further enhance the predictive skill of all models examined.

As illustrated in Figure 5.4 and also Figure 5.5, the ML scheme exhibits relatively high confidence in the extratropics and high latitudes while maintaining a non-zero contribution in the tropical regions that were used to design the Tiedtke scheme. The examination of the results from the twenty-year-long simulations in Figure 5.7 suggests that this confidence may be overestimated due to out-of-distributions estimates, highlighting the potential benefit of developing a separate convective triggering scheme to improve overall model performance. Moreover, training on a dataset which is closer to observational references for, e.g., the zonal mean precipitation (Figure 5.7) would also benefit the model developement.

As we developed a tunable ML-based scheme, future work should also prioritize proper tuning, exploring various settings of parameters such as p_0, p_1, the level of stochastic noise injection, and the weighting α of physical loss terms in the hybrid objective function to further optimize the scheme's performance. Furthermore, the confidence estimates produced by the ML model could be leveraged to develop a stochastic parameterization framework, transforming the current deterministic predictions into probabilistic outputs. Such a stochastic formulation

would better represent subgrid variability and improve the representation of uncertainty in climate and weather simulations.

Ultimately, our goal is to implement an ML-based convection scheme into ICON-XPP-MLe (where XPP stands for eXtended Predictions and Projections and MLe for machine learning enhanced) (Müller et al. 2025). Realizing this goal will require further work before the current proof-of-concept can be effectively deployed within this hybrid ESM. This will include systematic tuning of the scheme and hybrid ESM, potentially through automated methods such as the approach proposed by Grundner et al. (2025), further testing, and potentially interpolating the training data to the vertical levels of ICON-XPP-MLe. This would ensure seamless integration and optimal performance of the ML-based parameterization scheme within the broader modeling framework. Another important direction for future research is to assess the sensitivity of the ML scheme to horizontal resolution. We plan to evaluate its performance at higher resolutions, such as 80 km × 80 km, to determine its scalability and robustness across different model configurations. This will help clarify whether the learned relationships generalize across resolutions or require designing a scale-aware version of the scheme.

Additionally, a direct integration with the ICON-XPP-MLe modeling framework may be facilitated by incorporating ICON-specific simulation data into the training pipeline. A suitable dataset would have to fulfill several constraints regarding the length of the simulated time period and spatial extent, frequency of output, and scale separation, as mentioned in the introduction. Given such a dataset, the inclusion of ICON data may be achieved either through retraining the model on the ICON output or by applying transfer learning techniques to adapt the existing models further to the ICON model.

Together, these developments, ranging from stochastic extensions to resolution dependence studies and model-specific adaptation, will be crucial for advancing the reliability, robustness, and applicability of ML-based parameterizations in long-term climate simulations.

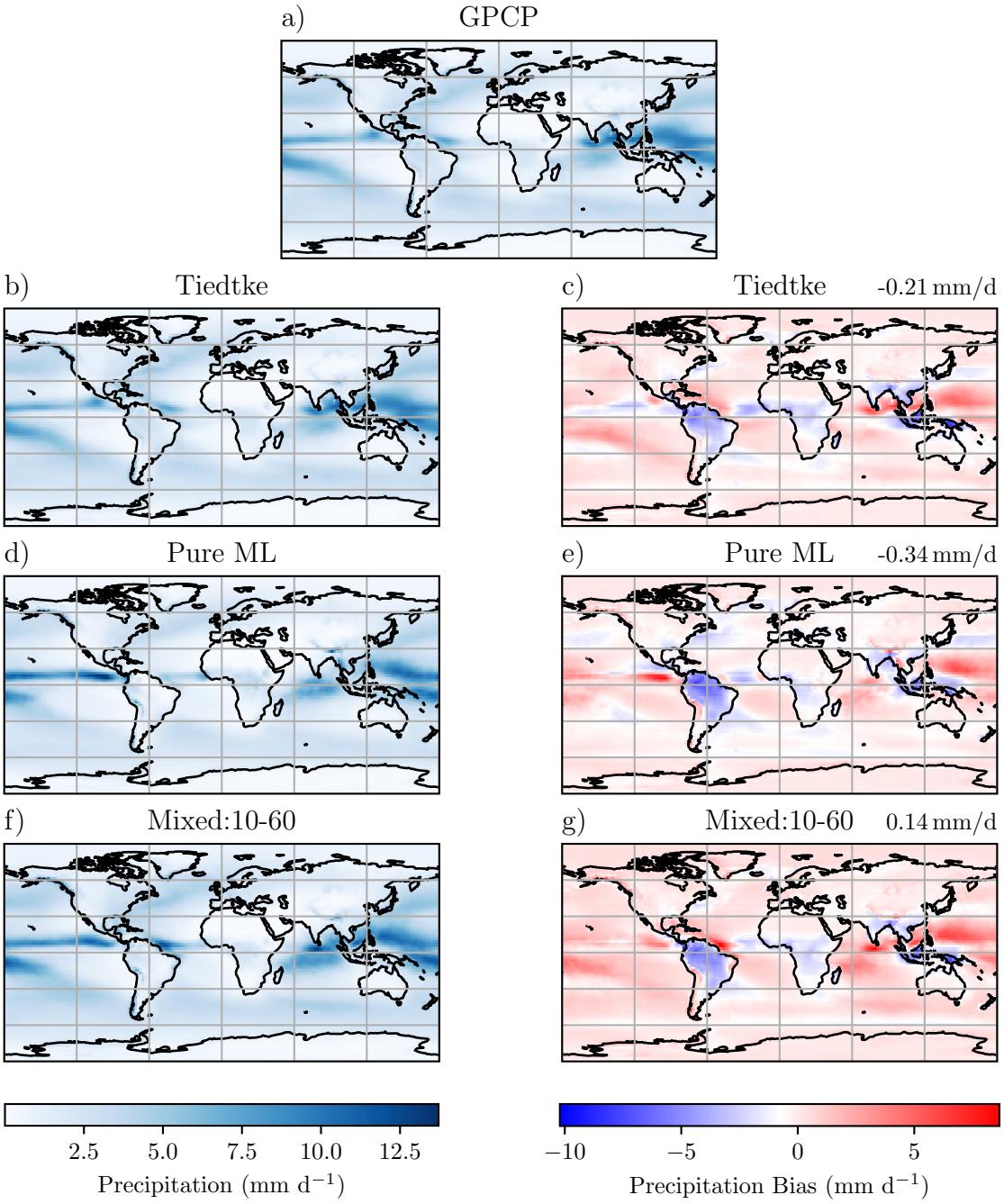

Figure 5.8.: The spatial distribution of 20-year averaged precipitation for different convection schemes in the left column and the bias with respect to GPCP in the right column. The first row (a) shows precipitation for the GPCP data, the Tiedtke scheme in the second row (b-c), the pure ML scheme in the third row (d-e), and the Mixed:10-60 scheme in the last row (f-g). In the upper right of each bias plot, the area-weighted mean bias is displayed. Adapted with permission from Heuer et al. (2025).

6. Conclusion

This chapter begins with a summary of the key findings and limitations of the two studies presented in this thesis. The key science questions introduced in Section 1.1 are answered and addressed in light of the results. Subsequently, an outlook is provided, highlighting promising avenues for future research and discussing potential improvements to the developed framework, with the goal of facilitating its integration into the operational version of the ICON model.

6.1. Overall Summary

While human-caused global warming is unequivocal (Eyring et al. 2021a; IPCC 2021), the exact projection remains uncertain (Forster et al. 2021). This is a problem for mitigation and adaptation planning as decision-makers require reliable projections to develop effective and resilient long-term strategies (Eyring et al. 2024b). From a broader perspective, this thesis contributes to the advancement of data-driven methods aimed at reducing uncertainty in climate projections and enhancing the understanding of the underlying physical mechanisms driving climate change.

Taken together, the two studies presented in this thesis contribute to a deeper understanding of the opportunities and limitations associated with ML in the context of convective parameterizations. They illustrate a maturing methodology in which ML is no longer applied as a black-box predictor but is more thoughtfully integrated into the modeling workflow with attention to causality, conservation laws, and numerical robustness. The regional, high-resolution focus of Chapter 4 provides insight into the physical mechanisms of convection and the risks of overfitting to non-causal correlations. In contrast, Chapter 5 demonstrates how global, statistically robust datasets, combined with carefully regularized architectures, can enable accurate, long-term applications across GCMs, bringing ML parameterizations closer to operational use in climate modeling. Apart from the core studies in this thesis, I also contributed to Yu et al. (2025) by testing the newly developed containerized pipeline for running and evaluating online simulations.

The first study (Chapter 4) focuses on the development of an ML-based parameterization trained directly on output from a high-resolution regional storm-resolving simulation of the tropical Atlantic using the ICON model at approximately ~2.5 km resolution over a two-month period: the NARVAL campaign simulations (Klocke et al. 2017; Stevens et al. 2019a). This approach enables the derivation of physically consistent subgrid fluxes through a carefully designed preprocessing procedure that follows energy, mass, and momentum conservation

principles, even with terrain-following coordinates, by explicitly accounting for horizontal density fluctuations. A filtering technique is applied with the goal of isolating convective contributions from other dynamical effects such as gravity waves so that the ML models specifically learn convective transport. Evaluating a suite of ML architectures, including U-Nets, convolutional neural networks, multilayer perceptrons, and ensemble tree methods (RF, ET, and GBT), demonstrates that the U-Net achieves superior performance in reconstructing 3D thermodynamic and momentum fluxes offline. However, applying an XAI technique, specifically a SHAP value analysis (Lundberg and Lee 2017), uncovers non-causal predictors related to precipitation, which are consequences rather than drivers of convection. This insight motivates an ablation experiment in which non-causal inputs are removed. While the resulting models exhibit slightly lower offline accuracy, they demonstrate dramatically improved behavior when coupled online. Indeed, when coupled globally, only the causally consistent models remain numerically stable over month-long integrations, underscoring the importance of embedding physical reasoning into the ML workflow as a guiding principle in model design. By applying an XAI technique, I directly addressed Key Science Question 1 ("How can we ensure that machine learning parameterizations learn physically consistent, causal, and interpretable relationships, rather than spurious correlations, when learning complex atmospheric convection?"), pointing out that XAI can be essential for identifying non-causal relationships and building trust in learned patterns by linking them to established physical mechanisms, such as the effects of wind shear on convective cells, the stabilizing effects of convection, and downgradient diffusive momentum transport. This trust, established through offline analysis, translates into stable simulations of at least 180 days when coupling the NN in regions corresponding to the tropical training data. These results also address Key Science Question 2 ("Can machine learning parameterizations of subgrid convection improve the representation of convective processes in coarse-resolution models while maintaining long-term numerical stability?") as the simulations show significantly improved predictions of extreme precipitation (see Figure 4.8). However, coupled runs exhibit a significant smoothing bias in the spatial distribution of column water vapor and mean precipitation, as well as biases in mean temperature. Moreover, most coupled integrations that extend the application of the ablated U-Net beyond tropical latitudes, outside the range of the training data, diverge within 180 days, underscoring the challenges of out-of-distribution generalization.

Motivated by the lessons learned in the first study, the second study (Chapter 5) shifts toward a global, longer-term perspective, aiming to develop a proof-of-concept ML-based parameterization capable of stable integration over climate-relevant timescales. To ensure broad representativeness and leverage the inherent scale separation in MMFs (see Section 2.4 and Chapter 5), this study uses the ClimSim dataset (Yu et al. 2025): a massive, global, superparameterized dataset, spanning 10 years. It was generated using the E3SM-MMF (Hannah et al. 2020), which embeds 2D SRMs within each grid column of a coarser GCM. The developed ML model builds on a BiLSTM architecture, inspired by the winner of the 5th place ("YA HB MS EK") in a Kaggle competition (Lin et al. 2024). To enhance physical consistency and robustness, I introduced three innovations: (1) a confidence-guided mixing strategy, in

which the magnitude of the ML model's uncertainty estimate guides mixing the model's predictions with those of the conventional scheme (e.g., Tiedtke (1989)), effectively forming a hybrid parameterization; (2) a physics-informed loss function that penalizes violations of energy, momentum and moisture conservation; and (3) additive noise during training, inspired by the "Engression" framework (Shen and Meinshausen 2024), to improve the robustness of the learned ML-based parameterization. The way in which I addressed Key Science Question 1 therefore changed in Chapter 5: by leveraging the confidence estimate of the NN and introducing inductive biases directly through the loss function design, the model was encouraged to learn physically consistent patterns and avoid out-of-distribution failure by mixing in the conventional parameterization to a larger extent when confidence is low. Most critically, in addressing both Key Science Question 2 and Key Science Question 3 ("To what extent can ML-based parameterizations be transferred across climate models, and how can training strategies be optimized to support robust hybrid climate modeling?"), incorporating the additive noise during training significantly enhances the model's robustness when transferring the ML-based scheme between different GCMs. This prevents catastrophic failures during extended integrations. As a result, coupled simulations using the ML scheme with the full ICON atmosphere model at ~160 km horizontal resolution remain stable for at least 20 years, demonstrating the successful long-term integration of an ML-based convection parameterization in a comprehensive GCM. These simulations not only maintain numerical stability but also show reduce mean climate biases compared to the conventional scheme. Improvements are evident in precipitation statistics, water vapor distribution, and near-surface air temperature when evaluated against observations. However, as seen in Chapter 5, the ClimSim dataset still has a relatively low spatial resolution with 1.5° and the dataset is not unbiased. For example, the zonal mean precipitation does not match the observational climatology derived from GPCP. Additionally, within the developed framework, conservation laws are implemented only as soft constraints; enforcing them strictly would require either a refined training dataset or a posteriori adjustments within ICON. This limitation may become increasingly important for century-scale simulations, particularly given that transferring ML parameterizations across climate models can lead to out-of-distribution predictions and violations of conservation laws. Furthermore, the hybrid scheme still depends partially on the conventional convection parameterization, and this dependence could be reduced through training on a more suitable dataset. Nevertheless, the online evaluation results presented in Chapter 5 demonstrate substantial progress over the shorter, less stable integrations reported in Chapter 4. Accordingly, Key Science Question 2 can be answered affirmatively: long-term, multi-decadal coupled simulations with an ML-based convection parameterization are feasible and can yield a more accurate representation of subgrid convection than a conventional scheme. Chapter 5 also addresses Key Science Question 3 by demonstrating that transferring knowledge across climate models is possible by deliberate model design choices such as confidence-guided mixing, loss function design, physics-informed training, and additive noise injection. Together with careful preprocessing of the training data to isolate the signal of interest (convection

through removal of radiative tendencies in Section 3.5.2), these elements enhance model robustness and enable the transferability across models.

Beyond technical innovation, this dissertation emphasizes a broader philosophical shift required to integrate data-driven methods into a field grounded in physical theory. Machine learning should not replace physical understanding but rather serve as a complementary tool within hybrid modeling frameworks. The ultimate goal is not merely to build faster or more accurate emulators, but to develop smart, reliable, and robust parameterizations that improve both the skill and efficiency of climate models. As we enter the era of exascale computing and kilometer-scale global models (Neumann et al. 2019), the insights gained here contribute to bridging the gap between explicit simulation and parameterized representations of convection.

6.2. Outlook

The results presented in this thesis extend and improve upon previous efforts toward integrating ML into climate modeling. Several promising directions remain to further enhance the robustness, generalizability, and physical fidelity of ML-based parameterizations. An immediate next step involves transferring the framework developed in Chapter 5 to the ICON-XPP-MLe model (where "MLe" denotes "machine learning enhanced"), which features different horizontal and vertical grid spacings than the configuration of ICON-A used in this work. This transition will likely require vertical interpolation and retuning of critical hyperparameters, including the mixing weights p_0 and p_1, stochastic noise amplitude, and the relative weighting of physics-informed loss terms. Such optimization could be achieved through automated strategies (Grundner et al. 2025). Depending on the performance, retraining or applying transfer learning using outputs from ICON simulations, aligned with the sub-timestep resolution strategy outlined in Chapter 4, may also be necessary.

Future work should explore stochastic formulations of ML-based parameterizations, leveraging model confidence estimates or other uncertainty quantification methods to better represent subgrid variability. These approaches could improve consistency with the scaling symmetries of the Navier-Stokes equations (Christensen et al. 2024; Palmer 2019) and enhance the representation of weather and climate variability (Buizza et al. 1999). Additionally, training ML models on more diverse climatic states, including, e.g., paleoclimate, present-day, and projected future conditions, would be computationally demanding but could be essential for ensuring reliable performance under a changing climate.

Incorporating causal learning principles (Iglesias-Suarez et al. 2024; Tang et al. 2025) offers a promising pathway toward discovering physically meaningful, interpretable relationships while reducing dependence on spurious correlations. Moreover, future iterations of training datasets such as ClimSim would greatly benefit from explicitly providing process-level tendencies, e.g., for radiation, turbulence, microphysics, and surface exchanges. This would facilitate the development of modular, process-based ML parameterizations. More broadly, rigorously assessed and verified training data, along with improved methods for separating processes

that are dynamically resolved in SRMs and MMFs, would further enhance model accuracy and reduce systematic errors.

Collectively, these advancements, spanning cross-model adaptability, stochastic modeling, causal inference, and enhanced training data design, are critical for realizing the full potential of ML in climate science. They represent essential steps toward developing reliable, robust, and physically consistent hybrid models capable of producing more accurate long-term climate projections.

Appendix

A. Supplementary Materials for Chapter 4

The work was already published as supporting information for Heuer et al. (2024). As indicated in Section 1.2, the author of this thesis created all the content, including text, figures, and tables, that is presented from this publication and implemented the code [1] to reproduce this study.

Some additional figures, presenting more details to the results shown in the main document, are presented in this section.

Figure A.1 displays the predicted subgrid fluxes against the true subgrid fluxes. The fluxes for the tracer species q_l, q_i, q_r, q_s are shown. The tendency to underestimate the true flux for high values which was previously observed is also visible in this figure.

The true and predicted precipitation distributions are displayed in Figure A.2. Although the difference in the R^2 value is just ~ 0.04 we see that the tail of the true distribution is much better captured by the U-Net in comparison to the GBT model.

[1] published under https://github.com/EyringMLClimateGroup/heuer23_ml_convection_parameterization (last access: 14.10.2025) and preserved (helgehr 2024)

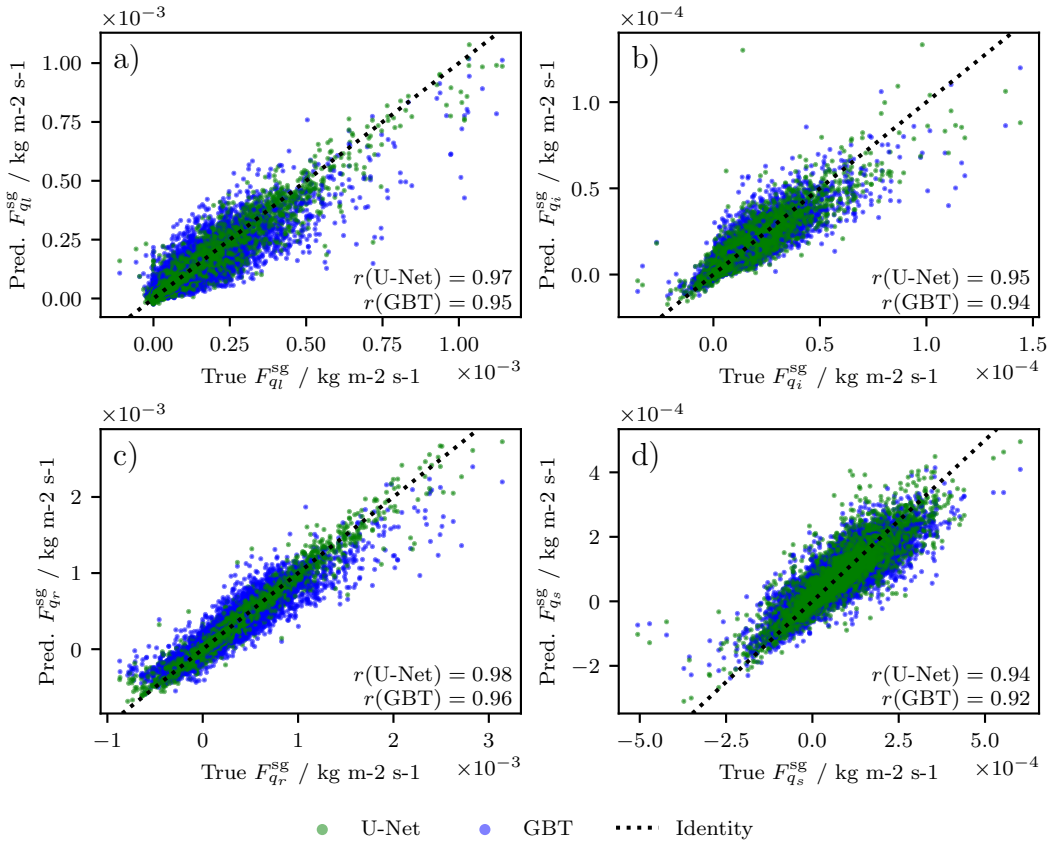

Figure A.1.: Scatterplots for the subgrid fluxes of the four remaining tracer species not shown in the main document. Data for the U-Net is shown in green, for the GBT in blue, and the diagonal is marked by a dotted line. The Pearson correlation coefficient is written in the lower right of each plot for both U-Net and GBT. Adapted with permission from Heuer et al. (2024).

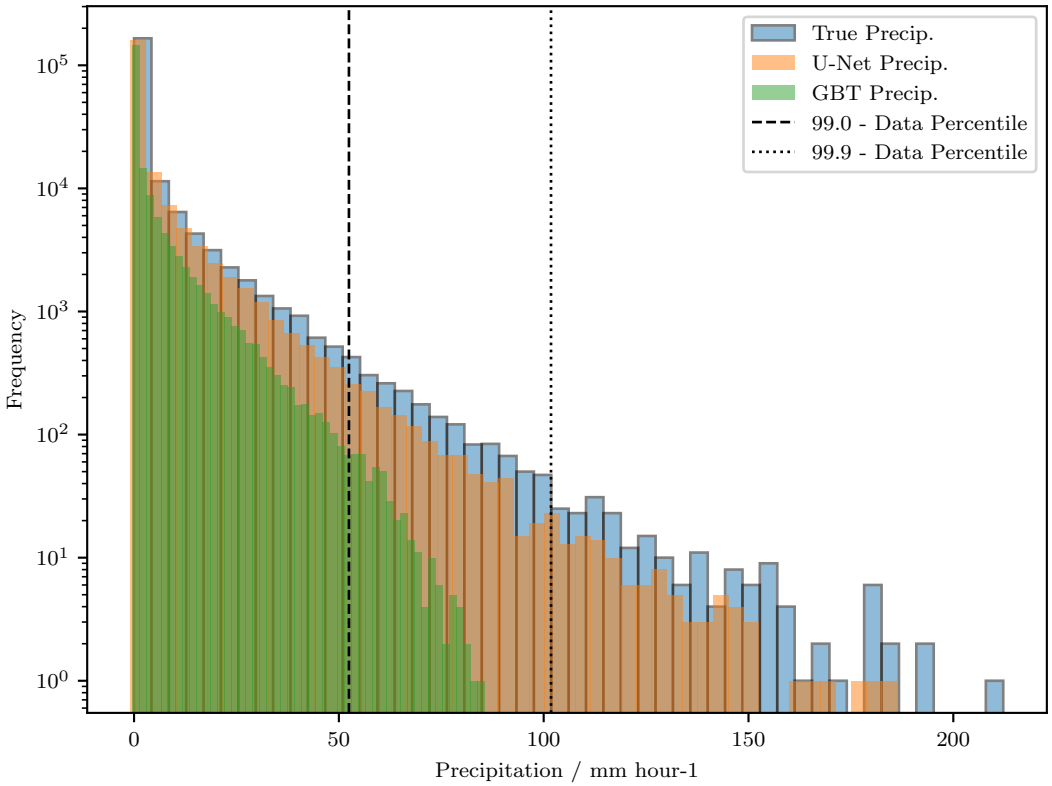

Figure A.2.: Distribution of the true precipitation and the predictions of U-Net and GBT. The 99th and the 99.9th percentile of the true precipitation are marked by the dashed and the dotted line, respectively. The precipitation R^2 scores are 0.897 for the U-Net and 0.860 for the GBT. Adapted with permission from Heuer et al. (2024).

The variance weighted RMSE of U-Net and GBT is shown in Figure A.3. We computed the RMSE across all variables and levels in this plot, excluded columns are shown as grey data points.

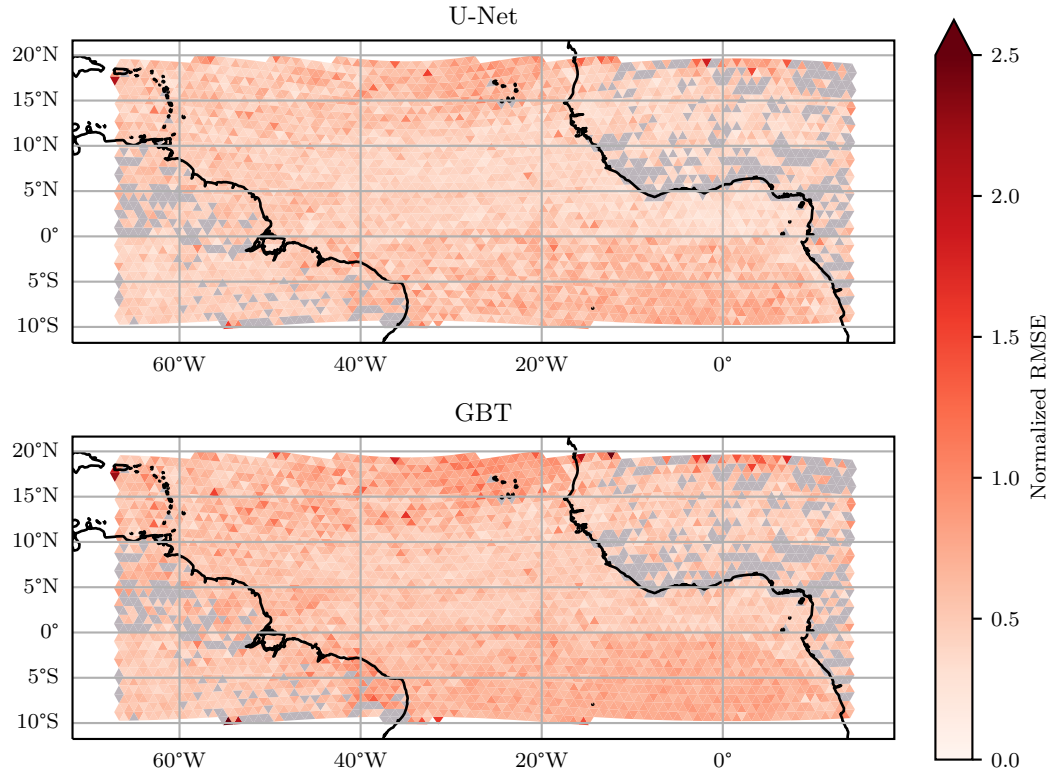

Figure A.3.: Spatial distribution of the by the variance weighted RMSE in the studied region. The top plot shows the data for the U-Net and on the bottom the result of the GBT model is displayed. Adapted with permission from Heuer et al. (2024).

Figure A.4 displays the weighted SHAP values and the feature importance for the non-ablated U-Net. By looking at the column for q_r/q_s we see that the model is heavily influenced by these variables and other variables have a comparatively lower influence on the target fluxes.

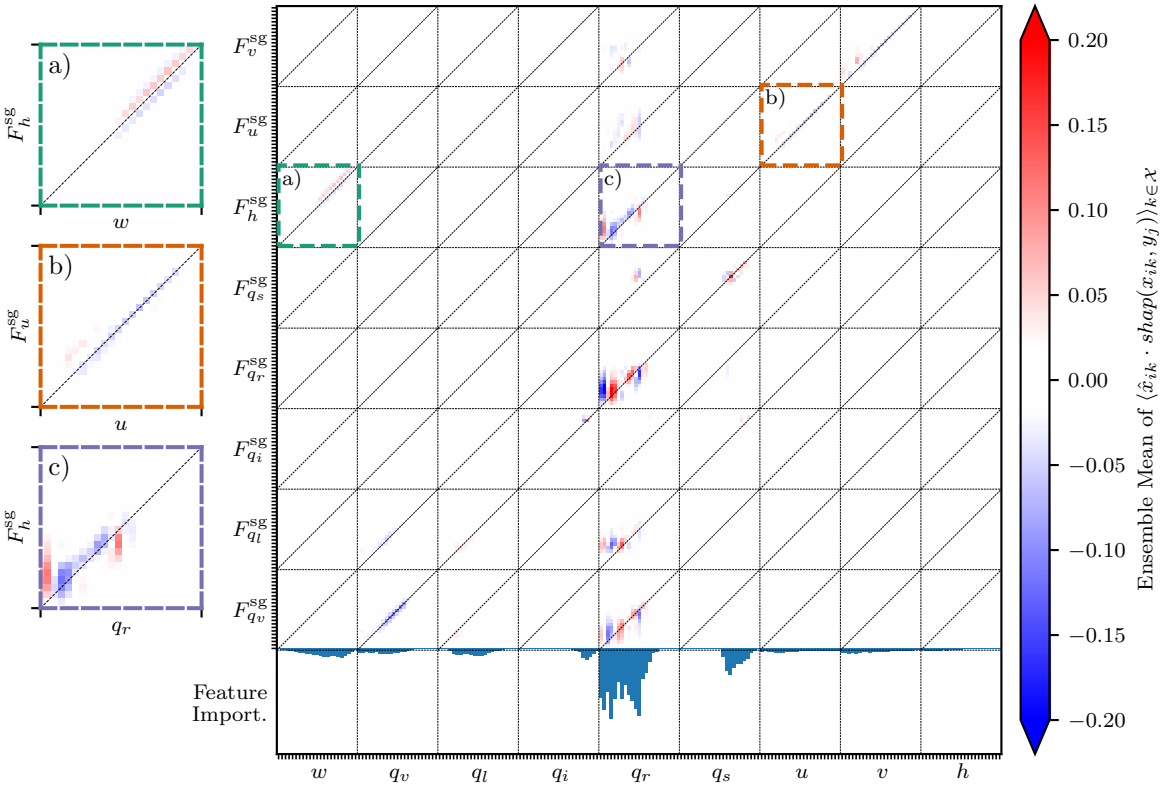

Figure A.4.: Ensemble mean of weighted SHAP values for the non-ablated U-Net model. The feature importance shown in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. The Insets a), b), and c) show a more detailed version of three specific variable pairs, the colors indicate which inset corresponds to which part of the large plot. Adapted with permission from Heuer et al. (2024).

Figure A.5 displays the weighted SHAP values and the feature importance for the non-ablated MLP. By looking at the column for q_r/q_s we see that the model is heavily influenced by these variables, as well, and other variables have a comparatively lower influence on the target fluxes.

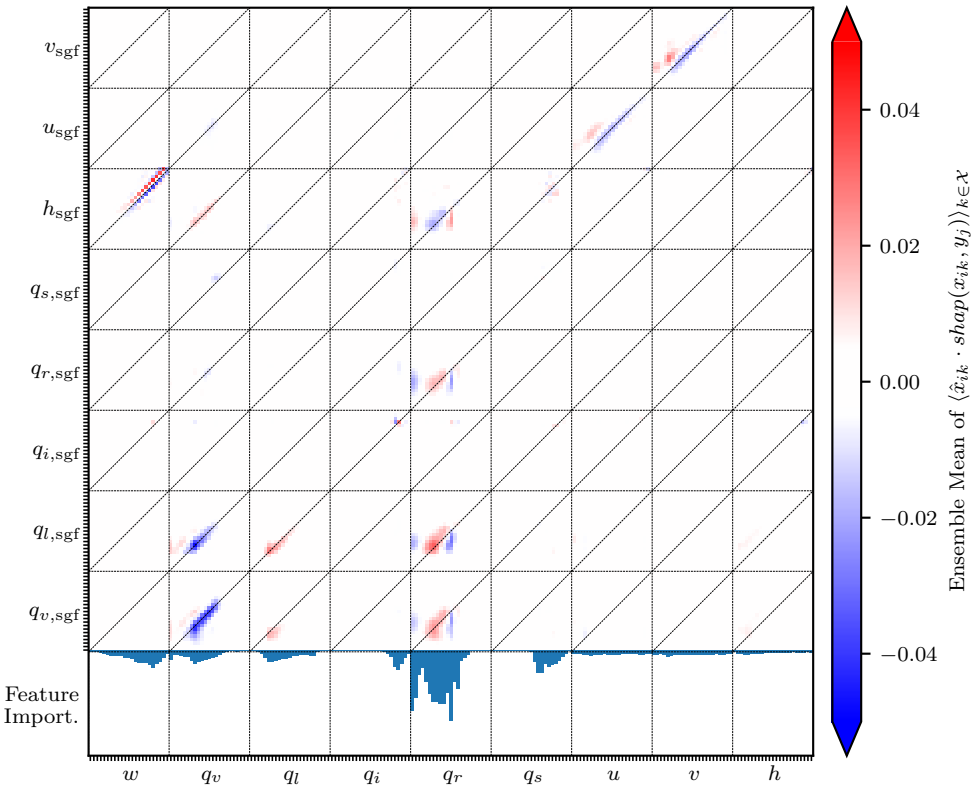

Figure A.5.: Ensemble mean of weighted SHAP values for the non-ablated MLP model. The feature importance shown in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. The Insets a), b), and c) show a more detailed version of three specific variable pairs, the colors indicate which inset corresponds to which part of the large plot. Adapted with permission from Heuer et al. (2024).

Figure A.6 shows a complexity-performance plot of the used non-tree-based models. Tree-based models are not compared in this plot as the number of parameters is not a meaningful measure of complexity for this model class. A Pareto frontier, defined as the set of points for which no other point exists with one improved metric and no metric worsened, is displayed in the figure, as well.

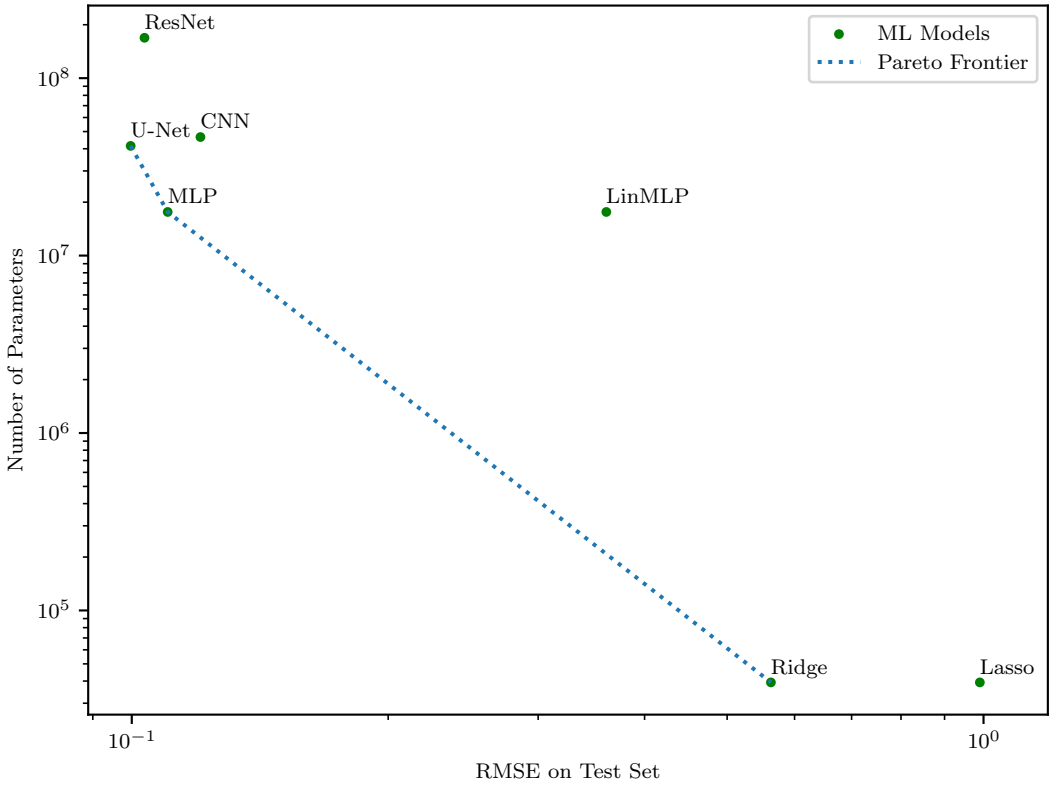

Figure A.6.: Complexity, measured by the number of parameters, is plotted against the RMSE on the test set. The Pareto frontier is visualized as the blue dotted line. Adapted with permission from Heuer et al. (2024).

The monthly mean spatial precipitation distribution of the high-resolution reference data (NARVAL) and the simulations with the convectional cumulus scheme, the ablated U-Net, and the full U-Net are shown in Figure A.7.

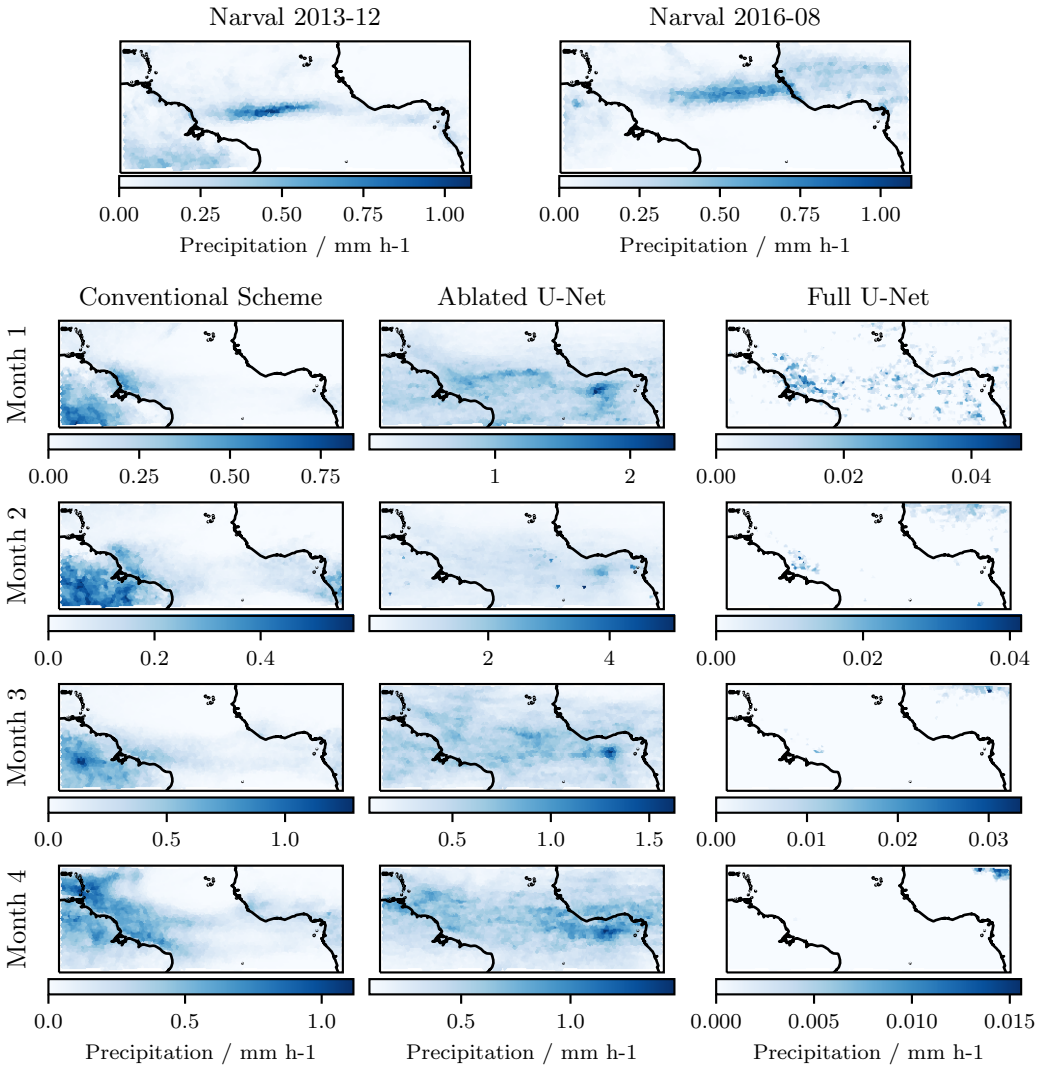

Figure A.7.: The monthly mean spatial precipitation distribution of the NARVAL data and the simulations with the convectional cumulus scheme, the ablated U-Net, and the full U-Net. For the latter three simulations the first four month of the simulations are shown. All data displayed is taken from the first ensemble member each. Adapted with permission from Heuer et al. (2025).

The data displayed in the top row and for the four rows below are not directly comparable as the initial times are different (2013/2016 for NARVAL vs. 1979 for the other plots). The position of the ITCZ, for example, is strongly dependent on the season (boreal summer vs. austral summer). However, one can see that the conventional scheme tends to underestimate the mean intensity and produces too much precipitation over land and very scarce precipitation over

the ocean where the high-resolution data shows most of precipitation. For the ablated U-Net we see higher precipitation values with more reasonable structures. Especially over the ocean there is a big difference to the conventional scheme, while the precipitation predictions look too spatially homogeneous, the ML-scheme predicts most of the precipitation over ocean, as does the high-resolution reference.

For month 2, a few grid cells with overly high mean precipitation are visible. This can be seen to a larger extent for the full U-Net, although in general, it predicts more than two orders of magnitude less precipitation than the ablated U-Net and the conventional scheme. The scarce places where the coupled full U-Net shows precipitation are constrained to one or a few grid cells, slightly similar to month 2 for the ablated U-Net but to a much larger extent.

A.1. Section S2

This section will give some really short background information on the non-deep learning models used in the study. As lowest complexity models we used linear methods such as Lasso (Tibshirani 2018) and Ridge (Hoerl and Kennard 1970) regression. These methods are a form of linear regression with additional L^1 and L^2 regularization terms. We also compared three different models based on ensembles of decision trees: RF (Breiman 2001), ET (Geurts et al. 2006), and GBT (Friedman 2002). An RF is a collection of decision trees fitted on subsets of the training data and feature set. The ET model is based on the same principle but does not sub-sample the training data set, and the splitting of individual nodes in the trees is not based on the minimization of the error but it first splits at random points for random features and only afterwards chooses the best split among these candidates (Pedregosa et al. 2011). GBTs are a part of the more general family of Gradient Boosting algorithms. This family is based on ensembles of weak learners which are fitted iteratively to the residual of the previously fitted model with respect to the target data. In the case of GBTs, the class chosen as a weak learner is a decision tree. In this study we chose to use an implementation called the Histogram-based Gradient Boosting Regression Tree (Ke et al. 2017). This model is much faster for large data sets than classic GBTs because it bins the input data first, which makes the splitting step computationally much more efficient (Alsabti et al. 1998).

A.2. Section S3

For the Hyperparameter optimization we used the `Ray Tune` library (Liaw et al. 2018). The procedure is illustrated in Figure A.8.

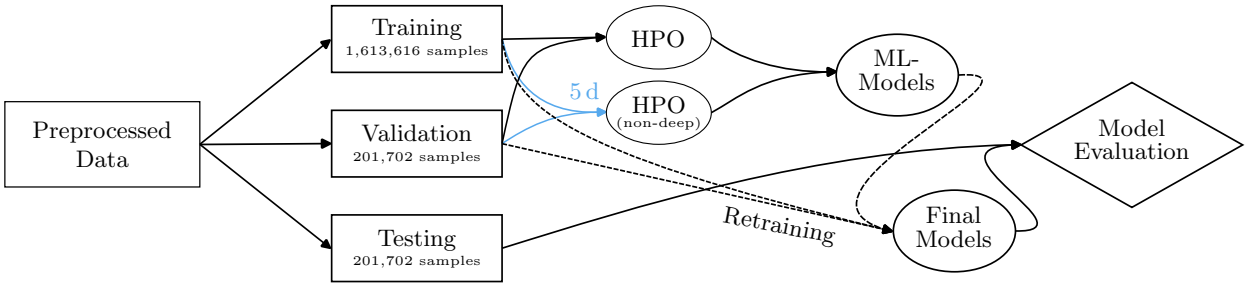

Figure A.8.: Visualization of training procedure. The preprocessed data set is split 80 %/10 %/10 % into training/validation/test data sets. The models are fit to the training data set, and validated with the validation set in the HPO. For the non-deep learning methods only 5 days are used for the HPO. The resulting hyperparameter-optimized models are then retrained on the full training and validation sets and evaluated mainly based on the coefficient of determination using the test data set. Adapted with permission from Heuer et al. (2025).

Deep Learning Models

The number of trainable parameters for the non-ablated deep learning architectures can be seen in Table A.1.

Table A.1.: Number of Trainable Parameters Used in the Various Deep Learning Models. Adapted with permission from Heuer et al. (2024).

Network	Number of parameters / 10^6
MLP	17.6
U-Net	41.5
CNN	46.6
ResNet	168.8

Table A.2 lists the hyperparameters which are common to all deep learning models. Additionally to the listed parameter ranges in Table A.2, we varied the parameter spaces described in the following for the deep learning architectures used. The parameters with the lowest error on the validation data set are marked by an asterisk.

Table A.2.: Common Parameters for the Hyperparameter Optimization of the Various Deep Learning Architectures. Adapted with permission from Heuer et al. (2024).

Parameter	Possible values
Learning rate (lr)	0.1, 0.01, 0.001, 0.0003, 0.0001
Activation function (σ)	relu, selu, gelu, sigmoid, leaky_relu
Batch size (bs)	512, 1024

MLP

For the MLP network we vary the number of layers n_l, the number of neurons in each hidden layer n_{hidden}, and whether there is a batch normalization layer after each hidden layer with the boolean parameter b.

The corresponding ranges are $n_l \in \{1, 2, 3, 4, 5^*, 6\}$, $n_{\text{hidden}} \in \{16, 32, 64, 128, 256, 512, 1024, 2048^*\}$, and $b \in \{0, 1^*\}$. The optimal values for the common hyperparameters were the learning rate 0.0003, the activation leaky_relu, and the batch size 1024.

The LinMLP architecture was chosen as the best performing MLP where all activation functions are replaced by identity functions.

CNN

The CNN network also has the parameters n_l and n_{hidden}. The number of neurons per hidden layer is optimized over the same range as the MLP model, and the number of layers is varied over $n_l \in \{0, 1^*, 2, 3, 4, 5\}$. Additionally, there is a parameter for the number of channels for the convolutional layer in the beginning of the net, n_{channels}, and for the kernel size of the convolution k.

The corresponding parameter spaces are $n_{\text{channels}} \in \{1, 8, 32, 64, 256, 512, 1024^*\}$ and $k \in \{2, 3, 4^*, 5\}$. Furthermore, the optimal values for the other hyperparameters were: lr = 0.0001, σ = selu, bs = 1024, and n_{hidden} = 2048.

ResNet

The ResNet Network is hyperparameter-optimized with respect to the number of blocks n_b, the number of hidden layers per block n_l, the number of neurons per hidden layer n_{hidden}, and whether there is a batch normalization layer after each hidden layer with the boolean parameter b.

The parameter ranges are $n_b \in \{2, 4, 8, 10^*, 14, 16\}$, $n_l \in \{1, 2, 3, 4^*\}$, $b \in \{0, 1^*\}$, and the parameter n_{hidden} does have the same range as the MLP model, with the optimal value being $n_{\text{hidden}} = 2048$. The optimal values for the other hyperparameters were: lr = 0.0001, σ = relu, and bs = 512.

U-Net

For the U-Net we also vary the number of blocks n_b, whether there is a batch norm layer after the first/second layer of each double convolution block b_1/b_2, and the number of channels in the first block n_{channels}, which is consequently doubled in each following block.

The parameter ranges are $n_b \in \{2^*, 3, 4, 5\}$, $b_1 \in \{0, 1^*\}$, $b_2 \in \{0^*, 1\}$, and $n_{\text{channels}} \in \{2^{i_m-2}, 2^{i_m-1}, 2^{i_m}\}$, where $i_m = \lfloor \log_2 (max_{\text{channels}}/2^{n_b}) \rfloor$ and $max_{\text{channels}} = 2048$. This ensures that the maximum number of channels in the lowest (having the most number of channels in the network) block of the U-Net does not have more channels than max_{channels}. The optimal values for the other hyperparameters were: lr = 0.0001, σ = leaky_relu, and bs = 1024, and $n_{\text{channels}} = 512$.

Non-Deep-Learning Models

Parameter names in this section are based on the names in the python ML-framework Scikit-learn (Pedregosa et al. 2011). Parameters with the lowest error on the validation data set are again marked by an asterisk. For Ridge and Lasso regression, we only vary the parameter α between 10^{-1} and 10^4 with 10 evenly spaced values on a log scale. The α parameter gives the weighting of the regularization term in Ridge and Lasso regression, and the optimal value for both types of regression is chosen as $\alpha = 0.1$.

For the tree-based methods RF and ET we optimize over the common parameters `bootstrap` $\in \{0, 1\}$, `max_features` $\in \{$`sqrt`$, 1\}$, `min_samples_split` $\in \{0.5, 0.1, 0.01\}$, and `n_estimators` $\in \{10, 100, 200\}$. The optimized RF uses no bootstrapping, the maximum number of features considered for a split is the square root of the number of features, the minimum number of samples to split a node is $0.01 \cdot$ `n_samples`, and the number of estimators is chosen as 200. For the ET model we end up with the same parameters except that all features are considered for splitting a node. The GBT model parameters are varied over `learning_rate` $\in \{0.5, 0.1^*, 0.01, 0.001\}$, the maximum number of leaves for each tree `max_leaf_nodes` $\in \{40^*, 30, 20\}$, the minimum number of samples per tree `min_samples_leaf` $\in \{20, 25^*, 30\}$, and the coefficient of the L^2-norm term `l2_regularization` $\in \{0^*, 0.01, 0.1\}$.

As tree-based methods can quickly become very large in memory for big data sets, we set the maximum tree depth of the methods to what was found best by the HPO on the subset of 5 days. This limits the size of the trees for subsequent training on the full data set.

B. Supplementary Materials for Chapter 5

The work was already published as supporting information for Heuer et al. (2025). As indicated in Section 1.2, the author of this thesis created all the content, including text, figures, and tables, that is presented from this publication and implemented the code [2] to reproduce this study.

B.1. Non-dimensionalization of Residual Fluxes

As written in Section 3.6.3, we start with the chosen scaling constants as they are defined in ICON (except t_0):

$$\begin{aligned} g_0 &= 9.806\,65\,\mathrm{m\,s^{-1}}, \\ t_0 &= 10\,\mathrm{s}, \\ \rho_{\mathrm{h2o}} &= 1000\,\mathrm{kg\,m^{-3}}, \\ c_p &= 1004.64\,\mathrm{J\,K^{-1}\,kg^{-1}}. \end{aligned}$$

We use these constants to derive scales for length l_0, temperature T_0, energy density e_0, mass flux m_0, velocity v_0, and pressure p_0:

$$l_0 = g_0\,t_0^2, \quad T_0 = \frac{e_0}{c_p}, \quad e_0 = g_0^2\,t_0^2, \quad m_0 = \rho_{\mathrm{h2o}}\,g_0\,t_0, \quad v_0 = g_0\,t_0, \quad p_0 = \rho_{\mathrm{h2o}}\,g_0^2\,t_0^2.$$

Furthermore, the latent heat of vaporization $L_v = 2.5008 \times 10^6\,\mathrm{J\,kg^{-1}}$ and sublimation $L_s = 2.8345 \times 10^6\,\mathrm{J\,kg^{-1}}$ are non-dimensionalized by dividing by e_0.

In ICON, the net column in/out fluxes for enthalpy, mass, zonal, and meridional momentum can be formulated as follows:

$$H_{\mathrm{res}} = \int_{z_{\mathrm{bot}}}^{z_{\mathrm{top}}} \rho \left(\frac{\partial T}{\partial t}\,c_p - \frac{\partial q_{\mathrm{l}}}{\partial t}\,L_{\mathrm{v}} - \frac{\partial q_{\mathrm{i}}}{\partial t}\,L_{\mathrm{s}} \right) dz - L_{\mathrm{v}}\,\mathcal{P}_{\mathrm{rain}} - L_{\mathrm{s}}\,\mathcal{P}_{\mathrm{snow}}, \tag{B.1}$$

$$m_{\mathrm{res}} = \int_{z_{\mathrm{bot}}}^{z_{\mathrm{top}}} \rho \left(\frac{\partial q_{\mathrm{v}}}{\partial t} + \frac{\partial q_{\mathrm{l}}}{\partial t} + \frac{\partial q_{\mathrm{i}}}{\partial t} \right) dz + \mathcal{P}_{\mathrm{rain}} + \mathcal{P}_{\mathrm{snow}}, \tag{B.2}$$

$$u_{\mathrm{res}} = \int_{z_{\mathrm{bot}}}^{z_{\mathrm{top}}} \rho\,\frac{\partial u}{\partial t}\,dz, \tag{B.3}$$

$$v_{\mathrm{res}} = \int_{z_{\mathrm{bot}}}^{z_{\mathrm{top}}} \rho\,\frac{\partial v}{\partial t}\,dz. \tag{B.4}$$

[2] published under https://github.com/EyringMLClimateGroup/heuer25james_ml_convection_climsim (last access: 14.10.2025) and preserved (helgehr 2025)

Using hydrostatic equilibrium for the background vertical coordinate:

$$dp = -\rho\, g_0\, dz, \tag{B.5}$$

we convert the vertical integration coordinate from elevation to pressure:

$$H_{\text{res}} = \int_{p_{\text{top}}}^{p_{\text{bot}}} \frac{1}{g_0} \left(\frac{\partial T}{\partial t} c_p - \frac{\partial q_{\text{l}}}{\partial t} L_{\text{v}} - \frac{\partial q_{\text{i}}}{\partial t} L_{\text{s}} \right) dp - L_{\text{v}} \mathcal{P}_{\text{rain}} - L_{\text{s}} \mathcal{P}_{\text{snow}}, \tag{B.6}$$

$$m_{\text{res}} = \int_{p_{\text{top}}}^{p_{\text{bot}}} \frac{1}{g_0} \left(\frac{\partial q_{\text{v}}}{\partial t} + \frac{\partial q_{\text{l}}}{\partial t} + \frac{\partial q_{\text{i}}}{\partial t} \right) dp + \mathcal{P}_{\text{rain}} + \mathcal{P}_{\text{snow}}, \tag{B.7}$$

$$u_{\text{res}} = \int_{p_{\text{top}}}^{p_{\text{bot}}} \frac{1}{g_0} \frac{\partial u}{\partial t}\, dp, \tag{B.8}$$

$$v_{\text{res}} = \int_{p_{\text{top}}}^{p_{\text{bot}}} \frac{1}{g_0} \frac{\partial v}{\partial t}\, dp. \tag{B.9}$$

Finally, substituting all dimensional quantities with their respective non-dimensional counterparts (marked by a tilde) times the corresponding physical scale yields the following non-dimensional fluxes of enthalpy, mass, zonal and meridional momentum as shown in Section 3.6.3:

$$\widetilde{H}_{\text{res}} = \int_{\tilde{p}_{\text{top}}}^{\tilde{p}_{\text{bot}}} \left(\frac{\partial \widetilde{T}}{\partial \widetilde{t}} - \frac{\partial \widetilde{q_{\text{l}}}}{\partial \widetilde{t}} \cdot \widetilde{L}_{\text{v}} - \frac{\partial \widetilde{q_{\text{i}}}}{\partial \widetilde{t}} \cdot \widetilde{L}_{\text{s}} \right) d\tilde{p} - \widetilde{L}_{\text{v}} \cdot \widetilde{\mathcal{P}}_{\text{rain}} - \widetilde{L}_{\text{s}} \cdot \widetilde{\mathcal{P}}_{\text{snow}}, \tag{B.10}$$

$$\widetilde{m}_{\text{res}} = \int_{\tilde{p}_{\text{top}}}^{\tilde{p}_{\text{bot}}} \left(\frac{\partial \widetilde{q_{\text{v}}}}{\partial \widetilde{t}} + \frac{\partial \widetilde{q_{\text{l}}}}{\partial \widetilde{t}} + \frac{\partial \widetilde{q_{\text{i}}}}{\partial \widetilde{t}} \right) d\tilde{p} + \widetilde{\mathcal{P}}_{\text{rain}} + \widetilde{\mathcal{P}}_{\text{snow}}, \tag{B.11}$$

$$\widetilde{u}_{\text{res}} = \int_{\tilde{p}_{\text{top}}}^{\tilde{p}_{\text{bot}}} \frac{\partial \widetilde{u}}{\partial \widetilde{t}}\, d\tilde{p}, \tag{B.12}$$

$$\widetilde{v}_{\text{res}} = \int_{\tilde{p}_{\text{top}}}^{\tilde{p}_{\text{bot}}} \frac{\partial \widetilde{v}}{\partial \widetilde{t}}\, d\tilde{p}. \tag{B.13}$$

B.2. The Hyperparameter Optimization Search Space and Offline R^2 Scores

Parameter	Search space	Used in "Trade-off"
`encode_dim` e	$\{10k \mid k \in \mathbb{N},\ 1 \leq k \leq 40\}$	280
`hidden_dim` h	$\{10k \mid k \in \mathbb{N},\ 1 \leq k \leq 40\}$	60
`iter_dim` it	$\{100 + 10k \mid k \in \mathbb{N},\ 0 \leq k \leq 80\}$	120
`lstm_layers`	$\{k \mid k \in \mathbb{N},\ 1 \leq k \leq 10\}$	4
`dropout_rate`	{0,0.01,0.02,0.03,0.05,0.07,0.1,0.13, 0.16,0.2,0.25,0.3}	0.02
`learning_rate`	$\{1 \times 10^{-3}, 5 \times 10^{-3}, 6.5 \times 10^{-3}, 1 \times 10^{-4}\}$	1×10^{-3}
`weight_decay`	$\{2 \times 10^{-4}, 1 \times 10^{-2}\}$	1×10^{-2}
`batch_dim` b	{256,512,1024,2048}	256
`scheduler`	{None,`consanh`,`reduce_plat`}	None
`optimizer`	∅	`AdamW`
`early_stopping_patience`	∅	6
`input_dim` i	∅	17
`column_height` l	∅	42
`scalar_out_dim` s	∅	6
`profile_out_dim` p	∅	2

Table B.1.: The parameter search space used for creating Figure 3.9 and the parameter setting for the "Trade-off" model. Additionally, some fixed Hyperparameters are indicated with an empty set as the search set. The scheduler `cosanh` is short for the PyTorch class `CosineAnnealingWarmRestarts` and `reduce_plat` for the class `ReduceLROnPlateau`. The `encode_dim` e, `hidden_dim` h, `iter_dim` it, `batch_dim` b, `input_dim` i, `column_height` l, `scalar_out_dim` s, and `profile_out_dim` p correspond to the dimensions displayed in Figure 3.7. Adapted with permission from Heuer et al. (2025).

α	offline R^2
0	0.896
0.01	0.894
0.1	0.892
0.5	0.884
0.9	0.631

Table B.2.: The overall R^2 scores for five models with different weighting factors of the physics informed loss terms. Adapted with permission from Heuer et al. (2025).

B.3. Additional Figures

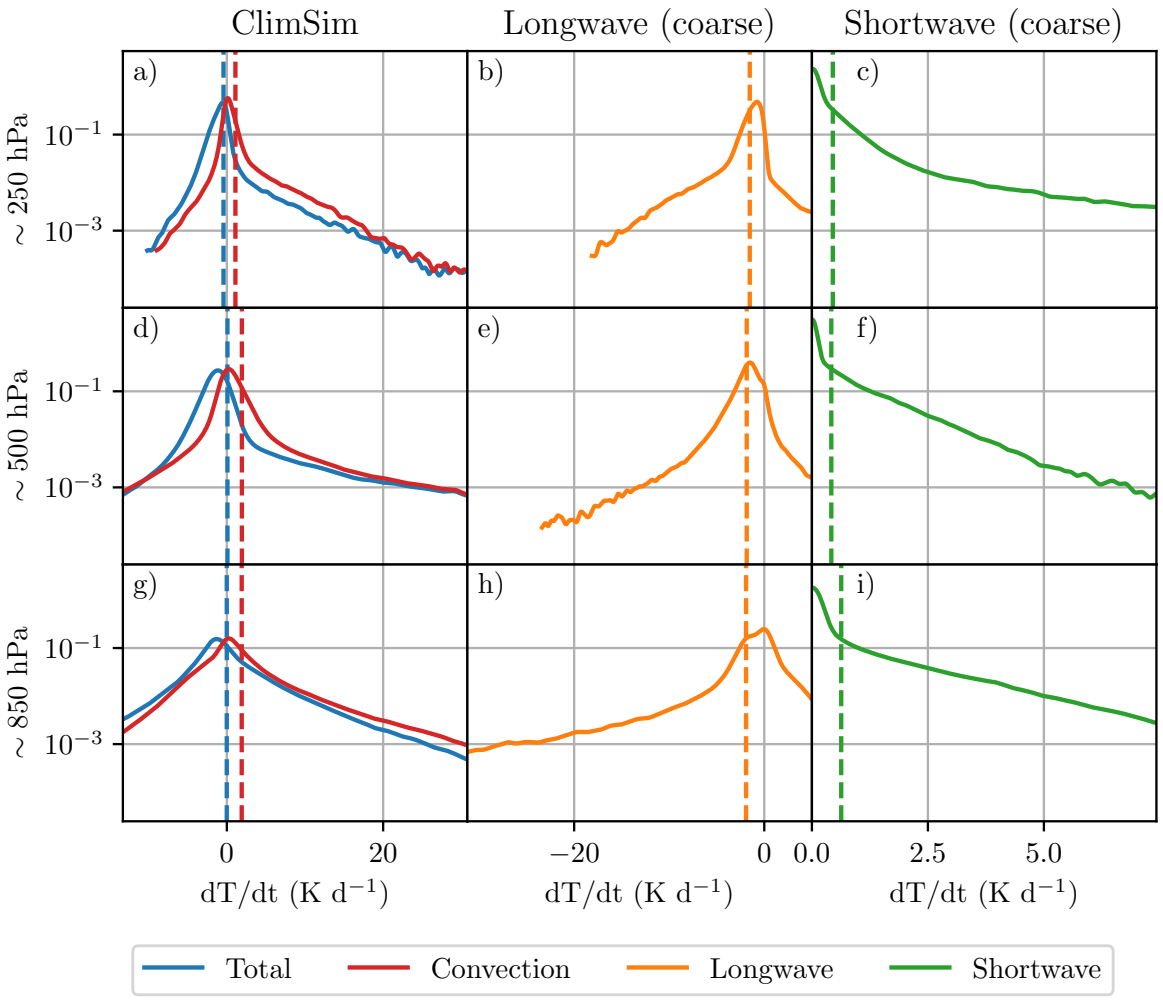

Figure B.1.: For three pressure levels (rows): (a) temperature tendency distributions before (blue, labeled "Total") and after (red, labeled "Convection") subtraction of the tendencies computed with RTE+RRTMGP. These radiative tendencies are decomposed into (b) longwave and (c) shortwave components. Mean values are shown with dashed vertical lines for all distributions. Adapted with permission from Heuer et al. (2025).

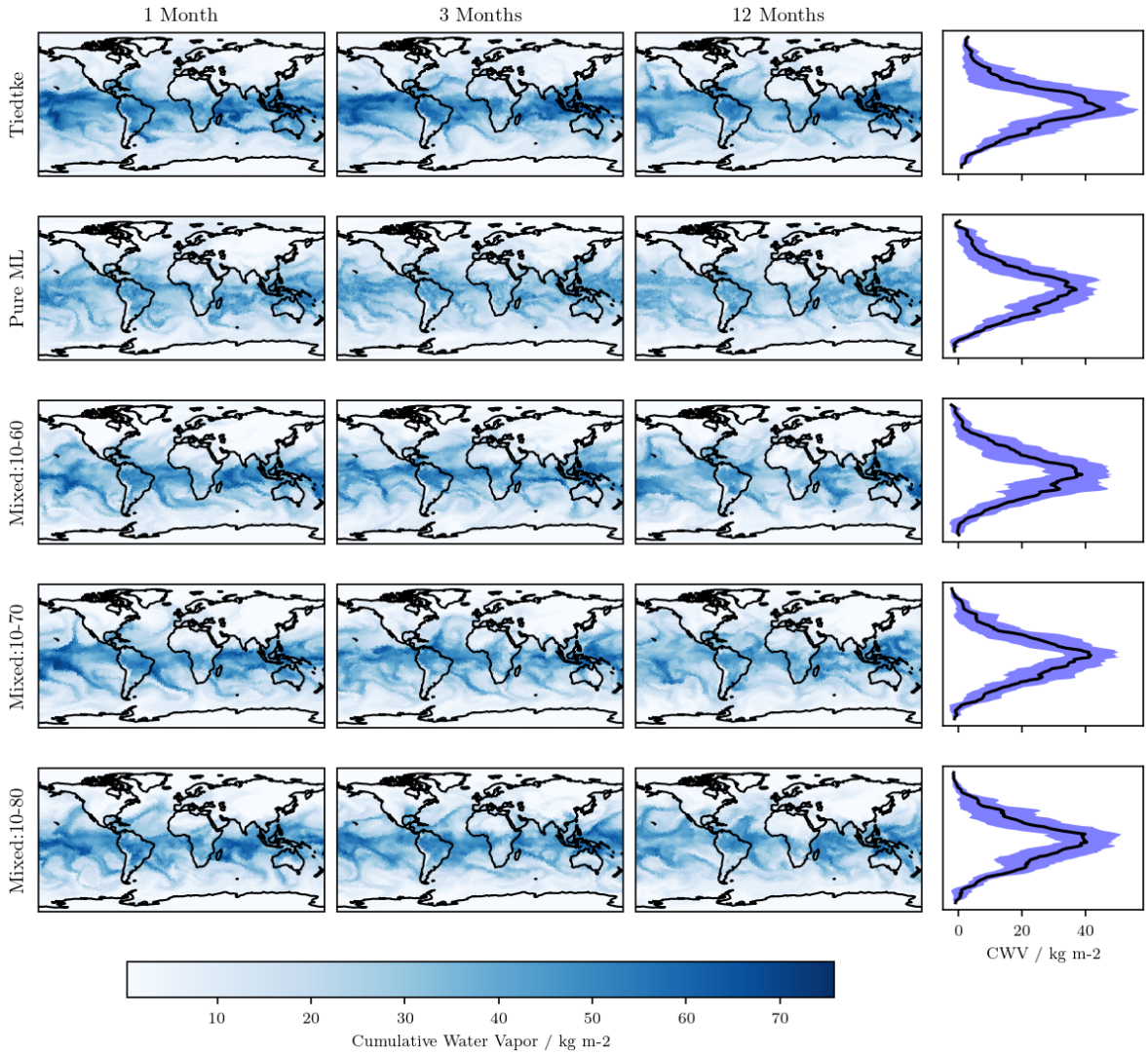

Figure B.2.: The column water vapor for three simulation snapshots after 1 month (first column), 3 months (second column), and 12 months (third column) of integration. The rows correspond to the five different coupled schemes. The last column shows the zonal mean and standard deviation of the CWV for the last shown timestep of every configuration. The y-axis corresponds here to the latitudes of the corresponding row. Adapted with permission from Heuer et al. (2025).

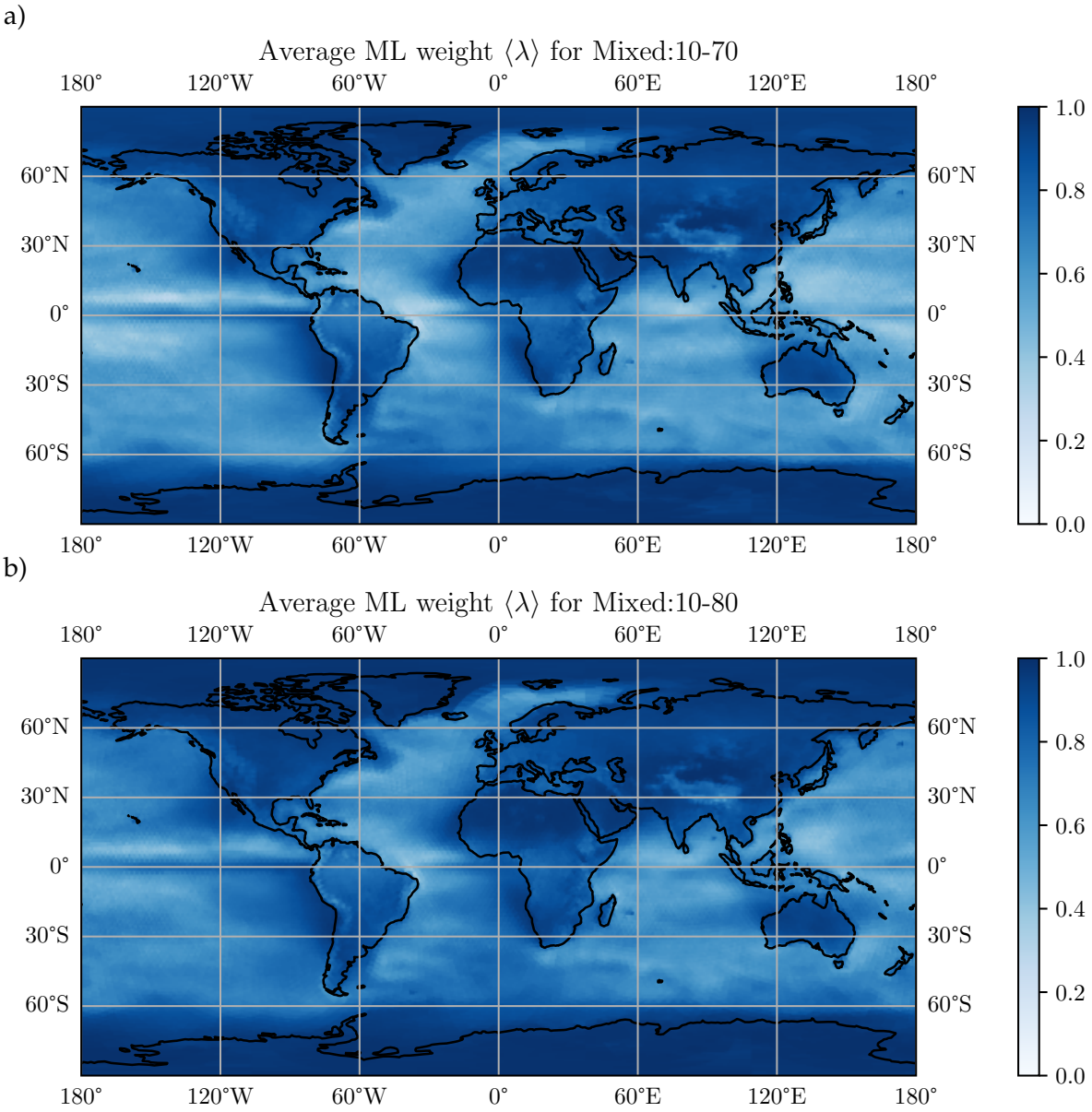

Figure B.3.: The spatial distribution of the temporal average ML weight $\langle\lambda\rangle$ over one year of simulation for the Mixed:10-70 and Mixed:10-80 models with a physics-informed weight $\alpha = 0.1$. The overall time averaged ML weights were $\langle\lambda\rangle \approx 0.71$ and $\langle\lambda\rangle \approx 0.76$, respectively. Adapted with permission from Heuer et al. (2025).

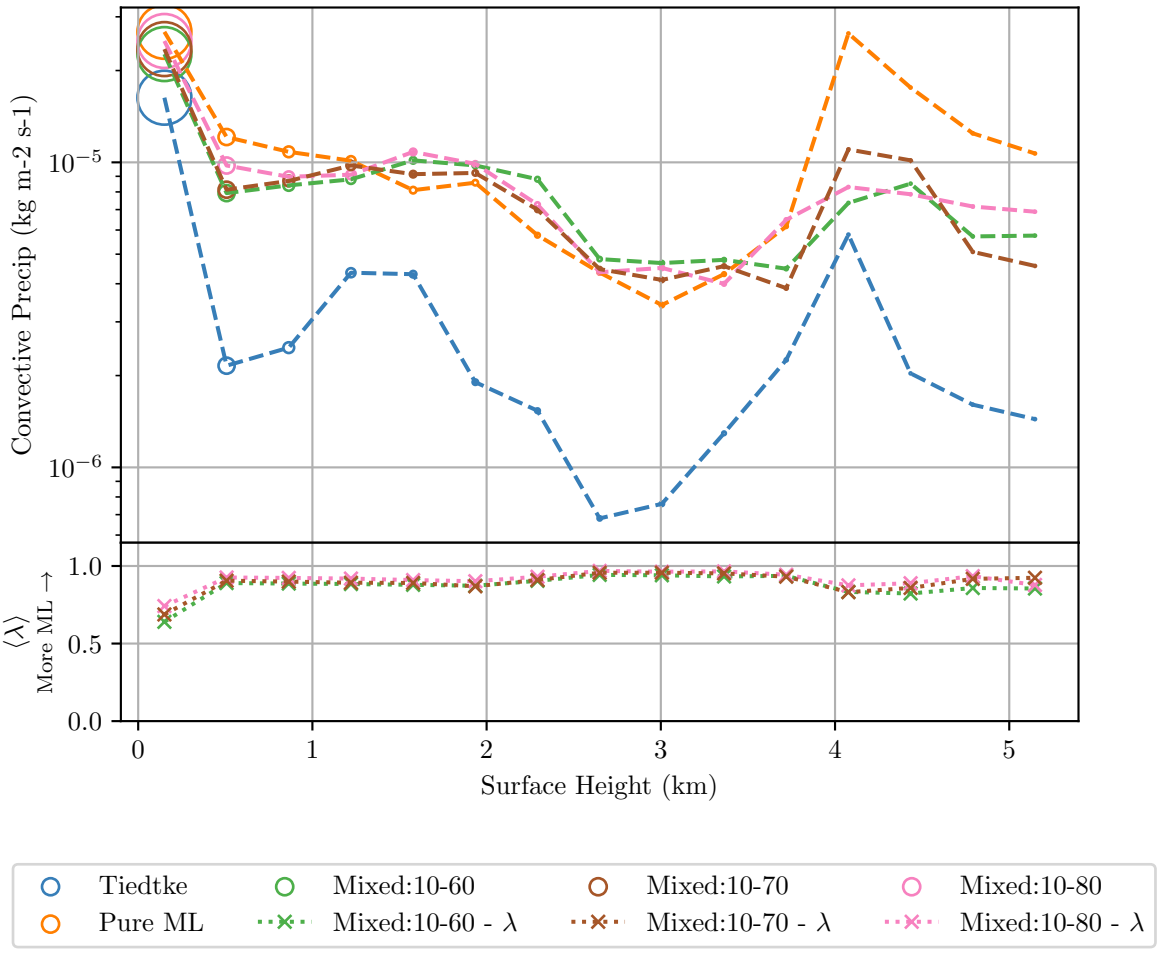

Figure B.4.: Conditionally averaged convective precipitation as a function of the surface height. Circles represent the convective precipitation (circle sizes indicate the number of samples in the respective region). Crosses in the lower plot represent the average ML weight $\langle\lambda\rangle$. Adapted with permission from Heuer et al. (2025).

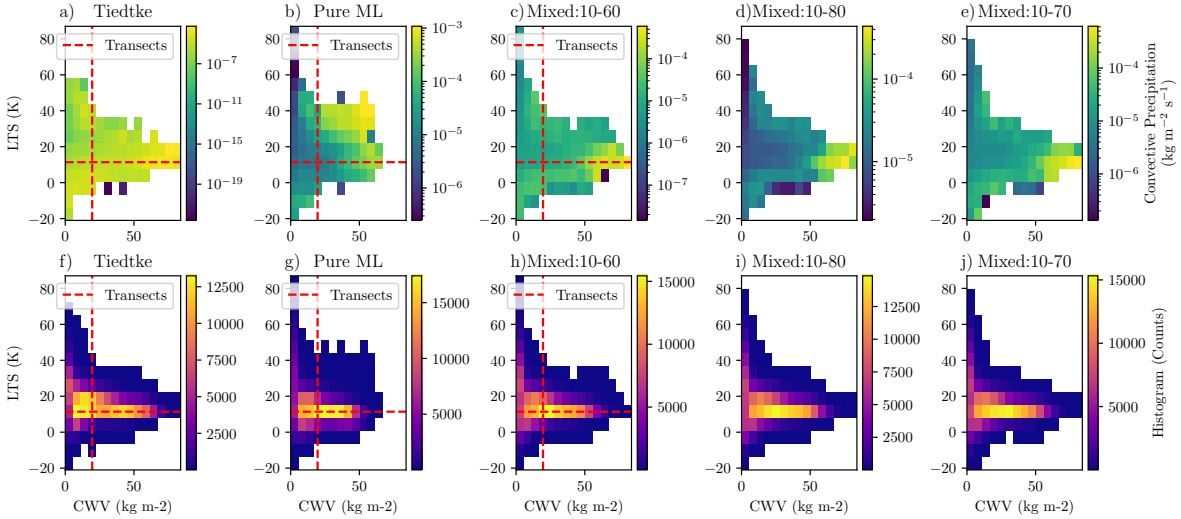

Figure B.5.: 2D histogram of LTS and CWV for 5 different coupled schemes in the top row (a-e). Additionally, the conditionally averaged convective precipitation for each bin above as a function of LTS and CWV is displayed in the lower row (f-j). Adapted with permission from Heuer et al. (2025).

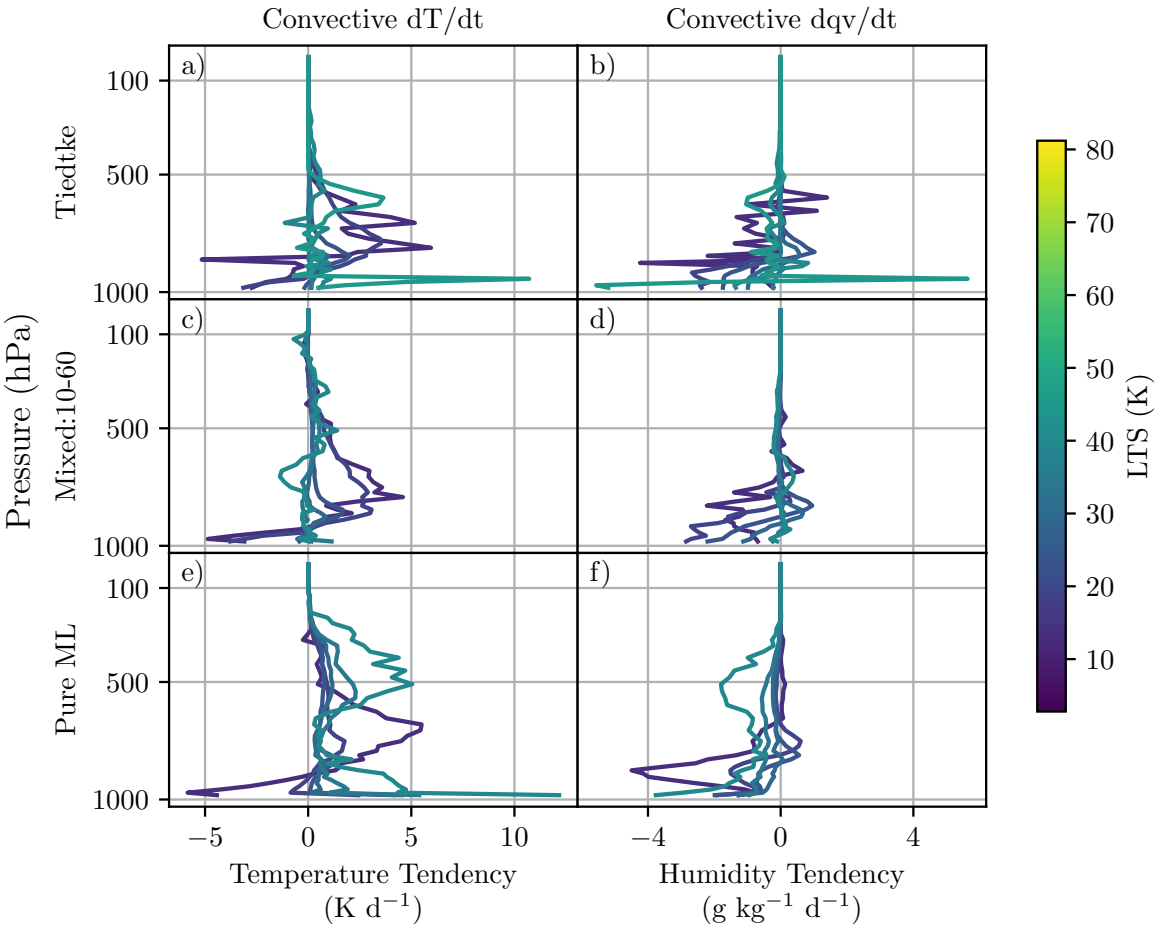

Figure B.6.: Conditional averages of convective heating rates (first column) and moistening rates (second column) as a function of height. The conditioning is based on LTS while we keep the value for the CWV fixed to $CWV = 19.6\,\mathrm{kg/m^2}$. Each row corresponds to a different coupled scheme: (a,b) for Tiedtke, (c,d) for Mixed:10-60, and (e,f) for the pure ML scheme. Conditional averaged curves are only computed for LTS conditions having at least ten samples. Adapted with permission from Heuer et al. (2025).

Area-weighted Mean Bias	Tiedtke	Pure ML	Mixed:10-60
T_{2m} (K)	0.50	1.03	**-0.26**
Precipitation ($\mathrm{mm\,d^{-1}}$)	-0.21	-0.34	**0.14**

Table B.3.: The mean bias for near-surface Temperature and Precipitation corresponding to Figures 5.8 and B.7. Adapted with permission from Heuer et al. (2025).

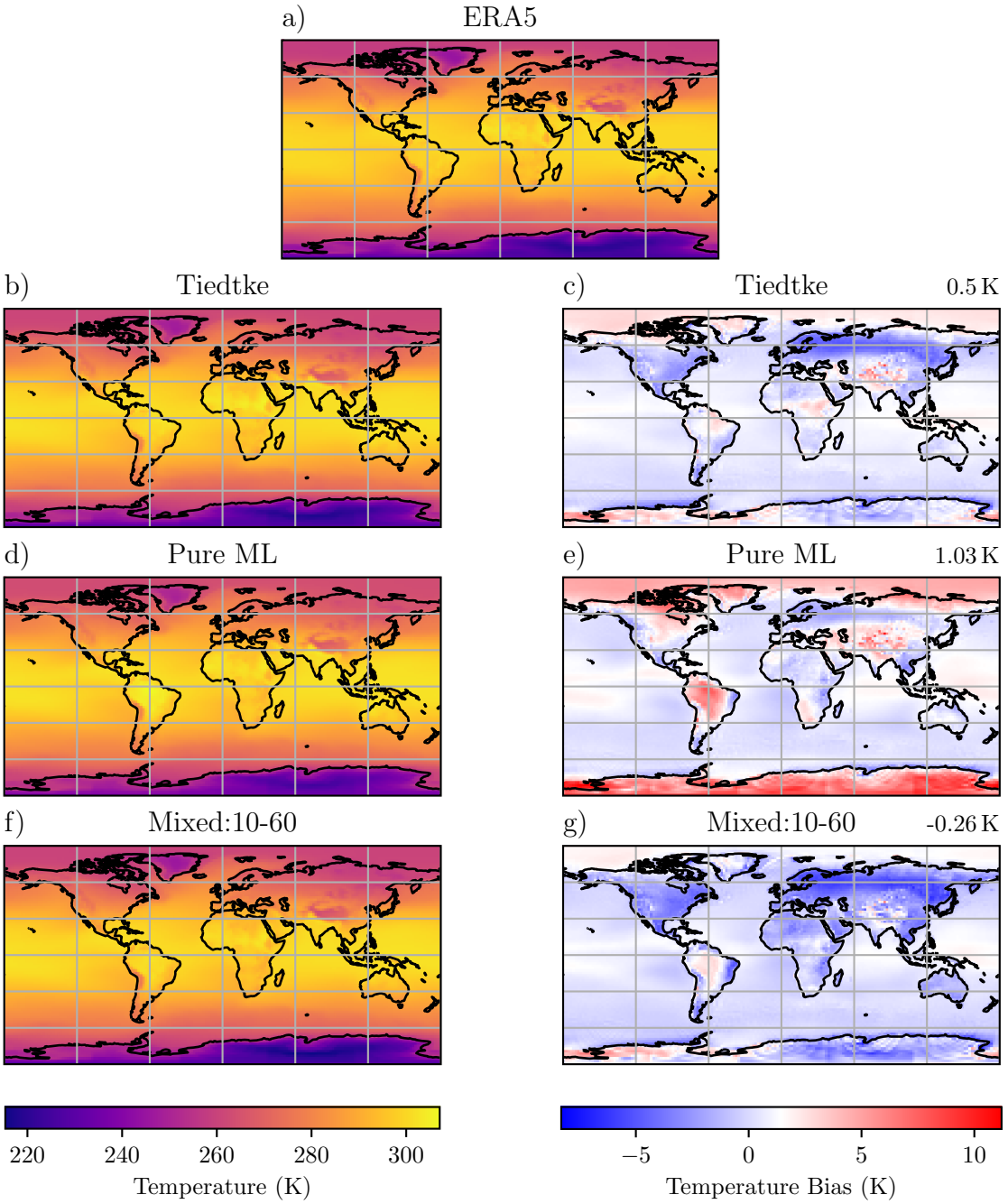

Figure B.7.: Spatial distribution of 20-year averaged near surface temperature T_{2m} for different convection schemes in the left column and the bias with respect to ERA5 in the right column. The first row (a) shows near surface temperature for the ERA5 data, the Tiedtke scheme in the second row (b-c), the pure ML scheme in the third row (d-e), and the Mixed:10-60 scheme in the last row (f-g). In the upper right of each bias plot, the area-weighted mean bias is displayed. Adapted with permission from Heuer et al. (2025).

Glossary

AI artificial intelligence v, vi

AMIP Atmospheric Model Intercomparison Project 35, 83

BiLSTM bidirectional long short-term memory 23, 46, 48, 49, 51, 54, 55, 82, 100, 134, 135

CAPE convective available potential energy 14, 18

CDF cumulative distribution function 41, 52, 134

CL confidence loss 46, 48, 50, 82, 134

ClimSim A Large Multi-scale Dataset and Framework for Hybrid ML-physics Climate Emulation (Yu et al. 2025) v, ix, 33, 35, 44–48, 56, 82, 85, 86, 92–95, 100–102, 134, 138, 139

ClimSim Convection ClimSim dataset (Yu et al. 2025) modified by approximate removing of radiatiative temperature tendencies with RTE+RRTMGP (Pincus et al. 2019, 2023). x, 45–47, 134

CNN convolutional neural network 21, 22, 42, 44, 61, 76, 77, 115, 116

CO_2 carbon dioxide 1, 57

CQE convective quasi-equilibrium 14

CRCP cloud-resolving convection parameterization 10

CWV column water vapor 84, 87–92, 122, 125, 126, 138, 140, 141

DLR German Aerospace Center (Deutsches Zentrum für Luft- und Raumfahrt; DLR) 34, 133

DWD German Weather Service 33, 34

DYAMOND Intercomparison project of global storm-resolving models: the DYnamics of the Atmospheric general circulation Modeled On Non-hydrostatic Domains (DYAMOND) (Stevens et al. 2019b) 6, 8, 81, 133

E3SM-MMF Energy Exascale Earth System Model-Multiscale Modeling Framework v, 10, 11, 35, 44–46, 82, 95, 100, 134

ECHAM6 Sixth generation of the Hamburg version of the ECMWF model (ECHAM6) (Stevens et al. 2013) 34, 35

ECMWF European Centre for Medium-Range Weather Forecasts 8, 15, 18, 47

ECS equilibrium climate sensitivity 1, 57

ERA5 ECMWF Reanalysis v5 27, 46, 47, 84, 128, 134, 138, 141

ESM Earth system model v, 1, 2, 5, 6, 35, 47, 57, 81, 97

ESMValTool ESMValTool is a community diagnostic and performance metrics tool for the evaluation of Earth system models (Andela et al. 2025; Righi et al. 2020) 47, 83

ET extremely randomized tree 20, 43, 60, 100, 114, 117

GBT gradient boosting tree 20, 43, 60–65, 69, 70, 76–78, 100, 105–108, 114, 117, 136, 139

GCM general circulation model 5–12, 15, 16, 25–31, 42, 57, 58, 70, 71, 73, 75–78, 95, 96, 99–101, 133

GME Global Model Europe 34

GPCP Global Precipitation Climatology Project 45–47, 84–86, 93, 94, 98, 101, 134, 138, 139, 143

greySnow 1st place in the LEAP - Atmospheric Physics Using AI (ClimSim) Kaggle competition: Lin et al. (2024) 48, 52, 82

h.p. hypohydrostatic rescaling - increases the horizontal length scale of convection in general circulation models and allows using a coarser horizontal grid spacing to resolve convective motions (Boos et al. 2016; Garner et al. 2007) 30, 31

HPO hyperparameter optimization 44, 59, 60, 62, 65, 115, 118, 134, 135, 140

ICON ICOsahedral Nonhydrostatic v, vi, 2–4, 6, 8, 15, 18, 19, 26, 29, 33–36, 38, 39, 41, 45–48, 51–53, 55, 56, 58, 59, 70–72, 76, 78, 80–84, 91, 93, 95–97, 99, 101, 102, 118, 133, 134, 137, 138

ICON-XPP A new Earth System model configuration framed into the ICON (Giorgetta et al. 2018; Zängl et al. 2015) architecture (XPP stands for eXtended Predictions and Projection) (Müller et al. 2025) 35, 102

ITCZ intertropical convergence zone 1, 10–12, 15, 36, 37, 45, 58, 64, 85, 86, 91, 93, 112

JSBACH The land component of the MPI Earth System Model (Jena Scheme for Biosphere-Atmosphere Coupling in Hamburg (JSBACH)) (Reick et al. 2021) 35

LFC level of free convection 18

LSTM long short-term memory 23, 54

LTS lower tropospheric stability 87–92, 125, 126, 138, 140, 141

MAC multiply-accumulate operations 54, 55, 135

MJO Madden-Julian oscillation 8, 11, 15

ML machine learning v, vi, 2–5, 19, 24–33, 35, 37–39, 42, 43, 45, 47, 49, 51, 53, 55, 56, 58, 59, 62, 65, 70–73, 75–103, 113, 117, 123, 124, 126–128, 133, 135, 137–141

MLP multilayer perceptron 21, 22, 42, 58, 61, 65, 76, 77, 110, 115–117, 133, 139

MMF multiscale modeling framework 2, 4, 5, 7, 10–12, 25, 29, 30, 35, 100, 103

MMM multi-model mean 85, 86, 93, 94, 138, 139

NARVAL Flight campaign with accompanying storm-resolving regional simulations (Klocke et al. 2017). Short for Next-Generation Aircraft Remote Sensing for Validation (or broader metaphor according to Stevens et al. (2019a)) 33, 36, 37, 41, 72, 73, 81, 99, 112, 134, 136, 140

NeuralGCM A fully differentiable hybrid GCM of Earth's atmosphere (Kochkov et al. 2024) 27

NextGEMS Next Generation Earth Modelling Systems 8, 81

NICAM Nonhydrostatic Icosahedral Atmospheric Model 8

NN neural network 21, 24, 25, 27, 45, 47, 48, 50, 52, 54, 55, 71, 77, 82, 87, 100, 101, 135

NWP numerical weather prediction 33–36

R^2 coefficient of determination 59–65, 76–78, 135, 136

ResNet residual neural network 22, 42, 43, 61, 76, 77, 117

RF random forest 20, 42, 43, 58, 60, 69, 100, 114, 117

RMSE root mean square error 59, 64, 75, 78, 84–86, 93, 94, 108, 111, 138, 139, 143

RNN recurrent neural network 23, 42

RTE+RRTMGP Radiative Transfer for Energetics + General circulation model applications—Parallel (Pincus et al. 2019, 2023). 45, 46, 95, 121, 134, 140

SAM System for Atmospheric Modeling 8, 35

SCM single column model 29, 30

SHAP SHapley Additive exPlanations v, 24, 65–70, 77, 78, 100, 109, 110, 136, 139

SPCAM Superparameterized Community Atmosphere Model 10, 11

SRM storm-resolving model 2, 5, 7–12, 28–31, 35, 44, 45, 82, 95, 100, 103

SST sea surface temperature 8, 83

U-Net A convolutional neural network with a U-shaped architecture (Ronneberger et al. 2015) 23, 42–44, 59–80, 100, 105–109, 112, 113, 115, 117, 134, 136, 137, 139, 140

XAI explainable artificial intelligence 24, 26, 58, 59, 100

YA HB MS EK 5th place in the LEAP - Atmospheric Physics Using AI (ClimSim) Kaggle competition: Lin et al. (2024) 48, 49, 54, 100, 135

Z Lab 2nd place in the LEAP - Atmospheric Physics Using AI (ClimSim) Kaggle competition: Lin et al. (2024) 52

List of Figures

2.1. Comparison of two horizontal resolutions (R2B4 and R2B10) of an atmospheric state modeled by the ICON model (Giorgetta et al. 2018; Zängl et al. 2015). Shown is a snapshot of the cloud cover distribution from the 05.02.2020. Data is taken from the DYAMOND intercopmarison project (Stevens et al. 2019b). Background image: NASA Earth Observatory (https://neo.gsfc.nasa.gov/view.php?datasetId=BlueMarbleNG). 6

2.2. Visualization of an ensemble of convective cumulus clouds (left) and the corresponding bulk mass-flux scheme representation (right). Updrafts and downdrafts are visualized as orange and blue arrows, respectively. Entrainment and detrainment are indicated by the violet arrows. The grid box is assumed to be much larger than the area depicted, with the environment subsidence region occupying most of it. 14

2.3. Visualization of an MLP with two hidden layers. The network has three input and two output neurons. Connection weights are represented by the W_i matrices. 22

2.4. Simplified visualization of one timestep with a coupled ML-based parameterization in a GCM for a representative atmospheric column. The column state at time t_i is called s_i. The tendencies computed by the ML scheme and other conventional parameterization are called $\frac{ds_{i,\text{ml}}}{dt}$ and $\frac{ds_{i,\text{phy}}}{dt}$, respectively. The actual implementation in ICON is more intricate, see Zängl et al. (2015) for details on, e.g., the distinction between *fast-physics* and *slow-physics* and *operator splitting*. Earth background image: NASA Earth Observatory (https://neo.gsfc.nasa.gov/view.php?datasetId=BlueMarbleNG). 26

3.1. The icosahedral grid structure of the ICON model with an R2B3 grid resolution, translating to a grid spacing of approximately 320 km. The chosen ortographic projection is centered at the location of DLR in Oberpfaffenhofen. Background image: NASA Earth Observatory (https://neo.gsfc.nasa.gov/view.php?datasetId=BlueMarbleNG). 34

3.2. Average number of high-resolution convective cells per displayed low-resolution column and time frame as defined in Equation (3.4) in the studied tropical Atlantic region over the entire considered period of time. In the west the coastline of South America and some Caribbean islands can be seen and in the east the coastline of Africa. The low resolution grid has an approximate horizontal resolution of $\Delta x \approx 80\,\text{km}$. Excluded columns are marked in grey. Adapted with permission from Heuer et al. (2024). 36
3.3. Summary of preprocessing steps. Starting from the original data, first the subgrid fluxes as well as 2D outputs, such as precipitation, were computed. After this, the data was coarse-grained and filtered for active convection. As a final preprocessing step, the data was rescaled and normalized. Adapted with permission from Heuer et al. (2024). 37
3.4. Probability distribution of convectively classified cells over altitude (green large dots) in the high-resolution data with ICON over the NARVAL region. The orange dashed line shows the CDF and the black dashed line represents the height up to which 99.9 % of the convective cells are found. The bottom scale corresponds to the probability and the top to the cumulative distribution. Adapted with permission from Heuer et al. (2024). 41
3.5. Visualization of the used U-Net architecture. The abbreviations DCB, Conv., Transp. Conv., and BN stand for double convolution block, convolutional layer, transpose convolutional layer, and batch normalization layer, respectively. The dotted lines mark the possibility for more blocks depending on the result of the HPO. The horizontal lines indicate skip connections. In the lower right of the figure, a more detailed visualization of the double convolution block is given. Adapted with permission from Heuer et al. (2024). 44
3.6. Overall training and evaluation pipeline of our hybrid model. x and y represent inputs and outputs of the ClimSim dataset, based on the E3SM-MMF model. $\dot{T}_{tot}$ is the total temperature tendency, and RTE+RRTMGP the ICON radiation scheme. The ClimSim dataset is first modified to separate radiative and convective subgrid tendencies, forming a new dataset, "ClimSim Convection". Afterward, we trained a BiLSTM model including a CL. Using CL, this model is mixed with the conventional "Tiedtke" cumulus convection scheme to predict convective tendencies as well as precipitation. In the mixing process, λ represents the fraction provided by the BiLSTM and $1-\lambda$ is the fraction from the conventional "Tiedtke" scheme, respectively. This mixed scheme predicts the tendencies due to convection in temperature $\dot{T}_{conv}$, water vapor, cloud liquid water, cloud ice ($\dot{q}_{\alpha,conv}$, $\alpha = v, l, i$), zonal wind $\dot{u}_{conv}$, and meridional wind $\dot{v}_{conv}$. Finally, we coupled the mixed scheme with the ICON model and evaluate these runs' emergent statistics with respect to observational datasets, including ERA5 and GPCP. Adapted with permission from Heuer et al. (2025). 46

3.7. The BiLSTM architecture developed by the 5th place Kaggle competition winner "YA HB MS EK", and used in the work presented in this article. Tensor dimensions are visualized in the lower right corner of the individual layers. The tensor dimensions shown in the figure are the batch dimension b, the column height level dimension l, the input dimension i, the encoding dimension e, hidden dimension h, iter dimension it, output scalar dimension s, and the output profile dimension p. In the blue-marked layers, the horizontal dotted lines indicate operations restricted to the last dimension, thereby preserving "vertical locality". Adapted with permission from Heuer et al. (2025). 49

3.8. ML weight λ as function of the predicted error percentile level. The tuning parameters p_0 and p_1 (here 20 % and 60 %) are marked by dashed and dotted lines, respectively. In blue and with slanted hatching, the area with $\lambda = 1$ (pure ML) is shown. $\lambda = 0$ (pure Tiedtke) is shown in orange and with horizontal hatching. Adapted with permission from Heuer et al. (2025). 53

3.9. Offline skill-complexity plane for various combinations of nine chosen hyperparameters of the BiLSTM on a smaller subset of the dataset with 3 million training and 1.5 million validation samples. The red dashed line shows the Pareto Front between the coefficient of determination R^2 and the number of MACs. The highlighted NN is selected for the remainder of this study because it strikes a suitable balance between skill and computational performance. Adapted with permission from Heuer et al. (2025). 54

4.1. R^2 on a test set for different types of models. All models were hyperparameter-optimized, and the best models were then trained on the whole data set. The deep learning methods are displayed on the left of the green dashed line and the non-deep learning methods on the right of it. The inset in the top right shows a zoomed-in version of the R^2 for the deep learning models. Adapted with permission from Heuer et al. (2024). 61

4.2. Root mean squared error during HPO on the validation set of the four different deep learning methods. The straight thick lines correspond to the median of the HPO ensemble, the shaded areas are drawn in between the first and third quartile. Additionally, ten realizations for each DL method are shown in similar colors. The legend shows the minimum of the validation loss for each of the methods. The scheduler of the HPO filters badly performing runs after 30 and 60 epochs, causing the steps in the profiles. For this task we used the `AsyncHyperBandScheduler` (Li et al. 2018) of the `Ray Tune` library (Liaw et al. 2018). Adapted with permission from Heuer et al. (2024). 62

4.3. Scatter plot for the subgrid fluxes of a) zonal, b) meridional momentum, c) liquid/ice water static energy, and d) specific humidity. Data for the U-Net is shown in green, for the GBT in blue, and the diagonal is marked by a dotted line. The Pearson correlation coefficient r between the true and the predicted subgrid flux is noted in the lower right corner of each plot for both U-Net and GBT. Adapted with permission from Heuer et al. (2024). 63
4.4. Average R^2 profile for all subgrid flux variables for a) the U-Net and b) GBT model. Data points where the GBT model actually has a higher R^2 than the U-Net are additionally marked by a red circle. Adapted with permission from Heuer et al. (2024). 64
4.5. Feature importances (i.e., the mean absolute values of the calculated SHAP values) of input variables for a) the full U-Net model, and b) the ablated (without qr,qs) U-Net model. The mean feature importance is visualized by the height of the bar, and the standard deviation over five different computations by the errorbars. The x-axis shows different height levels for each variable, increasing from left to right. Vertical lines separate the variables. The integrated fraction of feature importances over all vertical levels is written above each variable range. Adapted with permission from Heuer et al. (2024). 66
4.6. Ensemble mean of weighted SHAP values aggregated according to Equation (4.2) for the U-Net. The variables q_r, q_s were ablated. The height level for each variable is increasing from left to right / from bottom to top. The feature importance depicted in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. Insets a), b), and c) show a zoom into the plot for three specific variable pairs, the colors indicate which inset corresponds to which part of the large plot. Adapted with permission from Heuer et al. (2024). 68
4.7. Ensemble mean of weighted SHAP values aggregated according to Equation (4.2) for the GBT model. The variables q_r, q_s were ablated. The feature importance depicted in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. Adapted with permission from Heuer et al. (2024). 69
4.8. The precipitation distributions of the first two weeks over the tropics for the three simulations starting on 01.02.1979 for the full U-Net (configuration 4) in grey/dark green, the conventional cumulus scheme in green, and the ablated U-Net (configuration 3) in orange. Also, the precipitation distribution for the high-resolution data set (NARVAL) is displayed in blue. The 99.99th percentiles of each data set are marked by dashed lines in the corresponding color. Adapted with permission from Heuer et al. (2024). 72

4.9. The stability of the ablated vs. the full U-Net in form of a time series of the global mean air temperature on 2 m height over 180 days. For each defined configuration, the ten realizations are drawn in orange, purple, blue and green colors, respectively. Solely for the full U-Net coupled globally (Config 1), a second y-axis (also in orange) on the right side of the plot is introduced as this simulations shows a much higher reduction in 2m Temperature. To make this clearer, arrows are indicating the corresponding y-axis for each ensemble. An inset provides a close-up of the first 24 hours of the dynamics of configurations 1 and 2: the simulations with the full U-Net quickly become unstable. The data displayed in the inset has been saved with an output frequency of six minutes as opposed to the more stable simulations with an output frequency of six hours in the main plot. For all of the data except the inset, a rolling mean over 24 h was applied. Additionally, multi-model means over configurations 1, 3, 4, and the reference ensemble, respectively, are drawn as light green/yellow/violet/red-brown dashed lines. These colors are chosen as the complementary colors of the respective ensemble members and are marked in the legend as the second lower dashed line for each ensemble. For configuration 1, model blow-ups are marked by red crosses as to not obscure the other lines. Adapted with permission from Heuer et al. (2024). 74

4.10. The vertically integrated water vapor for three simulation snapshot with convection parameterized with the conventional physical cumulus convection scheme of the ICON model as a reference (ref), 1.) by the ablated U-Net, 2.) by the full U-Net, 3.) by the ablated U-Net applied only in the tropics, 4.) by the full U-Net applied only in the tropics. For row 3.) and 4.) the domain where the ML schemes are applied are marked by orange dashed lines. The last column shows the zonal mean and standard deviation of the vertically integrated Water Vapor (VIWV) for the last shown date (1979-03-01) of every configuration except the unstable one. The y-axis corresponds here to the latitudes of the corresponding row. Adapted with permission from Heuer et al. (2024). 80

5.1. Evaluation scores for coupled ICON runs, each dot represents a one-year long coupled ICON run at a horizontal resolution of 158 km × 158 km. The runs are colored according to their physics-informed loss weight α for the coupled ML schemes and the conventional Tiedtke scheme is colored in blue. Within each coloring group, the models have different values for p_0 and p_1. Panel (a) shows the spatial R^2 score of precipitation with respect to the observational dataset GPCP versus the R^2 score of CWV with respect to the mean of multiple observation sets as explained in Section 3.5.3. Panel (b) displays the R^2 score of near-surface (2 m) air temperature with respect to ERA5 versus the RMSE of zonal mean precipitation with respect to GPCP. In both panels, the Pareto front between the two skill metrics is marked with a dashed red line. Adapted with permission from Heuer et al. (2025). 84
5.2. Zonal mean precipitation in one-year-long runs (a) and precipitation distribution (b) for the pure ML scheme, the Tiedtke scheme, a mixed scheme (Mixed:10-60), and references; GPCP observations, CMIP6 MMM, and ClimSim for the mean precipitation, and ClimSim for the precipitation extremes. Adapted with permission from Heuer et al. (2025). 85
5.3. Mean absolute enthalpy residual (blue, left axis) and average ML weight λ during the one-year long online integration (purple, right axis) for a selection of tested models. The ten most-conserving (left in the plot) and least-conserving (right) models in terms of enthalpy conservation are displayed. In between the black dotted lines every 8^{th} model is displayed so that the figure is still readable. Additionally, models which are used for a deeper analysis in this section are marked by bold labels. Adapted with permission from Heuer et al. (2025). . . . 88
5.4. The spatial distribution of the temporally-averaged ML weight $\langle\lambda\rangle$ over one year of simulation for the Mixed:10-60 model with a physics-informed weight $\alpha = 0.1$. The overall time averaged ML weight was $\langle\lambda\rangle \approx 0.67$ for the coupled run. Adapted with permission from Heuer et al. (2025). 89
5.5. Conditionally averaged convective precipitation (top row) and average ML weight $\langle\lambda\rangle$ (lower row) as a function of CWV (a), LTS (b), and absolute latitude (c). Circles represent the convective precipitation (circle sizes indicate the number of samples in the respective region) and crosses the average ML weight $\langle\lambda\rangle$. All plots within one row share the same y-axis scale. Adapted with permission from Heuer et al. (2025). 90
5.6. Conditional averages of convective heating rates (first column) and moistening rates (second column) as a function of height. The conditioning is based on CWV while we keep the value for the LTS fixed to LTS = 11.4 K. Each row corresponds to a different coupled scheme: (a,b) for Tiedtke, (c,d) for Mixed:10-60, and (e,f) for the pure ML scheme. Conditional averaged curves are only computed for CWV conditions having at least ten samples. Adapted with permission from Heuer et al. (2025). 92

5.7. Zonal mean precipitation evaluated over twenty years for the observational dataset (GPCP), the Tiedtke scheme, the pure ML scheme, the Mixed:10-60 scheme, the Mixed:10-80 scheme, the CMIP6 MMM, and the ClimSim dataset. For ClimSim, the zonal mean precipitation is evaluated over its available 10-year simulation period. Adapted with permission from Heuer et al. (2025). 93
5.8. The spatial distribution of 20-year averaged precipitation for different convection schemes in the left column and the bias with respect to GPCP in the right column. The first row (a) shows precipitation for the GPCP data, the Tiedtke scheme in the second row (b-c), the pure ML scheme in the third row (d-e), and the Mixed:10-60 scheme in the last row (f-g). In the upper right of each bias plot, the area-weighted mean bias is displayed. Adapted with permission from Heuer et al. (2025). 98

A.1. Scatterplots for the subgrid fluxes of the four remaining tracer species not shown in the main document. Data for the U-Net is shown in green, for the GBT in blue, and the diagonal is marked by a dotted line. The Pearson correlation coefficient is written in the lower right of each plot for both U-Net and GBT. Adapted with permission from Heuer et al. (2024). 106
A.2. Distribution of the true precipitation and the predictions of U-Net and GBT. The 99th and the 99.9th percentile of the true precipitation are marked by the dashed and the dotted line, respectively. The precipitation R^2 scores are 0.897 for the U-Net and 0.860 for the GBT. Adapted with permission from Heuer et al. (2024). 107
A.3. Spatial distribution of the by the variance weighted RMSE in the studied region. The top plot shows the data for the U-Net and on the bottom the result of the GBT model is displayed. Adapted with permission from Heuer et al. (2024). . . 108
A.4. Ensemble mean of weighted SHAP values for the non-ablated U-Net model. The feature importance shown in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. The Insets a), b), and c) show a more detailed version of three specific variable pairs, the colors indicate which inset corresponds to which part of the large plot. Adapted with permission from Heuer et al. (2024). 109
A.5. Ensemble mean of weighted SHAP values for the non-ablated MLP model. The feature importance shown in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. The Insets a), b), and c) show a more detailed version of three specific variable pairs, the colors indicate which inset corresponds to which part of the large plot. Adapted with permission from Heuer et al. (2024). 110
A.6. Complexity, measured by the number of parameters, is plotted against the RMSE on the test set. The Pareto frontier is visualized as the blue dotted line. Adapted with permission from Heuer et al. (2024). 111

A.7. The monthly mean spatial precipitation distribution of the NARVAL data and the simulations with the convectional cumulus scheme, the ablated U-Net, and the full U-Net. For the latter three simulations the first four month of the simulations are shown. All data displayed is taken from the first ensemble member each. Adapted with permission from Heuer et al. (2025). 112
A.8. Visualization of training procedure. The preprocessed data set is split 80 %/10 %/10 % into training/validation/test data sets. The models are fit to the training data set, and validated with the validation set in the HPO. For the non-deep learning methods only 5 days are used for the HPO. The resulting hyperparameter-optimized models are then retrained on the full training and validation sets and evaluated mainly based on the coefficient of determination using the test data set. Adapted with permission from Heuer et al. (2025). . . . 115
B.1. For three pressure levels (rows): (a) temperature tendency distributions before (blue, labeled "Total") and after (red, labeled "Convection") subtraction of the tendencies computed with RTE+RRTMGP. These radiative tendencies are decomposed into (b) longwave and (c) shortwave components. Mean values are shown with dashed vertical lines for all distributions. Adapted with permission from Heuer et al. (2025). 121
B.2. The column water vapor for three simulation snapshots after 1 month (first column), 3 months (second column), and 12 months (third column) of integration. The rows correspond to the five different coupled schemes. The last column shows the zonal mean and standard deviation of the CWV for the last shown timestep of every configuration. The y-axis corresponds here to the latitudes of the corresponding row. Adapted with permission from Heuer et al. (2025). . . . 122
B.3. The spatial distribution of the temporal average ML weight $\langle\lambda\rangle$ over one year of simulation for the Mixed:10-70 and Mixed:10-80 models with a physics-informed weight $\alpha = 0.1$. The overall time averaged ML weights were $\langle\lambda\rangle \approx 0.71$ and $\langle\lambda\rangle \approx 0.76$, respectively. Adapted with permission from Heuer et al. (2025). . . 123
B.4. Conditionally averaged convective precipitation as a function of the surface height. Circles represent the convective precipitation (circle sizes indicate the number of samples in the respective region). Crosses in the lower plot represent the average ML weight $\langle\lambda\rangle$. Adapted with permission from Heuer et al. (2025). 124
B.5. 2D histogram of LTS and CWV for 5 different coupled schemes in the top row (a-e). Additionally, the conditionally averaged convective precipitation for each bin above as a function of LTS and CWV is displayed in the lower row (f-j). Adapted with permission from Heuer et al. (2025). 125

B.6. Conditional averages of convective heating rates (first column) and moistening rates (second column) as a function of height. The conditioning is based on LTS while we keep the value for the CWV fixed to $CWV = 19.6\,\mathrm{kg/m^2}$. Each row corresponds to a different coupled scheme: (a,b) for Tiedtke, (c,d) for Mixed:10-60, and (e,f) for the pure ML scheme. Conditional averaged curves are only computed for LTS conditions having at least ten samples. Adapted with permission from Heuer et al. (2025). 126
B.7. Spatial distribution of 20-year averaged near surface temperature T_{2m} for different convection schemes in the left column and the bias with respect to ERA5 in the right column. The first row (a) shows near surface temperature for the ERA5 data, the Tiedtke scheme in the second row (b-c), the pure ML scheme in the third row (d-e), and the Mixed:10-60 scheme in the last row (f-g). In the upper right of each bias plot, the area-weighted mean bias is displayed. Adapted with permission from Heuer et al. (2025). 128

List of Tables

5.1. The tropical precipitation asymmetry index A_P and the equatorial precipitation index E_P, and their biases, as well as the RMSE, with respect to GPCP for the data shown in Figure 5.2 (a). Adapted with permission from Heuer et al. (2025). 86

5.2. The tropical precipitation asymmetry index A_P and the equatorial precipitation index E_P, and their biases with respect to GPCP for the data shown in Figure 5.7. Adapted with permission from Heuer et al. (2025). 94

A.1. Number of Trainable Parameters Used in the Various Deep Learning Models. Adapted with permission from Heuer et al. (2024). 115

A.2. Common Parameters for the Hyperparameter Optimization of the Various Deep Learning Architectures. Adapted with permission from Heuer et al. (2024). . . 116

B.1. The parameter search space used for creating Figure 3.9 and the parameter setting for the "Trade-off" model. Additionally, some fixed Hyperparameters are indicated with an empty set as the search set. The scheduler `cosanh` is short for the PyTorch class `CosineAnnealingWarmRestarts` and `reduce_plat` for the class `ReduceLROnPlateau`. The `encode_dim` e, `hidden_dim` h, `iter_dim` it, `batch_dim` b, `input_dim` i, `column_height` l, `scalar_out_dim` s, and `profile_out_dim` p correspond to the dimensions displayed in Figure 3.7. Adapted with permission from Heuer et al. (2025). 120

B.2. The overall R^2 scores for five models with different weighting factors of the physics informed loss terms. Adapted with permission from Heuer et al. (2025). 120

B.3. The mean bias for near-surface Temperature and Precipitation corresponding to Figures 5.8 and B.7. Adapted with permission from Heuer et al. (2025). 127

References

Adam, O., Schneider, T., Brient, F., & Bischoff, T. (2016). Relation of the Double-ITCZ Bias to the Atmospheric Energy Budget in Climate Models. *Geophysical Research Letters*, *43*(14), 7670–7677. https://doi.org/10.1002/2016GL069465

Adler, R. F., Sapiano, M. R. P., Huffman, G. J., Wang, J.-J., Gu, G., Bolvin, D., Chiu, L., Schneider, U., Becker, A., Nelkin, E., Xie, P., Ferraro, R., & Shin, D.-B. (2018). The Global Precipitation Climatology Project (GPCP) Monthly Analysis (New Version 2.3) and a Review of 2017 Global Precipitation. *Atmosphere*, *9*(4). https://doi.org/10.3390/atmos9040138

Ahn, M.-S., & Kang, I.-S. (2018). A Practical Approach to Scale-Adaptive Deep Convection in a GCM by Controlling the Cumulus Base Mass Flux. *npj Climate and Atmospheric Science*, *1*(1), 13. https://doi.org/10.1038/s41612-018-0021-0

Alet, F., Price, I., El-Kadi, A., Masters, D., Markou, S., Andersson, T. R., Stott, J., Lam, R., Willson, M., Sanchez-Gonzalez, A., et al. (2025). Skillful Joint Probabilistic Weather Forecasting from Marginals. *arXiv preprint arXiv:2506.10772*.

Alsabti, K., Ranka, S., & Singh, V. (1998). CLOUDS: A Decision Tree Classifier for Large Datasets. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining*, 2–8.

Amiramjadi, M., Plougonven, R., Mohebalhojeh, A. R., & Mirzaei, M. (2023). Using Machine Learning to Estimate Nonorographic Gravity Wave Characteristics at Source Levels. *Journal of the Atmospheric Sciences*, *80*(2), 419–440.

Anber, U., Gentine, P., Wang, S., & Sobel, A. H. (2015). Fog and Rain in the Amazon. *Proceedings of the National Academy of Sciences*, *112*(37), 11473–11477. https://doi.org/10.1073/pnas.1505077112

Andela, B., Broetz, B., de Mora, L., Drost, N., Eyring, V., Koldunov, N., Lauer, A., Mueller, B., Predoi, V., Righi, M., Schlund, M., Vegas-Regidor, J., Zimmermann, K., Adeniyi, K., Arnone, E., Bellprat, O., Berg, P., Billows, C., Bock, L., Bodas-Salcedo, A., Caron, L.-P., Carvalhais, N., Cionni, I., Cortesi, N., Corti, S., Crezee, B., Davin, E. L., Davini, P., Deser, C., Diblen, F., Docquier, D., Dreyer, L., Ehbrecht, C., Earnshaw, P., Geddes, T., Gier, B., Gillett, E., Gonzalez-Reviriego, N., Goodman, P., Hagemann, S., Hardacre, C., von Hardenberg, J., Hassler, B., **Heuer**, **H.**, Hogan, E., Hunter, A., Kadow, C., Kindermann, S., Koirala, S., Kuehbacher, B., Lledó, L., Lejeune, Q., Lembo, V., Little, B., Loosveldt-Tomas, S., Lorenz, R., Lovato, T., Lucarini, V., Malinina, E., Massonnet, F., Mohr, C. W., Amarjiit, P., Pérez-Zanón, N., Phillips, A., Proft, M., Russell, J., Sandstad, M., Sellar, A., Senftleben, D., Serva, F., Sillmann, J., Stacke, T., Swaminathan, R., Tomkins, K., Torralba, V., Weigel, K., Sarauer, E., Roberts, C., Kalverla, P., Alidoost, S., Verhoeven,

S., Vreede, B., Smeets, S., Soares Siqueira, A., Kazeroni, R., Potter, J., Winterstein, F., Beucher, R., Kraft, J., Ruhe, L., Bonnet, P., Munday, G., & Chun, F. (2025). ESMValTool. https://doi.org/10.5281/zenodo.3401363

Ansel, J., Yang, E., He, H., Gimelshein, N., Jain, A., Voznesensky, M., Bao, B., Bell, P., Berard, D., Burovski, E., Chauhan, G., Chourdia, A., Constable, W., Desmaison, A., DeVito, Z., Ellison, E., Feng, W., Gong, J., Gschwind, M., Hirsh, B., Huang, S., Kalambarkar, K., Kirsch, L., Lazos, M., Lezcano, M., Liang, Y., Liang, J., Lu, Y., Luk, C. K., Maher, B., Pan, Y., Puhrsch, C., Reso, M., Saroufim, M., Siraichi, M. Y., Suk, H., Suo, M., Tillet, P., Wang, E., Wang, X., Wen, W., Zhang, S., Zhao, X., Zhou, K., Zou, R., Mathews, A., Chanan, G., Wu, P., & Chintala, S. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. *29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24)*. https://doi.org/10.1145/3620665.3640366

Arakawa, A., & Schubert, W. H. (1974). Interaction of a Cumulus Cloud Ensemble with the Large-Scale Environment, Part I. *Journal of the atmospheric sciences*, *31*(3), 674–701.

Arakawa, A., & Jung, J.-H. (2011). Multiscale Modeling of the Moist-Convective Atmosphere — A Review. *Atmospheric Research*, *102*(3), 263–285. https://doi.org/10.1016/j.atmosres.2011.08.009

Arakawa, A., Jung, J.-H., & Wu, C.-M. (2011). Toward Unification of the Multiscale Modeling of the Atmosphere. *Atmospheric Chemistry and Physics*, *11*(8), 3731–3742. https://doi.org/10.5194/acp-11-3731-2011

Arakawa, A. (1966). Computational Design for Long-Term Numerical Integration of the Equations of Fluid Motion: Two-dimensional Incompressible Flow. Part I. *Journal of computational physics*, *1*(1), 119–143.

Arakawa, A. (2004). The Cumulus Parameterization Problem: Past, Present, and Future. *Journal of Climate*, *17*(13), 2493–2525. https://doi.org/10.1175/1520-0442(2004)017<2493:RATCPP>2.0.CO;2

Atkinson, J., Elafrou, A., Kasoar, E., Wallwork, J. G., Meltzer, T., Clifford, S., Orchard, D., & Edsall, C. (2025). FTorch: A Library for Coupling PyTorch Models to Fortran. *Journal of Open Source Software*, *10*(107), 7602. https://doi.org/10.21105/joss.07602

Baba, Y., & Giorgetta, M. A. (2020). Tropical Variability Simulated in ICON-A with a Spectral Cumulus Parameterization. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001732.

Baba, Y. (2019). Spectral Cumulus Parameterization Based on Cloud-Resolving Model. *Climate Dynamics*, *52*(1), 309–334.

Balestriero, R., Pesenti, J., & LeCun, Y. (2021). Learning in High Dimension Always Amounts to Extrapolation. *arXiv e-prints*, arXiv:2110.09485. https://doi.org/10.48550/arXiv.2110.09485

Balogh, B., Saint-Martin, D., & Geoffroy, O. (2025). Online Test of a Neural Network Deep Convection Parameterization in ARP-GEM1. *Artificial Intelligence for the Earth Systems*, *4*(3), 240100.

Bechtold, P., Semane, N., Lopez, P., Chaboureau, J.-P., Beljaars, A., & Bormann, N. (2014). Representing Equilibrium and Nonequilibrium Convection in Large-Scale Models. *Journal of the atmospheric sciences*, *71*(2), 734–753. https://doi.org/10.1175/JAS-D-13-0163.1

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. *Journal of Advances in Modeling Earth Systems*, *14*(8), e2022MS003130.

Behrens, G., Beucler, T., Iglesias-Suarez, F., Yu, S., Gentine, P., Pritchard, M., Schwabe, M., & Eyring, V. (2025). Simulating Atmospheric Processes in Earth System Models and Quantifying Uncertainties with Deep Learning Multi-Member and Stochastic Parameterizations. *Journal of Advances in Modeling Earth Systems*, *17*(4), e2024MS004272.

Betts, A. K., & Miller, M. J. (1993). The Betts-Miller Scheme. In K. A. Emanuel & D. J. Raymond (Eds.), *The Representation of Cumulus Convection in Numerical Models* (pp. 107–121). American Meteorological Society.

Beucler, T., Cronin, T., & Emanuel, K. (2018). A Linear Response Framework for Radiative-Convective Instability. *Journal of Advances in Modeling Earth Systems*, *10*(8), 1924–1951. https://doi.org/10.1029/2018MS001280

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing Analytic Constraints in Neural Networks Emulating Physical Systems. *Physical review letters*, *126*(9), 098302.

Beucler, T., Ebert-Uphoff, I., Rasp, S., Pritchard, M., & Gentine, P. (2023). Machine Learning for Clouds and Climate. In *Clouds and Their Climatic Impacts* (pp. 325–345). American Geophysical Union (AGU). https://doi.org/10.1002/9781119700357.ch16

Beucler, T., Gentine, P., Yuval, J., Gupta, A., Peng, L., Lin, J., Yu, S., Rasp, S., Ahmed, F., O'Gorman, P. A., Neelin, J. D., Lutsko, N. J., & Pritchard, M. (2024). Climate-Invariant Machine Learning. *Science Advances*, *10*(6), eadj7250. https://doi.org/10.1126/sciadv.adj7250

Beucler, T., Grundner, A., Shamekh, S., Ukkonen, P., Chantry, M., & Lagerquist, R. (2025). Distilling Machine Learning's Added Value: Pareto Fronts in Atmospheric Applications. *Artificial Intelligence for the Earth Systems*, *4*(2), e240078. https://doi.org/10.1175/AIES-D-24-0078.1

Bi, K., Xie, L., Zhang, H., Chen, X., Gu, X., & Tian, Q. (2023). Accurate Medium-Range Global Weather Forecasting with 3D Neural Networks. *Nature*, *619*(7970), 533–538.

Blanchard, N., Pantillon, F., Chaboureau, J.-P., & Delanoë, J. (2021). Mid-Level Convection in a Warm Conveyor Belt Accelerates the Jet Stream. *Weather and Climate Dynamics*, *2*(1), 37–53. https://doi.org/10.5194/wcd-2-37-2021

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. *Journal of Geophysical Research: Atmospheres*, *125*(21). https://doi.org/10.1029/2019jd032321

Bony, S., & Dufresne, J.-L. (2005). Marine Boundary Layer Clouds at the Heart of Tropical Cloud Feedback Uncertainties in Climate Models. *Geophysical Research Letters*, *32*(20).

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., et al. (2015). Clouds, Circulation and Climate Sensitivity. *Nature Geoscience*, *8*(4), 261–268.

Boos, W. R., Fedorov, A., & Muir, L. (2016). Convective Self-Aggregation and Tropical Cyclogenesis under the Hypohydrostatic Rescaling. *Journal of the Atmospheric Sciences*, *73*(2), 525–544.

Bougeault, P. (1985). A Simple Parameterization of the Large-Scale Effects of Cumulus Convection. *Monthly Weather Review*, *113*(12), 2108–2121.

Breiman, L., Friedman, J., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*. Chapman and Hall/CRC.

Breiman, L. (1996). Bagging Predictors. *Machine learning*, *24*(2), 123–140.

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic Validation of a Neural Network Unified Physics Parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298.

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.

Bretherton, C. S., Peters, M. E., & Back, L. E. (2004). Relationships between Water Vapor Path and Precipitation over the Tropical Oceans. *Journal of Climate*, *17*(7), 1517–1528. https://doi.org/10.1175/1520-0442(2004)017<1517:RBWVPA>2.0.CO;2

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., Perkins, W. A., Clark, S. K., & Harris, L. (2022). Correcting Coarse-Grid Weather and Climate Models by Machine Learning from Global Storm-Resolving Simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. https://doi.org/10.1029/2021MS002794

Brient, F., & Schneider, T. (2016). Constraints on Climate Sensitivity from Space-Based Measurements of Low-Cloud Reflection. *Journal of Climate*, *29*(16), 5821–5835.

Bronstein, M. M., Bruna, J., Cohen, T., & Veličković, P. (2021). Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges.

Browning, K. A., Hardman, M. E., Harrold, T. W., & Pardoe, C. W. (1973). The Structure of Rainbands within a Mid-Latitude Depression. *Quarterly Journal of the Royal Meteorological Society*, *99*(420), 215–231.

Buizza, R., Milleer, M., & Palmer, T. N. (1999). Stochastic Representation of Model Uncertainties in the ECMWF Ensemble Prediction System. *Quarterly Journal of the Royal Meteorological Society*, *125*(560), 2887–2908.

Cambridge-ICCS. (2024). FTorch - A Library for Coupling (Py)Torch Machine Learning Models to Fortran.

Cauchy, A., et al. (1847). Méthode Générale Pour La Résolution Des Systemes d'équations Simultanées. *Comp. Rend. Sci. Paris*, *25*(1847), 536–538.

Ceppi, P., & Nowack, P. (2021). Observational Evidence That Cloud Feedback Amplifies Global Warming. *Proceedings of the National Academy of Sciences*, *118*(30), e2026290118. https://doi.org/doi:10.1073/pnas.2026290118

Chan, J. Y.-L., Leow, S. M. H., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., & Chen, Y.-L. (2022). Mitigating the Multicollinearity Problem and Its Machine Learning Approach: A Review. *Mathematics*, *10*(8). https://doi.org/10.3390/math10081283

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine Learning Emulation of Gravity Wave Drag in Numerical Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477. https://doi.org/10.1029/2021MS002477

Charney, J. G., Fjörtoft, R., & von Neumann, J. (1950). Numerical Integration of the Barotropic Vorticity Equation. *Tellus*, *2*(4), 237–254.

Chen, S. S., & Houze Jr, R. A. (1997). Diurnal Variation and Life-Cycle of Deep Convective Systems over the Tropical Pacific Warm Pool. *Quarterly Journal of the Royal Meteorological Society*, *123*(538), 357–388. https://doi.org/10.1002/qj.49712353806

Chen, J., Zhang, M., Zhang, T., Lin, W., & Xue, W. (2025). Stable Simulation of the Community Atmosphere Model Using Machine-Learning Physical Parameterization Trained with Experience Replay. *Journal of Advances in Modeling Earth Systems*, *17*(6), e2024MS004722.

Cheruy, F., Chevallier, F., Scott, N., & Chedin, A. (1995). A New Generation of Radiative Transfer Models for Climate Studies Based on Neural Networks. *1995 International Geoscience and Remote Sensing Symposium, IGARSS '95. Quantitative Remote Sensing for Science and Applications*, *1*, 535–537 vol.1. https://doi.org/10.1109/IGARSS.1995.520447

Chevallier, F., Chéruy, F., Scott, N. A., & Chédin, A. (1998). A Neural Network Approach for a Fast and Accurate Computation of a Longwave Radiative Budget. *Journal of Applied Meteorology*, *37*(11), 1385–1397. https://doi.org/10.1175/1520-0450(1998)037<1385:ANNAFA>2.0.CO;2

Chevallier, F., Morcrette, J.-J., Chéruy, F., & Scott, N. A. (2000). Use of a Neural-Network-Based Long-Wave Radiative-Transfer Scheme in the ECMWF Atmospheric Model. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 761–776.

Christensen, H. M., Kouhen, S., Miller, G., & Parthipan, R. (2024). Machine Learning for Stochastic Parametrization. *Environmental Data Science*, *3*, e38. https://doi.org/10.1017/eds.2024.45

Christopoulos, C., & Schneider, T. (2021). Assessing Biases and Climate Implications of the Diurnal Precipitation Cycle in Climate Models. *Geophysical Research Letters*, *48*(13), e2021GL093017. https://doi.org/10.1029/2021GL093017

Christopoulos, C., Lopez-Gomez, I., Beucler, T., Cohen, Y., Kawczynski, C., Dunbar, O. R., & Schneider, T. (2024). Online Learning of Entrainment Closures in a Hybrid Ma-

chine Learning Parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(11), e2024MS004485.

Colin, M., Sherwood, S., Geoffroy, O., Bony, S., & Fuchs, D. (2019). Identifying the Sources of Convective Memory in Cloud-Resolving Simulations. *Journal of the atmospheric sciences*, *76*(3), 947–962.

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 Formulation and Atmospheric Simulation of the Community Atmosphere Model Version 3 (CAM3). *Journal of Climate*, *19*(11), 2144–2161.

Council, N. R. (1979). *Carbon Dioxide and Climate: A Scientific Assessment*. The National Academies Press. https://doi.org/10.17226/12181

Craig, G. C., & Dörnbrack, A. (2008). Entrainment in Cumulus Clouds: What Resolution Is Cloud-Resolving? *Journal of the Atmospheric Sciences*, *65*(12), 3978–3988.

Doswell, C. A., & Evans, J. S. (2003). Proximity Sounding Analysis for Derechos and Supercells: An Assessment of Similarities and Differences. *Atmospheric Research*, *67–68*, 117–133. https://doi.org/10.1016/S0169-8095(03)00047-4

Drake, J. B. (2014). *Climate Modeling for Scientists and Engineers*. SIAM.

Dunne, J. P., Hewitt, H. T., Arblaster, J. M., Bonou, F., Boucher, O., Cavazos, T., Dingley, B., Durack, P. J., Hassler, B., Juckes, M., Miyakawa, T., Mizielinski, M., Naik, V., Nicholls, Z., O'Rourke, E., Pincus, R., Sanderson, B. M., Simpson, I. R., & Taylor, K. E. (2025). An Evolving Coupled Model Intercomparison Project Phase 7 (CMIP7) and Fast Track in Support of Future Climate Assessment. *Geoscientific Model Development*, *18*(19), 6671–6700. https://doi.org/10.5194/gmd-18-6671-2025

E3SM Project. (2018). Energy Exascale Earth System Model (E3SM). https://doi.org/10.11578/E3SM/dc.20180418.36

Easterbrook, S. M. (2023). *Computing the Climate: How We Know What We Know About Climate Change*. Cambridge University Press.

Edwards, P. N. (2011). History of Climate Modeling. *Wiley Interdisciplinary Reviews: Climate Change*, *2*(1), 128–139.

Emanuel, K. A., & Raymond, D. J. (1993). The Representation of Cumulus Convection in Numerical Models. *Meteorological Monographs*, *24*(46), 1–246. https://doi.org/10.1175/0065-9401-24.46.1

Espinosa, Z. I., Sheshadri, A., Cain, G. R., Gerber, E. P., & DallaSanta, K. J. (2022). Machine Learning Gravity Wave Parameterization Generalizes to Capture the QBO and Response to Increased CO2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022GL098174

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.

Eyring, V., Gillett, N., Rao, K. A., Barimalala, R., Parrillo, M. B., Bellouin, N., Cassou, C., Durack, P., Kosaka, Y., McGregor, S., Min, S., Morgenstern, O., & Sun, Y. (2021a).

Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. 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.). Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. https://doi.org/10.1017/9781009157896.005

Eyring, V., Mishra, V., Griffith, G. P., Chen, L., Keenan, T., Turetsky, M. R., Brown, S., Jotzo, F., Moore, F. C., & van der Linden, S. (2021b). Reflections and Projections on a Decade of Climate Science. *Nature Climate Change*, *11*(4), 279–285. https://doi.org/10.1038/s41558-021-01020-x

Eyring, V., Collins, W. D., Gentine, P., Barnes, E. A., Barreiro, M., Beucler, T., Bocquet, M., Bretherton, C. S., Christensen, H. M., Dagon, K., Gagne, D. J., Hall, D., Hammerling, D., Hoyer, S., Iglesias-Suarez, F., Lopez-Gomez, I., McGraw, M. C., Meehl, G. A., Molina, M. J., Monteleoni, C., Mueller, J., Pritchard, M. S., Rolnick, D., Runge, J., Stier, P., Watt-Meyer, O., Weigel, K., Yu, R., & Zanna, L. (2024a). Pushing the Frontiers in Climate Modelling and Analysis with Machine Learning. *Nature Climate Change*, *14*(9), 916–928. https://doi.org/10.1038/s41558-024-02095-y

Eyring, V., Gentine, P., Camps-Valls, G., Lawrence, D. M., & Reichstein, M. (2024b). AI-empowered next-Generation Multiscale Climate Modelling for Mitigation and Adaptation. *Nature Geoscience*, *17*(10), 963–971. https://doi.org/10.1038/s41561-024-01527-w

Falcon, W., & The PyTorch Lightning team. (2019). PyTorch Lightning. https://doi.org/10.5281/zenodo.3828935

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., Lunt, D., Mauritsen, T., Palmer, M., Watanabe, M., Wild, M., & Zhang, H. (2021). The Earth's Energy Budget, Climate Feedbacks, and Climate Sensitivity. 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. 923–1054). Cambridge University Press. https://doi.org/10.1017/9781009157896.009

Fosser, G., Gaetani, M., Kendon, E. J., Adinolfi, M., Ban, N., Belušić, D., Caillaud, C., Careto, J. A. M., Coppola, E., Demory, M.-E., de Vries, H., Dobler, A., Feldmann, H., Goergen, K., Lenderink, G., Pichelli, E., Schär, C., Soares, P. M. M., Somot, S., & Tölle, M. H. (2024). Convection-Permitting Climate Models Offer More Certain Extreme Rainfall Projections. *npj Climate and Atmospheric Science*, *7*(1), 51. https://doi.org/10.1038/s41612-024-00600-w

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *Annals of statistics*, 1189–1232.

Friedman, J. H. (2002). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

Garner, S. T., Frierson, D. M. W., Held, I. M., Pauluis, O., & Vallis, G. K. (2007). Resolving Convection in a Global Hypohydrostatic Model. *Journal of the Atmospheric Sciences*, *64*(6), 2061–2075.

Gelaro, R., McCarty, W., Suárez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Silva, A. M. da, Gu, W., Kim, G.-K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., & Zhao, B. (2017). The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *Journal of Climate*, *30*(14), 5419–5454. https://doi.org/10.1175/JCLI-D-16-0758.1

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

Géron, A. (2022). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. " O'Reilly Media, Inc."

Gettelman, A., Salby, M. L., & Sassi, F. (2002). Distribution and Influence of Convection in the Tropical Tropopause Region. *Journal of Geophysical Research: Atmospheres*, *107*(D10), ACL 6-1-ACL 6–12. https://doi.org/10.1029/2001JD001048

Gettelman, A., & Rood, R. B. (2016). *Demystifying Climate Models: A Users Guide to Earth System Models*. Springer Nature.

Gettelman, A., Gagne, D. J., Chen, C.-C., Christensen, M. W., Lebo, Z. J., Morrison, H., & Gantos, G. (2021). Machine Learning the Warm Rain Process. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002268.

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely Randomized Trees. *Machine Learning*, *63*(1), 3–42. https://doi.org/10.1007/s10994-006-6226-1

Giglio, D., Gille, S. T., Cornuelle, B. D., Subramanian, A. C., Turk, F. J., Hristova-Veleva, S., & Northcott, D. (2022). Annual Modulation of Diurnal Winds in the Tropical Oceans. *Remote Sensing*, *14*(459). https://doi.org/10.3390/rs14030459

Giorgetta, M. A., Brokopf, R., Crueger, T., Esch, M., Fiedler, S., Helmert, J., Hohenegger, C., Kornblueh, L., Köhler, M., Manzini, E., Mauritsen, T., Nam, C., Raddatz, T., Rast, S., Reinert, D., Sakradzija, M., Schmidt, H., Schneck, R., Schnur, R., Silvers, L., Wan, H., Zängl, G., & Stevens, B. (2018). ICON-A, the Atmosphere Component of the ICON Earth System Model: I. Model Description. *Journal of Advances in Modeling Earth Systems*, *10*(7), 1613–1637. https://doi.org/10.1029/2017MS001242

Giorgetta, M. A., Sawyer, W., Lapillonne, X., Adamidis, P., Alexeev, D., Clément, V., Dietlicher, R., Engels, J. F., Esch, M., Franke, H., et al. (2022). The ICON-A Model for Direct QBO Simulations on GPUs (Version Icon-Cscs: Baf28a514). *Geoscientific Model Development*, *15*(18), 6985–7016.

Golaz, J.-C., Van Roekel, L. P., Zheng, X., Roberts, A. F., Wolfe, J. D., Lin, W., Bradley, A. M., Tang, Q., Maltrud, M. E., Forsyth, R. M., et al. (2022). The DOE E3SM Model Version 2:

Overview of the Physical Model and Initial Model Evaluation. *Journal of Advances in Modeling Earth Systems*, *14*(12), e2022MS003156.

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

Grabowski, W. W., & Smolarkiewicz, P. K. (1999). CRCP: A Cloud Resolving Convection Parameterization for Modeling the Tropical Convecting Atmosphere. *Physica D: Nonlinear Phenomena*, *133*(1–4), 171–178.

Grabowski, W. W. (2001). Coupling Cloud Processes with the Large-Scale Dynamics Using the Cloud-Resolving Convection Parameterization (CRCP). *Journal of the Atmospheric Sciences*, *58*(9), 978–997.

Gregory, D., & Rowntree, P. R. (1990). A Mass Flux Convection Scheme with Representation of Cloud Ensemble Characteristics and Stability-Dependent Closure. *Monthly Weather Review*, *118*(7), 1483–1506. https://doi.org/10.1175/1520-0493(1990)118<1483:AMFCSW>2.0.CO;2

Grell, G. A., & Dévényi, D. (2002). A Generalized Approach to Parameterizing Convection Combining Ensemble and Data Assimilation Techniques. *Geophysical Research Letters*, *29*(14), 38–1.

Grundner, A., Beucler, T., Gentine, P., Iglesias-Suarez, F., Giorgetta, M. A., & Eyring, V. (2022). Deep Learning Based Cloud Cover Parameterization for ICON. *Journal of Advances in Modeling Earth Systems*, *14*(12), e2021MS002959. https://doi.org/10.1029/2021MS002959

Grundner, A., Beucler, T., Gentine, P., & Eyring, V. (2024). Data-Driven Equation Discovery of a Cloud Cover Parameterization. *Journal of Advances in Modeling Earth Systems*, *16*(3), e2023MS003763.

Grundner, A., Beucler, T., Savre, J., Lauer, A., Schlund, M., & Eyring, V. (2025). Reduced Cloud Cover Errors in a Hybrid AI-Climate Model Through Equation Discovery And Automatic Tuning.

Gupta, A., Sheshadri, A., Roy, S., Gaur, V., Maskey, M., & Ramachandran, R. (2024). Machine Learning Global Simulation of Nonlocal Gravity Wave Propagation. *arXiv preprint arXiv:2406.14775*.

Gupta, A., Sheshadri, A., Roy, S., Schmude, J., Gaur, V., Leong, W. J., Maskey, M., & Ramachandran, R. (2025). Finetuning AI Foundation Models to Develop Subgrid-Scale Parameterizations: A Case Study on Atmospheric Gravity Waves. *arXiv preprint arXiv:2509.03816*.

Gurrola-Ramos, J., Dalmau, O., & Alarcon, T. E. (2021). A Residual Dense U-Net Neural Network for Image Denoising. *IEEE access : practical innovations, open solutions*, *9*, 31742–31754.

Haarsma, R. J., Roberts, M. J., Vidale, P. L., Senior, C. A., Bellucci, A., Bao, Q., Chang, P., Corti, S., Fučkar, N. S., Guemas, V., et al. (2016). High Resolution Model Intercomparison Project (HighResMIP v1. 0) for CMIP6. *Geoscientific Model Development*, *9*(11), 4185–4208.

Hafner, K., Iglesias-Suarez, F., Shamekh, S., Gentine, P., Giorgetta, M. A., Pincus, R., & Eyring, V. (2025a). Interpretable Machine Learning-Based Radiation Emulation for Icon. *Journal of Geophysical Research: Machine Learning and Computation*, *2*(4), e2024JH000501.

Hafner, K., Iglesias-Suarez, F., Shamekh, S., Gentine, P., Giorgetta, M. A., Pincus, R., & Eyring, V. (2025b). Stable Machine Learning Based Radiation Emulation for ICON. *Authorea Preprints*.

Hafner, K., Shamekh, S., Bertoli, G., Lauer, A., Pincus, R., Savre, J., & Eyring, V. (2025c). *Representing Subgrid-Scale Cloud Effects in a Radiation Parameterization Using Machine Learning: MLe-radiation v1. 0*.

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A Moist Physics Parameterization Based on Deep Learning. *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. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2022MS003508. https://doi.org/10.1029/2022MS003508

Hannah, W. M., Jones, C. R., Hillman, B. R., Norman, M. R., Bader, D. C., Taylor, M. A., Leung, L. R., Pritchard, M. S., Branson, M. D., Lin, G., Pressel, K. G., & Lee, J. M. (2020). Initial Results From the Super-Parameterized E3SM. *Journal of Advances in Modeling Earth Systems*, *12*(1), e2019MS001863. https://doi.org/10.1029/2019MS001863

Hannah, W., Pressel, K., Ovchinnikov, M., & Elsaesser, G. (2022). Checkerboard Patterns in E3SMv2 and E3SM-MMFv2. *Geoscientific Model Development*, *15*(15), 6243–6257. https://doi.org/10.5194/gmd-15-6243-2022

Hannah, W., & Pressel, K. (2022). A Method for Transporting Cloud-Resolving Model Variance in a Multiscale Modeling Framework. *Geoscientific Model Development*, *15*(24), 8999–9013.

Harder, P., Watson-Parris, D., Stier, P., Strassel, D., Gauger, N. R., & Keuper, J. (2022). Physics-Informed Learning of Aerosol Microphysics. *Environmental Data Science*, *1*, e20.

Hardiman, S. C., Scaife, A. A., van Niekerk, A., Prudden, R., Owen, A., Adams, S. V., Dunstan, T., Dunstone, N. J., & Madge, S. (2023). Machine Learning for Nonorographic Gravity Waves in a Climate Model. *Artificial intelligence for the Earth systems*, *2*(4), e220081.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Heikes, R. P., Randall, D. A., & Konor, C. S. (2013). Optimized Icosahedral Grids: Performance of Finite-Difference Operators and Multigrid Solver. *Monthly Weather Review*, *141*(12), 4450–4469.

helgehr. (2024). *EyringMLClimateGroup/heuer23_ml_convection_paramet erization: Interpretable multiscale Machine Learning-Based Parameterizations of Convection for ICON* (Version v1.0). Zenodo. https://doi.org/10.5281/zenodo.12773936

helgehr. (2025). *EyringMLClimateGroup/heuer25james_ml_convection_climsim: Beyond the Training Data: Confidence-Guided Mixing of Parameterizations in a Hybrid AI-Climate Model* (Version v1.0). Zenodo. https://doi.org/10.5281/zenodo.17234569

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., De Chiara, G., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A.,

Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., de Rosnay, P., Rozum, I., Vamborg, F., Villaume, S., & Thépaut, J.-N. (2020). The ERA5 Global Reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049. https://doi.org/10.1002/qj.3803

Herzegh, P. H., & Hobbs, P. V. (1980). The Mesoscale and Microscale Structure and Organization of Clouds and Precipitation in Midlatitude Cyclones. II: Warm-Frontal Clouds. *Journal of Atmospheric Sciences*, *37*(3), 597–611. https://doi.org/10.1175/1520-0469(1980)037<0597:TMAMSA>2.0.CO;2

Heuer, **H.**, Schwabe, M., Gentine, P., Giorgetta, M. A., & Eyring, V. (2024). Interpretable Multiscale Machine Learning-Based Parameterizations of Convection for ICON. *Journal of Advances in Modeling Earth Systems*, *16*(8), e2024MS004398. https://doi.org/10.1029/2024MS004398

Heuer, **H.**, Beucler, T., Schwabe, M., Savre, J., Schlund, M., & Eyring, V. (2025). Beyond the Training Data: Confidence-guided Mixing of Parameterizations in a Hybrid AI-climate Model [Under Review for the Journal of Advances in Modeling Earth Systems]. *arXiv preprint arXiv:2510.08107*. https://doi.org/10.48550/arXiv.2510.08107

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural computation*, *9*(8), 1735–1780.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics : a journal of statistics for the physical, chemical, and engineering sciences*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

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., Giorgetta, M., Gutjahr, O., Haak, H., Hanke, M., Ilyina, T., Jahns, T., Jungclaus, J., Kern, M., Klocke, D., Kluft, L., Kölling, T., Kornblueh, L., Kosukhin, S., Kroll, C., Lee, J., Mauritsen, T., Mehlmann, C., Mieslinger, T., Naumann, A. K., Paccini, L., Peinado, A., Praturi, D. S., Putrasahan, D., Rast, S., Riddick, T., Roeber, N., Schmidt, H., Schulzweida, U., Schütte, F., Segura, H., Shevchenko, R., Singh, V., Specht, M., Stephan, C. C., von Storch, J. S., Vogel, R., Wengel, C., Winkler, M., Ziemen, F., Marotzke, J., & 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

Holloway, C. E., & Neelin, J. D. (2009). Moisture Vertical Structure, Column Water Vapor, and Tropical Deep Convection. *Journal of the Atmospheric Sciences*, *66*(6), 1665–1683. https://doi.org/10.1175/2008JAS2806.1

Honnert, R., Efstathiou, G. A., Beare, R. J., Ito, J., Lock, A., Neggers, R., Plant, R. S., Shin, H. H., Tomassini, L., & Zhou, B. (2020). The Atmospheric Boundary Layer and the "Gray Zone" of Turbulence: A Critical Review. *Journal of Geophysical Research: Atmospheres*, *125*(13), e2019JD030317.

Houze Jr, R. A., Locatelli, J. D., & Hobbs, P. V. (1976). Dynamics and Cloud Microphysies of the Rainbands in an Occluded Frontal System. *Journal of Atmospheric Sciences*, *33*(10), 1921–1936.

Houze Jr, R. A., & Betts, A. K. (1981). Convection in GATE. *Reviews of Geophysics*, *19*(4), 541–576.

Houze, R. A. (1997). Stratiform Precipitation in Regions of Convection: A Meteorological Paradox? *Bulletin of the American Meteorological Society*, *78*(10), 2179–2196. https://doi.org/10.1175/1520-0477(1997)078<2179:SPIROC>2.0.CO;2

Hu, Z., Subramaniam, A., Kuang, Z., Lin, J., Yu, S., Hannah, W. M., Brenowitz, N. D., Romero, J., & Pritchard, M. S. (2025). Stable Machine-Learning Parameterization of Subgrid Processes in a Comprehensive Atmospheric Model Learned from Embedded Convection-Permitting Simulations. *Journal of Advances in Modeling Earth Systems*, *17*(7), e2024MS004618.

Hwang, Y.-T., & Frierson, D. M. W. (2013). Link between the Double-Intertropical Convergence Zone Problem and Cloud Biases over the Southern Ocean. *Proceedings of the National Academy of Sciences*, *110*(13), 4935–4940. https://doi.org/10.1073/pnas.1213302110

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. *Journal of Geophysical Research: Atmospheres*, *129*(4), e2023JD039202.

IPCC. (2013). *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press. https://doi.org/10.1017/CBO9781107415324

IPCC. (2021). *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (V. Masson-Delmotte, P. Zhai, A. Pirani, S. L. Connors, C. Péan, S. Berger, N. Caud, Y. Chen, L. Goldfarb, M. I. Gomis, M. Huang, K. Leitzell, E. Lonnoy, J. B. R. Matthews, T. K. Maycock, T. Waterfield, O. Yelekçi, R. Yu, & B. Zhou, Eds.). Cambridge University Press. https://doi.org/10.1017/9781009157896

Johnson, R. H., Rickenbach, T. M., Rutledge, S. A., Ciesielski, P. E., & Schubert, W. H. (1999). Trimodal Characteristics of Tropical Convection. *Journal of Climate*, *12*(8), 2397–2418. https://doi.org/10.1175/1520-0442(1999)012<2397:TCOTC>2.0.CO;2

Jones, C., Waliser, D. E., Lau, K. M., & Stern, W. (2004). Global Occurrences of Extreme Precipitation and the Madden–Julian Oscillation: Observations and Predictability. *Journal of Climate*, *17*(23), 4575–4589. https://doi.org/10.1175/3238.1

Judt, F., Klocke, D., Rios-Berrios, R., Vanniere, B., Ziemen, F., Auger, L., Biercamp, J., Bretherton, C., Chen, X., Düben, P., et al. (2021). Tropical Cyclones in Global Storm-Resolving Models. *Journal of the Meteorological Society of Japan. Ser. II*, *99*(3), 579–602.

Judt, F. (2018). Insights into Atmospheric Predictability through Global Convection-Permitting Model Simulations. *Journal of the Atmospheric Sciences*, *75*(5), 1477–1497. https://doi.org/10.1175/JAS-D-17-0343.1

Jülich Supercomputing Centre. (2021). JUWELS Cluster and Booster: Exascale Pathfinder with Modular Supercomputing Architecture at Juelich Supercomputing Centre. *Journal of large-scale research facilities*, *7*(A138). https://doi.org/10.17815/jlsrf-7-183

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly Accurate Protein Structure Prediction with AlphaFold. *nature*, *596*(7873), 583–589.

Jung, J.-H., & Arakawa, A. (2010). Development of a Quasi-3D Multiscale Modeling Framework: Motivation, Basic Algorithm and Preliminary Results. *Journal of Advances in Modeling Earth Systems*, *2*(4).

Jungclaus, J. H., Lorenz, S. J., Schmidt, H., Brovkin, V., Brüggemann, N., Chegini, F., Crüger, T., De-Vrese, P., Gayler, V., Giorgetta, M. A., et al. (2022). The ICON Earth System Model Version 1.0. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002813.

Kain, J. S., & Fritsch, J. M. (1990). A One-Dimensional Entraining/Detraining Plume Model and Its Application in Convective Parameterization. *Journal of Atmospheric Sciences*, *47*(23), 2784–2802.

Kain, J. S., & Fritsch, J. M. (1993). Convective Parameterization for Mesoscale Models: The Kain-Fritsch Scheme. In *The Representation of Cumulus Convection in Numerical Models* (pp. 165–170). Springer.

Kalthoff, N., Adler, B., Barthlott, Ch., Corsmeier, U., Mobbs, S., Crewell, S., Träumner, K., Kottmeier, Ch., Wieser, A., Smith, V., & Di Girolamo, P. (2009). The Impact of Convergence Zones on the Initiation of Deep Convection: A Case Study from COPS. *Atmospheric Research*, *93*(4), 680–694. https://doi.org/10.1016/j.atmosres.2009.02.010

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). Lightgbm: A Highly Efficient Gradient Boosting Decision Tree. *Advances in neural information processing systems*, *30*.

Kendon, E. J., Prein, A. F., Senior, C. A., & Stirling, A. (2021). Challenges and Outlook for Convection-Permitting Climate Modelling. *Philosophical Transactions of the Royal Society A*, *379*(2195), 20190547.

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.

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.

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.

Kim, Y.-J., Eckermann, S. D., & Chun, H.-Y. (2003). An Overview of the Past, Present and Future of Gravity-wave Drag Parametrization for Numerical Climate and Weather Prediction Models. *Atmosphere-Ocean*, *41*(1), 65–98. https://doi.org/10.3137/ao.410105

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *arXiv e-prints*, arXiv:1412.6980. https://doi.org/10.48550/arXiv.1412.6980

Kirshbaum, D. J. (2022). Large-Eddy Simulations of Convection Initiation over Heterogeneous, Low Terrain. *Journal of the Atmospheric Sciences*.

Klocke, D., Brueck, M., Hohenegger, C., & Stevens, B. (2017). Rediscovery of the Doldrums in Storm-Resolving Simulations over the Tropical Atlantic. *Nature Geoscience*, *10*(12), 891–896.

Kochkov, D., Yuval, J., Langmore, I., Norgaard, P., Smith, J., Mooers, G., Klöwer, M., Lottes, J., Rasp, S., Düben, P., et al. (2024). Neural General Circulation Models for Weather and Climate. *Nature*, *632*(8027), 1060–1066.

Koldunov, N., Kölling, T., Pedruzo-Bagazgoitia, X., Rackow, T., Redler, R., Sidorenko, D., Wieners, K.-H., & Ziemen, F. A. (2023). nextGEMS: Output of the Model Development Cycle 3 Simulations for ICON and IFS. Doi: 10.26050. *WDCC/nextGEMS_cyc3*.

Kooperman, G. J., Pritchard, M. S., Burt, M. A., Branson, M. D., & Randall, D. A. (2016). Robust Effects of Cloud Superparameterization on Simulated Daily Rainfall Intensity Statistics across Multiple Versions of the C Ommunity E Arth S Ystem M Odel. *Journal of Advances in Modeling Earth Systems*, *8*(1), 140–165.

Korn, P., Brüggemann, N., Jungclaus, J. H., Lorenz, S. J., Gutjahr, O., Haak, H., Linardakis, L., Mehlmann, C., Mikolajewicz, U., Notz, D., et al. (2022). ICON-O: The Ocean Component of the ICON Earth System Model—Global Simulation Characteristics and Local Telescoping Capability. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2021MS002952.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New Approach to Calculation of Atmospheric Model Physics: Accurate and Fast Neural Network Emulation of Longwave Radiation in a Climate Model. *Monthly Weather Review*, *133*(5), 1370–1383.

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., Tolman, H. L., & Belochitski, A. A. (2008). Neural Network Approach for Robust and Fast Calculation of Physical Processes in Numerical Environmental Models: Compound Parameterization with a Quality Control of Larger Errors. *Neural Networks*, *21*(2–3), 535–543.

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. *Advances in Artificial Neural Systems*, *2013*.

Kuang, Z., Blossey, P. N., & Bretherton, C. S. (2005). A New Approach for 3D Cloud-resolving Simulations of Large-scale Atmospheric Circulation. *Geophysical Research Letters*, *32*(2).

Kuang, Z. (2018). Linear Stability of Moist Convecting Atmospheres. Part I: From Linear Response Functions to a Simple Model and Applications to Convectively Coupled Waves. *Journal of the atmospheric sciences*, *75*(9), 2889–2907. https://doi.org/10.1175/JAS-D-18-0092.1

Kühnlein, C., Deconinck, W., Klein, R., Malardel, S., Piotrowski, Z. P., Smolarkiewicz, P. K., Szmelter, J., & Wedi, N. P. (2019). FVM 1.0: A Nonhydrostatic Finite-Volume Dynamical Core for the IFS. *Geoscientific Model Development*, *12*(2), 651–676.

Kuo, H.-L. (1965). On Formation and Intensification of Tropical Cyclones through Latent Heat Release by Cumulus Convection. *Journal of Atmospheric Sciences*, *22*(1), 40–63.

Kuo, H.-L. (1974). Further Studies of the Parameterization of the Influence of Cumulus Convection on Large-Scale Flow. *Journal of Atmospheric Sciences*, *31*(5), 1232–1240.

Kwa, A., Clark, S. K., Henn, B., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., Perkins, W. A., Harris, L., & Bretherton, C. S. (2023). Machine-Learned Climate Model Corrections from a Global Storm-Resolving Model: Performance across the Annual Cycle. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003400. https://doi.org/10.1029/2022MS003400

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., Ravuri, S., Ewalds, T., Eaton-Rosen, Z., Hu, W., et al. (2023). Learning Skillful Medium-Range Global Weather Forecasting. *Science*, *382*(6677), 1416–1421.

LEAP. (2023). ClimSim_high-Res (Revision D251368). *Hugging Face*. https://doi.org/10.57967/hf/0739

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten Digit Recognition with a Back-Propagation Network. *Advances in neural information processing systems*, *2*.

LeCun, Y., Bottou, L., Orr, G. B., & Müller, K.-R. (2002). Efficient Backprop. In *Neural Networks: Tricks of the Trade* (pp. 9–50). Springer.

LeCun, Y. (1985). Une Procedure d'apprentissage Ponr Reseau a Seuil Asymetrique. *Proceedings of cognitiva 85*, 599–604.

Lee, J.-Y., Marotzke, J., Bala, G., Cao, L., Corti, S., Dunne, J. P., Engelbrecht, F., Fischer, E., Fyfe, J. C., Jones, C., Maycock, A., Mutemi, J., Niaye, O., Panickal, S., Zhou, T., & Christensen, H. M. (2021). Future Global Climate: Scenario-Based Projections and near-Term Information. In *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 553–672). Cambridge University Press.

Lee, J., Hannah, W. M., & Bader, D. C. (2023). Representation of Atmosphere-Induced Heterogeneity in Land–Atmosphere Interactions in E3SM–MMFv2. *Geoscientific Model Development*, *16*(24), 7275–7287. https://doi.org/10.5194/gmd-16-7275-2023

LeMone, M. A. (1983). Momentum Transport by a Line of Cumulonimbus. *Journal of Atmospheric Sciences*, *40*(7), 1815–1834. https://doi.org/10.1175/1520-0469(1983)040<1815:MTBALO>2.0.CO;2

Leuenberger, D., Koller, M., Fuhrer, O., & Schär, C. (2010). A Generalization of the SLEVE Vertical Coordinate. *Monthly Weather Review*, *138*(9), 3683–3689.

Li, F., Rosa, D., Collins, W. D., & Wehner, M. F. (2012). "Super-parameterization": A Better Way to Simulate Regional Extreme Precipitation? *Journal of Advances in Modeling Earth Systems*, *4*(2).

Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., & Talwalkar, A. (2018). A System for Massively Parallel Hyperparameter Tuning. *arXiv e-prints*, arXiv:1810.05934. https://doi.org/10.48550/arXiv.1810.05934

Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J. E., & Stoica, I. (2018). Tune: A Research Platform for Distributed Model Selection and Training. *arXiv preprint arXiv:1807.05118*.

Limon, G. C., & Jablonowski, C. (2023). Probing the Skill of Random Forest Emulators for Physical Parameterizations via a Hierarchy of Simple CAM6 Configurations. *Journal of Advances in Modeling Earth Systems*, *15*(6), e2022MS003395. https://doi.org/10.1029/2022MS003395

Lin, J.-L., Lee, M.-I., Kim, D., Kang, I.-S., & Frierson, D. M. W. (2008). The Impacts of Convective Parameterization and Moisture Triggering on AGCM-Simulated Convectively Coupled Equatorial Waves. *Journal of Climate*, *21*(5), 883–909. https://doi.org/10.1175/2007JCLI1790.1

Lin, J., Hu, Z., Yu, S., Pritchard, M., Gupta, R., Zheng, T., Hannah, W., Mansfield, L., Qu, Y., Geleta, M., Lopez, M., Rudolph, M., Chow, A., & Reade, W. (2024). LEAP - Atmospheric Physics Using AI (ClimSim).

Lin, J., Yu, S., Peng, L., Beucler, T., Wong-Toi, E., Hu, Z., Gentine, P., Geleta, M., & Pritchard, M. (2025). Navigating the Noise: Bringing Clarity to ML Parameterization Design With O O (100) Ensembles. *Journal of Advances in Modeling Earth Systems*, *17*(4), e2024MS004551.

Lin, S.-J. (2004). A "Vertically Lagrangian" Finite-Volume Dynamical Core for Global Models. *Monthly Weather Review*, *132*(10), 2293–2307.

Lindzen, R. S. (1981). Some Remarks on Cumulus Parameterization. Rep. on NASA-GISS Workshop: Clouds in Climate: Modelling and Satellite Observational Studies, 42–51.

Linnainmaa, S. (1970). *The Representation of the Cumulative Rounding Error of an Algorithm as a Taylor Expansion of the Local Rounding Errors* [Doctoral dissertation, Master's Thesis (in Finnish), Univ. Helsinki].

Liu, N., Pritchard, M. S., Jenney, A. M., & Hannah, W. M. (2023). Understanding Precipitation Bias Sensitivities in E3SM-multi-scale Modeling Framework from a Dilution Framework. *Journal of Advances in Modeling Earth Systems*, *15*(4), e2022MS003460.

Loshchilov, I., & Hutter, F. (2019). Decoupled Weight Decay Regularization.

Lu, Z., & Chen, Y. (2022). Single Image Super-Resolution Based on a Modified U-net with Mixed Gradient Loss. *signal, image and video processing*, *16*(5), 1143–1151.

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc.

Mahajan, S., Passarella, L. S., Tang, Q., Keen, N. D., Caldwell, P. M., van Roekel, L. P., & Golaz, J.-C. (2023). ENSO Diversity and the Simulation of Its Teleconnections to Winter Precipitation Extremes over the US in High Resolution Earth System Models. *Geophysical Research Letters*, *50*(11). https://doi.org/10.1029/2022gl102657

Majda, A. J. (2007). Multiscale Models with Moisture and Systematic Strategies for Superparameterization. *Journal of the Atmospheric Sciences*, *64*(7), 2726–2734. https://doi.org/10.1175/JAS3976.1

Manabe, S., Smagorinsky, J., & Strickler, R. F. (1965). Simulated Climatology of a General Circulation Model with a Hydrologic Cycle. *Monthly Weather Review*, *93*(12), 769–798.

Manabe, S., & Wetherald, R. T. (1967). Thermal Equilibrium of the Atmosphere with a Given Distribution of Relative Humidity. *Journal of Atmospheric Sciences*, *24*(3), 241–259. https://doi.org/10.1175/1520-0469(1967)024<0241:TEOTAW>2.0.CO;2

Manabe, S., & Bryan, K. (1969). Climate Calculations with a Combined Ocean-Atmosphere Model. *Journal of Atmospheric Sciences*, *26*(4), 786–789.

Manabe, S., Bryan, K., & Spelman, M. J. (1975). A Global Ocean-Atmosphere Climate Model. Part I. The Atmospheric Circulation. *Journal of Physical Oceanography*, *5*(1), 3–29.

Mansfield, L. A., & Sheshadri, A. (2024). Uncertainty Quantification of a Machine Learning Subgrid-Scale Parameterization for Atmospheric Gravity Waves. *Journal of Advances in Modeling Earth Systems*, *16*(7), e2024MS004292.

Martinez-Villalobos, C., & Neelin, J. D. (2019). Why Do Precipitation Intensities Tend to Follow Gamma Distributions? *Journal of the Atmospheric Sciences*, *76*(11), 3611–3631. https://doi.org/10.1175/JAS-D-18-0343.1

Matsuoka, D., Watanabe, S., Sato, K., Kawazoe, S., Yu, W., & Easterbrook, S. (2020). Application of Deep Learning to Estimate Atmospheric Gravity Wave Parameters in Reanalysis Data Sets. *Geophysical Research Letters*, *47*(19), e2020GL089436.

Miura, H., Satoh, M., Nasuno, T., Noda, A. T., & Oouchi, K. (2007). A Madden-Julian Oscillation Event Realistically Simulated by a Global Cloud-Resolving Model. *Science*, *318*(5857), 1763–1765.

Möbis, B., & Stevens, B. (2012). Factors Controlling the Position of the Intertropical Convergence Zone on an Aquaplanet. *Journal of Advances in Modeling Earth Systems*, *4*(4). https://doi.org/10.1029/2012MS000199

Molnar, C. (2025). *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable* (3rd ed.).

Moncrieff, M. W. (2019). Toward a Dynamical Foundation for Organized Convection Parameterization in GCMs. *Geophysical Research Letters*, *46*(23), 14103–14108.

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

Mooers, G., Tuyls, J., Mandt, S., Pritchard, M., & Beucler, T. G. (2021b). Generative Modeling of Atmospheric Convection. *Proceedings of the 10th International Conference on Climate Informatics*, 98–105. https://doi.org/10.1145/3429309.3429324

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. *Scientific Reports*, *13*(1), 22365. https://doi.org/10.1038/s41598-023-49455-w

Moorthi, S., & Suarez, M. J. (1992). Relaxed Arakawa-Schubert. A Parameterization of Moist Convection for General Circulation Models. *Monthly Weather Review*, *120*(6), 978–1002.

Morcrette, C., Cave, T., Reid, H., da Silva Rodrigues, J., Deveney, T., Kreusser, L., Van Weverberg, K., & Budd, C. (2025). Scale-Aware Parameterization of Cloud Fraction and Condensate for a Global Atmospheric Model Machine-Learned From Coarse-Grained Kilometer-Scale Simulations. *Journal of Advances in Modeling Earth Systems*, *17*(4), e2024MS004651. https://doi.org/10.1029/2024MS004651

Moseley, C., Hohenegger, C., Berg, P., & Haerter, J. O. (2016). Intensification of Convective Extremes Driven by Cloud–Cloud Interaction. *Nature Geoscience*, *9*(10), 748–752. https://doi.org/10.1038/ngeo2789

Müller, W. A., Lorenz, S., Pham, T. V., Schneidereit, A., Brokopf, R., Brovkin, V., Brüggemann, N., Chegini, F., Dommenget, D., Fröhlich, K., Früh, B., Gayler, V., Haak, H., Hagemann, S., Hanke, M., Ilyina, T., Jungclaus, J., Köhler, M., Korn, P., Kornblüh, L., Kroll, C., Krüger, J., Castro-Morales, K., Niemeier, U., Pohlmann, H., Polkova, I., Potthast, R., Riddick, T., Schlund, M., Stacke, T., Wirth, R., Yu, D., & Marotzke, J. (2025). The ICON-based Earth System Model for Climate Predictions and Projections (ICON XPP v1.0). *EGUsphere*, *2025*, 1–60. https://doi.org/10.5194/egusphere-2025-2473

Nadiga, B. T., Sun, X., & Nash, C. (2022). Stochastic Parameterization of Column Physics Using Generative Adversarial Networks. *Environmental Data Science*, *1*, e22.

Nair, V., & Hinton, G. E. (2010). Rectified Linear Units Improve Restricted Boltzmann Machines. *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 807–814.

Neale, R. B., Chen, C.-C., Gettelman, A., Lauritzen, P. H., Park, S., Williamson, D. L., Conley, A. J., Garcia, R., Kinnison, D., Lamarque, J.-F., et al. (2010). Description of the NCAR Community Atmosphere Model (CAM 5.0). *NCAR Tech. Note Ncar/tn-486+ STR*, *1*(1), 1–12.

Neelin, J. D., Peters, O., Lin, J. W.-B., Hales, K., & Holloway, C. E. (2008). Rethinking Convective Quasi-Equilibrium: Observational Constraints for Stochastic Convective Schemes in Climate Models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *366*(1875), 2579–2602.

Neelin, J. D. (2010). *Climate Change and Climate Modeling*. Cambridge University Press.

Neumann, P., Düben, P., Adamidis, P., Bauer, P., Brück, M., Kornblueh, L., Klocke, D., Stevens, B., Wedi, N., & Biercamp, J. (2019). Assessing the Scales in Numerical Weather and Climate Predictions: Will Exascale Be the Rescue? *Philosophical Transactions of the Royal Society A*, *377*(2142), 20180148.

Nordeng, T. E. (1994). Extended Versions of the Convective Parametrization Scheme at ECMWF and Their Impact on the Mean and Transient Activity of the Model in the Tropics. *Research Department Technical Memorandum*, *206*, 1–41.

O'Gorman, P. A., & Dwyer, J. G. (2018). Using Machine Learning to Parameterize Moist Convection: Potential for Modeling of Climate, Climate Change, and Extreme Events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563.

Otness, K., Zanna, L., & Bruna, J. (2023). Data-Driven Multiscale Modeling of Subgrid Parameterizations in Climate Models. *arXiv e-prints*, arXiv:2303.17496. https://doi.org/10.48550/arXiv.2303.17496

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), 8888811.

Ouala, S., Chapron, B., Collard, F., Gaultier, L., & Fablet, R. (2024). Online Calibration of Deep Learning Sub-Models for Hybrid Numerical Modeling Systems. *Communications Physics*, *7*(1), 402.

Palmer, T. (2014). Climate Forecasting: Build High-Resolution Global Climate Models. *Nature*, *515*(7527), 338–339.

Palmer, T. N. (2019). Stochastic Weather and Climate Models. *Nature Reviews Physics*, *1*(7), 463–471. https://doi.org/10.1038/s42254-019-0062-2

Parker, D. B. (1985). Learning Logic Technical Report Tr-47. *Center of Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge, MA*.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., & Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 8024–8035). Curran Associates, Inc.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-Learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perkins, W. A., Brenowitz, N. D., Bretherton, C. S., & Nugent, J. M. (2024). Emulation of Cloud Microphysics in a Climate Model. *Journal of Advances in Modeling Earth Systems*, *16*(4), e2023MS003851.

Phillips, N. A. (1956). The General Circulation of the Atmosphere: A Numerical Experiment. *Quarterly Journal of the Royal Meteorological Society*, *82*(352), 123–164.

Pincus, R., Mlawer, E. J., & Delamere, J. S. (2019). Balancing Accuracy, Efficiency, and Flexibility in Radiation Calculations for Dynamical Models. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3074–3089. https://doi.org/10.1029/2019MS001621

Pincus, R., Iacono, M. J., Alexeev, D., Adamidis, P., Hillman, B. R., Norman, M., Pfister, E., Polonsky, I. N., Romero, N. A., Kosukhin, S. S., & Wehe, A. (2023). RTE+RRTMGP.

Prein, A. F., Langhans, W., Fosser, G., Ferrone, A., Ban, N., Goergen, K., Keller, M., Tölle, M., Gutjahr, O., Feser, F., et al. (2015). A Review on Regional Convection-Permitting Climate Modeling: Demonstrations, Prospects, and Challenges. *Reviews of geophysics*, *53*(2), 323–361.

Prill, F., Reinert, D., Rieger, D., & Zängl, G. (2022). ICON Tutorial. *ICON*, *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

Putman, W. M., & Suarez, M. (2011). Cloud-System Resolving Simulations with the NASA Goddard Earth Observing System Global Atmospheric Model (GEOS-5). *Geophysical Research Letters*, *38*(16).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al. (2018). Improving Language Understanding by Generative Pre-Training.

Ramadhan, A., Marshall, J., Souza, A., Lee, X. K., Piterbarg, U., Hillier, A., LeClaire Wagner, G., Rackauckas, C., Hill, C., Campin, J.-M., & Ferrari, R. (2020). Capturing Missing Physics in Climate Model Parameterizations Using Neural Differential Equations. *arXiv e-prints*, arXiv:2010.12559. https://doi.org/10.48550/arXiv.2010.12559

Randall, D., Khairoutdinov, M., Arakawa, A., & Grabowski, W. (2003). Breaking the Cloud Parameterization Deadlock. *Bulletin of the American Meteorological Society*, *84*(11), 1547–1564.

Randall, D. A. (2013). Beyond Deadlock. *Geophysical Research Letters*, *40*(22), 5970–5976.

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep Learning to Represent Subgrid Processes in Climate Models. *Proceedings of the National Academy of Sciences*, *115*(39), 9684–9689.

Rasp, S., & Thuerey, N. (2021). Data-driven Medium-range Weather Prediction with a Resnet Pretrained on Climate Simulations: A New Model for Weatherbench. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002405.

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., Sanchez-Gonzalez, A., Yang, V., Carver, R., Agrawal, S., Chantry, M., Ben Bouallegue, Z., Dueben, P., Bromberg, C., Sisk, J., Barrington, L., Bell, A., & Sha, F. (2023). WeatherBench 2: A Benchmark for the next Generation of Data-Driven Global Weather Models, arXiv:2308.15560. https://doi.org/10.48550/arXiv.2308.15560

Rasp, S. (2020). Coupled Online Learning as a Way to Tackle Instabilities and Biases in Neural Network Parameterizations: General Algorithms and Lorenz 96 Case Study (v1.0). *Geoscientific Model Development*, *13*(5), 2185–2196. https://doi.org/10.5194/gmd-13-2185-2020

Ray, D., Ramaswamy, H., Patel, D. V., & Oberai, A. A. (2022). The Efficacy and Generalizability of Conditional GANs for Posterior Inference in Physics-Based Inverse Problems.

Reick, C. H., Gayler, V., Goll, D., Hagemann, S., Heidkamp, M., Nabel, J. E., Raddatz, T., Roeckner, E., Schnur, R., & Wilkenskjeld, S. (2021). JSBACH 3-the Land Component of the MPI Earth System Model: Documentation of Version 3.2.

Righi, M., Andela, B., Eyring, V., Lauer, A., Predoi, V., Schlund, M., Vegas-Regidor, J., Bock, L., Brötz, B., de Mora, L., Diblen, F., Dreyer, L., Drost, N., Earnshaw, P., Hassler, B., Koldunov, N., Little, B., Loosveldt Tomas, S., & Zimmermann, K. (2020). Earth System

Model Evaluation Tool (ESMValTool) v2.0 – Technical Overview. *Geoscientific Model Development*, *13*(3), 1179–1199. https://doi.org/10.5194/gmd-13-1179-2020

Roberts, C. D., Senan, R., Molteni, F., Boussetta, S., Mayer, M., & Keeley, S. P. E. (2018). Climate Model Configurations of the ECMWF Integrated Forecasting System (ECMWF-IFS Cycle 43r1) for HighResMIP. *Geoscientific Model Development*, *11*(9), 3681–3712. https://doi.org/10.5194/gmd-11-3681-2018

Roca, R., Fiolleau, T., & Bouniol, D. (2017). A Simple Model of the Life Cycle of Mesoscale Convective Systems Cloud Shield in the Tropics. *Journal of Climate*, *30*(11), 4283–4298. https://doi.org/10.1175/JCLI-D-16-0556.1

Roehrig, R., Beau, I., Saint-Martin, D., Alias, A., Decharme, B., Guérémy, J.-F., Voldoire, A., Abdel-Lathif, A. Y., Bazile, E., Belamari, S., et al. (2020). The CNRM Global Atmosphere Model ARPEGE-Climat 6.3: Description and Evaluation. *Journal of Advances in Modeling Earth Systems*, *12*(7), e2020MS002075.

Romps, D. M., & Charn, A. B. (2015). Sticky Thermals: Evidence for a Dominant Balance between Buoyancy and Drag in Cloud Updrafts. *Journal of the Atmospheric Sciences*, *72*(8), 2890–2901. https://doi.org/10.1175/JAS-D-15-0042.1

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 234–241.

Ross, A., Li, Z., Perezhogin, P., Fernandez-Granda, C., & Zanna, L. (2023). Benchmarking of Machine Learning Ocean Subgrid Parameterizations in an Idealized Model. *Journal of Advances in Modeling Earth Systems*, *15*(1).

Rotunno, R., Klemp, J. B., & Weisman, M. L. (1988). A Theory for Strong, Long-Lived Squall Lines. *Journal of Atmospheric Sciences*, *45*(3), 463–485. https://doi.org/10.1175/1520-0469(1988)045<0463:ATFSLL>2.0.CO;2

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning Representations by Back-Propagating Errors. *nature*, *323*(6088), 533–536.

Sanford, C., Kwa, A., Watt-Meyer, O., Clark, S. K., Brenowitz, N., McGibbon, J., & Bretherton, C. (2023). Improving the Reliability of ML-Corrected Climate Models With Novelty Detection. *Journal of Advances in Modeling Earth Systems*, *15*(11), e2023MS003809. https://doi.org/10.1029/2023MS003809

Sarauer, E., Schwabe, M., Weiss, P., Lauer, A., Stier, P., & Eyring, V. (2024). Physics-Informed Machine Learning-Based Cloud Microphysics Parameterization for Earth System Models. *Physics-informed Machine Learning-based Cloud Microphysics parameterization for Earth System Models*.

Sarauer, E., Schwabe, M., Weiss, P., Lauer, A., Stier, P., & Eyring, V. (2025). A Physics-Informed Machine Learning Parameterization for Cloud Microphysics in ICON. *Environmental Data Science*, *4*, e40.

Satoh, M., Tomita, H., Yashiro, H., Miura, H., Kodama, C., Seiki, T., Noda, A. T., Yamada, Y., Goto, D., Sawada, M., et al. (2014). The Non-Hydrostatic Icosahedral Atmospheric Model: Description and Development, Progress in Earth and Planetary Science, 1, 18.

Satoh, M., Stevens, B., Judt, F., Khairoutdinov, M., Lin, S.-J., Putman, W. M., & Düben, P. (2019). Global Cloud-Resolving Models. *Current Climate Change Reports*, *5*(3), 172–184.

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 Syst. Dynam.*, *11*(4), 1233–1258. https://doi.org/10.5194/esd-11-1233-2020

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017). Climate Goals and Computing the Future of Clouds. *Nature Climate Change*, *7*(1), 3–5. https://doi.org/10.1038/nclimate3190

Schneider, T., Leung, L. R., & Wills, R. C. (2024). Opinion: Optimizing Climate Models with Process Knowledge, Resolution, and Artificial Intelligence. *Atmospheric Chemistry and Physics*, *24*(12), 7041–7062.

Schröder, M., Danne, O., Falk, U., Niedorf, A., Preusker, R., Trent, T., Brockmann, C., Fischer, J., Hegglin, M., Hollmann, R., & Pinnock, S. (2023). A Combined High Resolution Global TCWV Product from Microwave and near Infrared Imagers - COMBI. *Satellite Application Facility on Climate Monitoring (CM SAF)*. https://doi.org/10.5676/EUM_SAF_CM/COMBI/V001

Schulzweida, U. (2022). CDO User Guide. *Zenodo*. https://doi.org/10.5281/zenodo.7112925

Schumacher, C., & Funk, A. (2023). Assessing Convective-Stratiform Precipitation Regimes in the Tropics and Extratropics with the GPM Satellite Radar. *Geophysical Research Letters*, *50*(14), e2023GL102786. https://doi.org/10.1029/2023GL102786

Shamekh, S., & Gentine, P. (2023). Learning Atmospheric Boundary Layer Turbulence. *Authorea Preprints*.

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/doi:10.1073/pnas.2216158120

Shapley, L. S. (1951). Notes on the N-Person Game—Ii: The Value of an n-Person Game.

Shen, X., & Meinshausen, N. (2024). Engression: Extrapolation through the Lens of Distributional Regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, qkae108. https://doi.org/10.1093/jrsssb/qkae108

Shenk, W. E. (1974). Cloud Top Height Variability of Strong Convective Cells. *Journal of Applied Meteorology and Climatology*, *13*(8), 917–922. https://doi.org/10.1175/1520-0450(1974)013<0917:CTHVOS>2.0.CO;2

Sherwood, S. C., Bony, S., & Dufresne, J.-L. (2014). Spread in Model Climate Sensitivity Traced to Atmospheric Convective Mixing. *Nature*, *505*(7481), 37–42. https://doi.org/10.1038/nature12829

Skamarock, W. C., Klemp, J. B., Duda, M. G., Fowler, L. D., Park, S.-H., & Ringler, T. D. (2012). A Multiscale Nonhydrostatic Atmospheric Model Using Centroidal Voronoi Tesselations and C-grid Staggering. *Monthly Weather Review*, *140*(9), 3090–3105.

Smagorinsky, J., Manabe, S., & Holloway Jr, J. L. (1965). Numerical Results from a Nine-Level General Circulation Model of the Atmosphere. *Monthly weather review*, *93*(12), 727–768.

Smagorinsky, J. (1963). General Circulation Experiments with the Primitive Equations: I. The Basic Experiment. *Monthly weather review*, *91*(3), 99–164.

Smolarkiewicz, P. K., Deconinck, W., Hamrud, M., Kühnlein, C., Mozdzynski, G., Szmelter, J., & Wedi, N. P. (2016). A Finite-Volume Module for Simulating Global All-Scale Atmospheric Flows. *Journal of Computational Physics*, *314*, 287–304.

Song, H.-J., Roh, S., & Park, H. (2021). Compound Parameterization to Improve the Accuracy of Radiation Emulator in a Numerical Weather Prediction Model. *Geophysical Research Letters*, *48*(20), e2021GL095043.

Stensrud, D. J. (2007). *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models*. Cambridge University Press.

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelmen, A., Golaz, J.-C., Bodas-Salcedo, A., Suzuki, K., Gabriel, P., & Haynes, J. (2010). Dreary State of Precipitation in Global Models. *Journal of Geophysical Research: Atmospheres*, *115*(D24). https://doi.org/10.1029/2010JD014532

Stevens, B., & Bony, S. (2013). What Are Climate Models Missing? *science*, *340*(6136), 1053–1054.

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., Salzmann, M., Schmidt, H., Bader, J., & Block, K. (2013). Atmospheric Component of the MPI-M Earth System Model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172.

Stevens, B., Ament, F., Bony, S., Crewell, S., Ewald, F., Gross, S., Hansen, A., Hirsch, L., Jacob, M., & Kölling, T. (2019a). A High-Altitude Long-Range Aircraft Configured as a Cloud Observatory: The NARVAL Expeditions. *Bulletin of the American Meteorological Society*, *100*(6), 1061–1077.

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., Düben, P., Judt, F., Khairoutdinov, M., & Klocke, D. (2019b). DYAMOND: The DYnamics of the Atmospheric General Circulation Modeled On Non-hydrostatic Domains. *Progress in Earth and Planetary Science*, *6*(1), 1–17.

Stevens, B., Acquistapace, C., Hansen, A., Heinze, R., Klinger, C., Klocke, D., Rybka, H., Schubotz, W., Windmiller, J., Adamidis, P., et al. (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.

Stull, R. B. (1988). *An Introduction to Boundary Layer Meteorology*. https://doi.org/10.1007/978-94-009-3027-8

Sui, C.-H., & Yanai, M. (1986). Cumulus Ensemble Effects on the Large-Scale Vorticity and Momentum Fields of GATE. Part I: Observational Evidence. *Journal of Atmospheric Sciences*, *43*(15), 1618–1642.

Sukovich, E. M., Ralph, F. M., Barthold, F. E., Reynolds, D. W., & Novak, D. R. (2014). Extreme Quantitative Precipitation Forecast Performance at the Weather Prediction Center from 2001 to 2011. *Weather and Forecasting*, *29*(4), 894–911. https://doi.org/10.1175/WAF-D-13-00061.1

Sun, Y. Q., Pahlavan, H. A., Chattopadhyay, A., Hassanzadeh, P., Lubis, S. W., Alexander, M. J., Gerber, E. P., Sheshadri, A., & Guan, Y. (2024). Data Imbalance, Uncertainty Quantification, and Transfer Learning in Data-Driven Parameterizations: Lessons from

the Emulation of Gravity Wave Momentum Transport in WACCM. *Journal of Advances in Modeling Earth Systems*, *16*(7), e2023MS004145.

Takasuka, D., Satoh, M., Miyakawa, T., Kodama, C., Klocke, D., Stevens, B., Vidale, P. L., & Terai, C. R. (2024). A Protocol and Analysis of Year-Long Simulations of Global Storm-Resolving Models and Beyond. *Progress in Earth and Planetary Science*, *11*(1), 66.

Tallec, C., & Ollivier, Y. (2018). Can Recurrent Neural Networks Warp Time?

Tang, S., Wang, S., Jiang, J., & Zheng, Y. (2025). Incorporating Causality into Deep Learning Architectures to Improve Flash Drought Forecasts. *Water Resources Research*, *61*(10), e2024WR039470.

Tao, W.-K., Chern, J.-D., Atlas, R., Randall, D., Khairoutdinov, M., Li, J.-L., Waliser, D. E., Hou, A., Lin, X., Peters-Lidard, C., et al. (2009). A Multiscale Modeling System: Developments, Applications, and Critical Issues. *Bulletin of the American Meteorological Society*, *90*(4), 515–534.

Tibshirani, R. (2018). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tiedtke, M. I. C. H. A. E. L. (1989). A Comprehensive Mass Flux Scheme for Cumulus Parameterization in Large-Scale Models. *Monthly weather review*, *117*(8), 1779–1800.

Tomita, H., Miura, H., Iga, S.-i., Nasuno, T., & Satoh, M. (2005). A Global Cloud-Resolving Simulation: Preliminary Results from an Aqua Planet Experiment. *Geophysical Research Letters*, *32*(8).

Tompkins, A. M. (2000). The Impact of Dimensionality on Long-Term Cloud-Resolving Model Simulations. *Monthly Weather Review*, *128*(5), 1521–1535.

Tulich, S. N. (2015). A Strategy for Representing the Effects of Convective Momentum Transport in Multiscale Models: Evaluation Using a New Superparameterized Version of the Weather Research and Forecast Model (SP-WRF). *Journal of Advances in Modeling Earth Systems*, *7*(2), 938–962.

Ueyama, R., & Deser, C. (2008). A Climatology of Diurnal and Semidiurnal Surface Wind Variations over the Tropical Pacific Ocean Based on the Tropical Atmosphere Ocean Moored Buoy Array. *Journal of Climate*, *21*(4), 593–607. https://doi.org/10.1175/JCLI1666.1

Ukkonen, P., & Chantry, M. (2024). Representing Sub-Grid Processes in Weather and Climate Models via Sequence Learning. *Authorea Preprints*.

Ukkonen, P., & Chantry, M. (2025). Vertically Recurrent Neural Networks for Sub-Grid Parameterization. *Journal of Advances in Modeling Earth Systems*, *17*(6), e2024MS004833. https://doi.org/10.1029/2024MS004833

Ukkonen, P. (2022). Exploring Pathways to More Accurate Machine Learning Emulation of Atmospheric Radiative Transfer. *Journal of Advances in Modeling Earth Systems*, *14*(4), e2021MS002875.

Walters, D., Baran, A. J., Boutle, I., Brooks, M., Earnshaw, P., Edwards, J., Furtado, K., Hill, P., Lock, A., Manners, J., Morcrette, C., Mulcahy, J., Sanchez, C., Smith, C., Stratton, R.,

Tennant, W., Tomassini, L., Van Weverberg, K., Vosper, S., Willett, M., Browse, J., Bushell, A., Carslaw, K., Dalvi, M., Essery, R., Gedney, N., Hardiman, S., Johnson, B., Johnson, C., Jones, A., Jones, C., Mann, G., Milton, S., Rumbold, H., Sellar, A., Ujiie, M., Whitall, M., Williams, K., & Zerroukat, M. (2019). The Met Office Unified Model Global Atmosphere 7.0/7.1 and JULES Global Land 7.0 Configurations. *Geoscientific Model Development*, *12*(5), 1909–1963. https://doi.org/10.5194/gmd-12-1909-2019

Wang, Y., Zhang, G. J., & Craig, G. C. (2016). Stochastic Convective Parameterization Improving the Simulation of Tropical Precipitation Variability in the NCAR CAM5. *Geophysical Research Letters*, *43*(12), 6612–6619. https://doi.org/10.1002/2016GL069818

Wang, J., Balaprakash, P., & Kotamarthi, R. (2019). Fast Domain-Aware Neural Network Emulation of a Planetary Boundary Layer Parameterization in a Numerical Weather Forecast Model. *Geoscientific Model Development*, *12*(10), 4261–4274.

Wang, P., Yuval, J., & O'Gorman, P. A. (2022a). Non-Local Parameterization of Atmospheric Subgrid Processes with Neural Networks. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2022MS002984.

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

Wang, L.-Y., & Tan, Z.-M. (2023). Deep Learning Parameterization of the Tropical Cyclone Boundary Layer. *Journal of Advances in Modeling Earth Systems*, e2022MS003034.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., & Bretherton, C. S. (2021). Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555.

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., Harris, L., & Bretherton, C. S. (2024). Neural Network Parameterization of Subgrid-Scale Physics From a Realistic Geography Global Storm-Resolving Simulation. *Journal of Advances in Modeling Earth Systems*, *16*(2), e2023MS003668. https://doi.org/10.1029/2023MS003668

Wedi, N. P., Bauer, P., Denoninck, W., Diamantakis, M., Hamrud, M., Kuhnlein, C., Malardel, S., Mogensen, K., Mozdzynski, G., & Smolarkiewicz, P. K. (2015). The Modelling Infrastructure of the Integrated Forecasting System: Recent Advances and Future Challenges.

Weisman, M. L., Skamarock, W. C., & Klemp, J. B. (1997). The Resolution Dependence of Explicitly Modeled Convective Systems. *Monthly Weather Review*, *125*(4), 527–548.

Werbos, P. J. (2005). Applications of Advances in Nonlinear Sensitivity Analysis. *System Modeling and Optimization: Proceedings of the 10th IFIP Conference New York City, USA, August 31–September 4, 1981*, 762–770.

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.

Wu, S., & Easterbrook, S. M. (2025). Bridging the Gap: Climate Scientists' Concerns and Expectations for Machine Learning. *Proceedings of the ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies*, 284–296.

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.

Yanai, M., Chu, J. H., Stark, T. E., & Nitta, T. (1976). Response of Deep and Shallow Tropical Maritime Cumuli to Large-Scale Processes. *Journal of Atmospheric Sciences*, *33*(6), 976–991.

Yang, Q., Hannah, W. M., & Leung, L. R. (2022). Convective Momentum Transport and Its Impact on the Madden-Julian Oscillation in E3SM-MMF. *Journal of Advances in Modeling Earth Systems*, *14*(11), e2022MS003206.

Yano, J.-I., & Plant, R. S. (2012). Convective Quasi-Equilibrium. *Reviews of geophysics*, *50*(4). https://doi.org/10.1029/2011RG000378

Yao, Y., Zhong, X., Zheng, Y., & Wang, Z. (2023). A Physics-Incorporated Deep Learning Framework for Parameterization of Atmospheric Radiative Transfer. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003445. https://doi.org/10.1029/2022MS003445

Yu, F., & Koltun, V. (2015). Multi-Scale Context Aggregation by Dilated Convolutions. *arXiv e-prints*, arXiv:1511.07122. https://doi.org/10.48550/arXiv.1511.07122

Yu, S., Hannah, W., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., Behrens, G., Busecke, J., Loose, N., Stern, C., Beucler, T., Harrop, B., Hillman, B., Jenney, A., Ferretti, S. L., Liu, N., Anandkumar, A., Brenowitz, N., Eyring, V., Geneva, N., Gentine, P., Mandt, S., Pathak, J., Subramaniam, A., Vondrick, C., Yu, R., Zanna, L., Zheng, T., Abernathey, R., Ahmed, F., Bader, D., Baldi, P., Barnes, E., Bretherton, C., Caldwell, P., Chuang, W., Han, Y., HUANG, Y. U., Iglesias-Suarez, F., Jantre, S., Kashinath, K., Khairoutdinov, M., Kurth, T., Lutsko, N., Ma, P.-L., Mooers, G., Neelin, J. D., Randall, D., Shamekh, S., Taylor, M., Urban, N., Yuval, J., Zhang, G., & Pritchard, M. (2023). ClimSim: A Large Multi-Scale Dataset for Hybrid Physics-ML Climate Emulation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, & S. Levine (Eds.), *Advances in Neural Information Processing Systems* (pp. 22070–22084, Vol. 36). Curran Associates, Inc.

Yu, S., Hu, Z., Subramaniam, A., Hannah, W., Peng, L., Lin, J., Bhouri, M. A., Gupta, R., Lütjens, B., Will, J. C., Behrens, G., Busecke, J. J. M., Loose, N., Stern, C. I., Beucler, T., Harrop, B., **Heuer**, **H.**, Hillman, B. R., Jenney, A., Liu, N., White, A., Zheng, T., Kuang, Z., Ahmed, F., Barnes, E., Brenowitz, N. D., Bretherton, C., Eyring, V., Ferretti, S., Lutsko, N., Gentine, P., Mandt, S., Neelin, J. D., Yu, R., Zanna, L., Urban, N. M., Yuval, J., Abernathey, R., Baldi, P., Chuang, W., Huang, Y., Iglesias-Suarez, F., Jantre, S., Ma, P.-L., Shamekh, S., Zhang, G., & Pritchard, M. (2025). ClimSim-online: A Large Multi-Scale Dataset and Framework for Hybrid Physics-ML Climate Emulation. *Journal of Machine Learning Research*, *26*(142), 1–85. http://jmlr.org/papers/v26/24-1014.html

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.

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of Neural Networks for Stable, Accurate and Physically Consistent Parameterization of Subgrid Atmospheric Processes with Good Performance at Reduced Precision. *Geophysical Research Letters*, *48*(6), e2020GL091363.

Yuval, J., & O'Gorman, P. A. (2023). Neural-Network Parameterization of Subgrid Momentum Transport in the Atmosphere. *Journal of Advances in Modeling Earth Systems*, *15*(4), e2023MS003606. https://doi.org/10.1029/2023MS003606

Yuval, J., Langmore, I., Kochkov, D., & Hoyer, S. (2024). Neural General Circulation Models Optimized to Predict Satellite-Based Precipitation Observations. *arXiv preprint arXiv:2412.11973*.

Zängl, G., Reinert, D., Rípodas, P., & Baldauf, M. (2015). The ICON (ICOsahedral Non-hydrostatic) Modelling Framework of DWD and MPI-M: Description of the Non-Hydrostatic Dynamical Core. *Quarterly Journal of the Royal Meteorological Society*, *141*(687), 563–579. https://doi.org/10.1002/qj.2378

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., & Taylor, K. E. (2020). Causes of Higher Climate Sensitivity in CMIP6 Models. *Geophysical Research Letters*, *47*(1), e2019GL085782.

Zhang, G. J., & 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.

Zhang, G. J., & McFarlane, N. A. (2019). Sensitivity of Climate Simulations to the Parameterization of Cumulus Convection in the Canadian Climate Centre General Circulation Model. In *Data, Models and Analysis* (pp. 145–168). Routledge.

Zhang, Y., & Rossow, W. B. (2023). Global Radiative Flux Profile Data Set: Revised and Extended. *Journal of Geophysical Research: Atmospheres*, *128*(5), e2022JD037340. https://doi.org/10.1029/2022JD037340

Zhong, X., Yu, X., & Li, H. (2024). Machine Learning Parameterization of the Multi-Scale Kain–Fritsch (MSKF) Convection Scheme and Stable Simulation Coupled in the Weather Research and Forecasting (WRF) Model Using WRF–ML v1. 0. *Geoscientific Model Development*, *17*(9), 3667–3685.

Acknowledgments

Funding for this thesis was provided by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187). This thesis used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID 1179 (USMILE). Furthermore, the author gratefully acknowledges the Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS (Jülich Supercomputing Centre 2021) at the Jülich Supercomputing Centre (JSC).

I would like to thank Prof. Dr. Markus Rapp and Prof. Dr. Veronika Eyring for the opportunity to work on my doctoral thesis at the Institute of Atmospheric Physics at DLR Oberpfaffenhofen in the "Earth System Model Evaluation and Analysis (EVA)" department.

I would like to begin by expressing my great appreciation to my supervisor and primary examiner, Prof. Dr. Veronika Eyring, for her support, guidance, and mentorship throughout my Ph.D. project. I am profoundly grateful for the opportunities she provided me to participate in international conferences, summer schools, and research visits, which enabled me to collaborate with scientists around the world and broaden my scientific perspective. I sincerely thank my secondary examiner, Prof. Dr. Pierre Gentine, for his insightful advice, stimulating ideas, and the possibility to visit his research group at Columbia University.

Special thanks go to Dr. Mierk Schwabe, whose mentorship has been instrumental from the very beginning of this project. I am really grateful for her constant availability, thoughtful feedback, scientific guidance, and for teaching me the fundamentals of academic writing. I am also deeply thankful to Prof. Dr. Tom Beucler for our close collaboration on the second study. His innovative ideas, deep insights, and general scientific knowledge were invaluable to the work, and I also truly appreciate giving me the opportunity to meet and collaborate with his network of outstanding international researchers.

I extend my appreciation to all my co-authors: Dr. Mierk Schwabe, Prof. Dr. Tom Beucler, Dr. Marco Giorgetta, Dr. Julien Savre, Dr. Manuel Schlund, Prof. Dr. Pierre Gentine, and Prof. Dr. Veronika Eyring. For their collaboration, expertise, and the significant contributions they made to improving the quality of our manuscripts.

I am grateful to Dr. Mierk Schwabe, Dr. Julien Savre, Dr. Katja Weigel, and my brother Fin Heuer for their careful proofreading of this thesis and their constructive suggestions, which greatly enhanced its clarity and coherence. I also thank Dr. Manuel Schlund for developing the excellent LaTeX template used for this thesis, it made the writing process much smoother.

To my colleagues in the EVA department and my fellow PhD students: thank you for the stimulating scientific exchanges, the enjoyable coffee breaks, and the many shared group activities, including our visits to the beer garden. It has been a privilege to work alongside such a talented and supportive community.

Finally, I would like to express my heartfelt thanks to my friends and family, especially my parents, Sabine and Eckbert, and to my brother Fin, for their encouragement, support, and belief in me. Knowing I could always count on their support has made this journey possible.