JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2024MS004398

Interpretable Multiscale Machine Learning-Based Parameterizations of Convection for ICON

Helge Heuer[1], Mierk Schwabe[1], Pierre Gentine[2], Marco A. Giorgetta[3], and Veronika Eyring[1,4]

[1]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Wessling, Germany, [2]Center for Learning the Earth with Artificial Intelligence and Physics (LEAP), Columbia University, New York, NY, USA, [3]Max Planck Institute for Meteorology, Hamburg, Germany, [4]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

Key Points:

- We train/benchmark machine learning models on convective fluxes derived from realistic coarse-grained data of storm-resolving simulations
- Shapley values reveal that the best offline model, a U-Net, learns non-causal links to precipitation and shows poor online performance
- A model, without non-causal precipitation connections, runs more stable coupled to ICON and indicates better precipitation predictions

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

H. Heuer,
helge.heuer@dlr.de

Citation:

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*, e2024MS004398. https://doi.org/10.1029/2024MS004398

Received 12 APR 2024
Accepted 1 AUG 2024

Abstract Machine learning (ML)-based parameterizations have been developed for Earth System Models (ESMs) with the goal to better represent subgrid-scale processes or to accelerate computations. ML-based parameterizations within hybrid ESMs have successfully learned subgrid-scale processes from short high-resolution simulations. However, most studies used a particular ML method to parameterize the subgrid tendencies or fluxes originating from the compound effect of various small-scale processes (e.g., radiation, convection, gravity waves) in mostly idealized settings or from superparameterizations. Here, we use a filtering technique to explicitly separate convection from these processes in simulations with the Icosahedral Non-hydrostatic modeling framework (ICON) in a realistic setting and benchmark various ML algorithms against each other offline. We discover that an unablated U-Net, while showing the best offline performance, learns reverse causal relations between convective precipitation and subgrid fluxes. While we were able to connect the learned relations of the U-Net to physical processes this was not possible for the non-deep learning-based Gradient Boosted Trees. The ML algorithms are then coupled online to the host ICON model. Our best online performing model, an ablated U-Net excluding precipitating tracer species, indicates higher agreement for simulated precipitation extremes and mean with the high-resolution simulation compared to the traditional scheme. However, a smoothing bias is introduced both in water vapor path and mean precipitation. Online, the ablated U-Net significantly improves stability compared to the non-ablated U-Net and runs stable for the full simulation period of 180 days. Our results hint to the potential to significantly reduce systematic errors with hybrid ESMs.

Plain Language Summary Due to their computational costs, it is currently not feasible to run more accurate high-resolution climate models on a global domain on climate (century) time-scales. However, high-accuracy climate simulations are needed for more robust and detailed projections of our future climate. Here, we develop and evaluate various machine learning-based convection parameterizations learned on reconstructed and coarse-grained high-resolution subgrid fluxes to solve this problem, and benchmark their performance. The data set is chosen from simulations of the Icosahedral Non-hydrostatic modeling framework (ICON) in a realistic setting of the tropical Atlantic and at storm-resolving resolutions. We focus only on convective subgrid fluxes that are isolated from other components. We improve the best ML algorithms further by excluding variables that cause unphysical correlations. Finally, we explain the learned relations of the best data-driven schemes based on physical process understanding, test their performance when coupled to the ICON model, and achieve stable coupled simulations for 180 days as well as improved precipitation predictions.

1. Introduction

General Circulation Models (GCMs) have been used since the late 1960s to answer scientific questions about our climate (Manabe & Wetherald, 1967; Phillips, 1956) and to project its expected changes, which are already felt across the globe (Eyring, Gillett, et al., 2021). Over time, these models gradually included more and more aspects and processes of the climate system and have evolved into Earth System Models (ESMs), including the carbon cycle and biogeochemical processes. However, the uncertainty of the simulated equilibrium climate sensitivity (ECS), that is, the response of global surface air temperature to a doubling of CO2 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–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–4°C (Forster et al., 2021).

A large portion of this uncertainty is attributed to cloud feedbacks (Ceppi & 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 & 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, Mishra, et al., 2021), such as biases in precipitation patterns (Christopoulos & Schneider, 2021; Fosser et al., 2024; Stephens et al., 2010), in the position and shape of the intertropical convergence zone (ITCZ) (Stevens, Satoh, et al., 2019), 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., 2020; Stevens, Satoh, et al., 2019).

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 & 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 multilayer perceptron (MLP) neural networks in a simplified aquaplanet setup to replace the super-parameterized physics in the SuperParameterized Community Atmosphere Model (SPCAM3) (Collins et al., 2006). Random Forests (RFs) have been used as well (O'Gorman & Dwyer, 2018; Yuval & 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 & O'Gorman, 2023). Wang et al. (2022) 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, 2023). 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 (Data set 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 explainable Artificial Intelligence (AI) 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 explainable AI 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

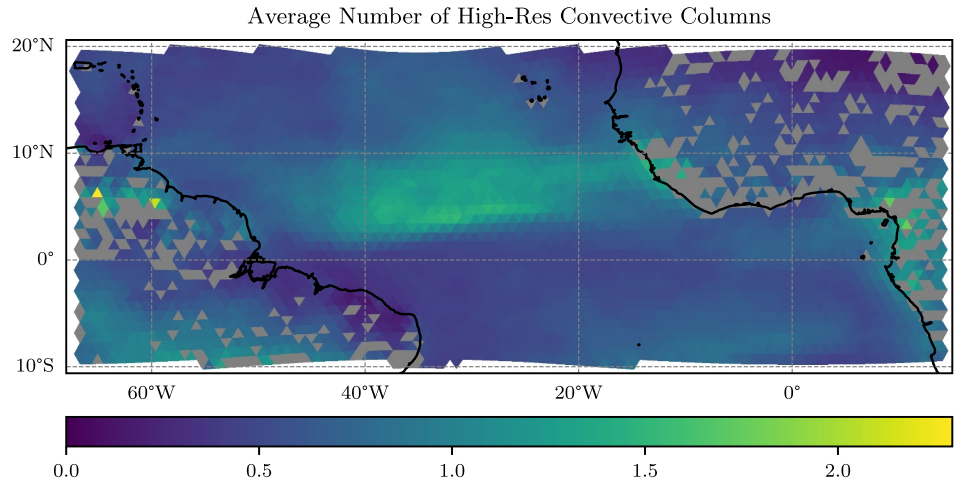

Figure 1. Average number of high-resolution convective cells per displayed low-resolution column and time frame as defined in Equation 4 in the studied tropical Atlantic region over the entire considered period of time. In the west the coastline of South America and some Caribbean islands can be seen and in the east the coastline of Africa. The low resolution grid has an approximate horizontal resolution of $\Delta x \approx 80$ km. Excluded columns are marked in gray.

in the retrieved models and can be used to evaluate the ML model, going beyond common metrics such as the root mean squared 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 paper is structured as follows. First, in Section 2 we describe the data, preprocessing, and coarse-graining method. Afterward, we introduce the machine learning methods in Section 3. Results of the offline evaluation/benchmarking of different machine learning models are then shown and their predictions interpreted using an explainable AI technique in Section 4. We will conclude Section 4 with an online stability test of the developed U-Net parameterizations. Finally, we discuss our results and give a conclusion of our work.

2. Data and Preprocessing

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, Ament, et al., 2019). 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 (December 2013/January 2014), and one from the boreal summer (August 2016). We use simulation data accompanying both expeditions. The horizontal resolution of the used simulations is $\Delta x \approx 2.5$ km (`R2B10` grid), and is available with an hourly output frequency. The simulations were performed with the Icosahedral Non-hydrostatic modeling framework (ICON) model (Giorgetta et al., 2018; Zängl et al., 2015), and for each day of the 2-month data set the simulations where initialized at 0000 UTZ and run for 36 hr. For this simulation the ICON model was used in its numerical weather prediction 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).

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 1 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 data set as described later in the coarse-graining section (Section 2.2) are marked in gray. The figure shows a clear pattern of the ITCZ (compare Stevens, Ament, et al., 2019; Figure 2) with an increased number of convective cells. Additionally, many convective cells can be found along the coast and over

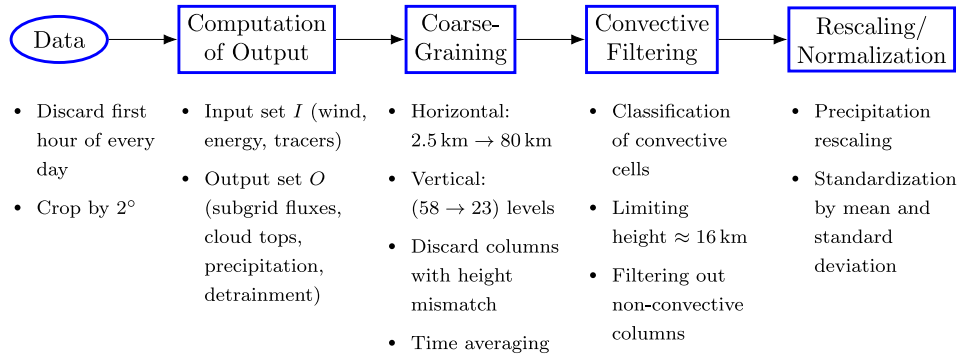

Figure 2. Summary of preprocessing steps. Starting from the original data, first the subgrid fluxes as well as 2D outputs, such as precipitation, were computed. After this, the data was coarse-grained and filtered for active convection. As a final preprocessing step, the data was rescaled and normalized.

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 1.

As a first preprocessing step we discarded the first hour of every day in the data set 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 1 was already cropped by the mentioned 2°.

To give a short overview of the preprocessing steps described below, Figure 2 depicts an overview of the various steps used, beginning with the original data set.

2.1. 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

$$h = c_p T + zg - L_v \cdot (q_c + q_r) - L_s \cdot (q_i + q_s + q_g), \tag{1}$$

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

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

The first eight variables with notation "$F_{\mathrm{var}}^{\mathrm{sg}}$" 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 ($q_{l,\mathrm{detr}}$, $q_{i,\mathrm{detr}}$), 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.

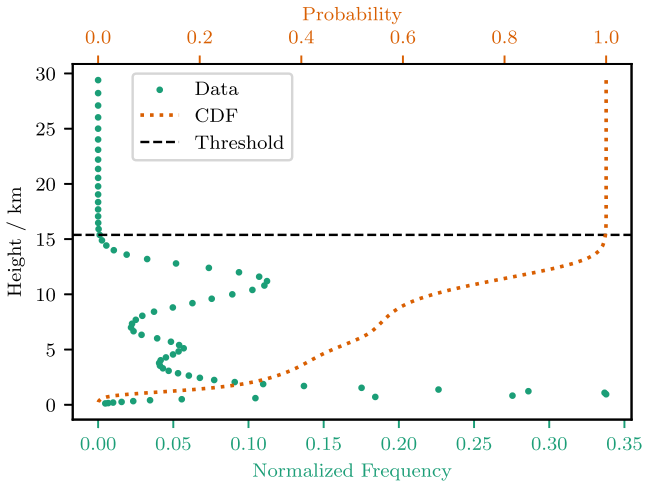

Figure 3. Probability distribution of convectively classified cells over altitude (green large dots) in the high-resolution data with ICON over the NARVAL region. The orange dashed line shows the cumulative distribution function (CDF) and the black dashed line represents the height up to which 99.9% of the convective cells are found. The bottom scale corresponds to the probability and the top to the cumulative distribution.

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:

$$F_u^{sg} = \overline{\rho w u} - \overline{\rho}\,\overline{w}\,\overline{u} = \overline{\rho}\,\overline{w'u'} + \overline{w}\,\overline{\rho'u'} + \overline{u}\,\overline{\rho'w'} + \overline{\rho'w'u'}. \quad (2)$$

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 2). The term on the right hand side in Equation 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 neighboring 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). 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 4). While there are different ways to estimate the detrainment of liquid/ice (Arakawa & Schubert, 1974; Zhang & McFarlane, 2019) we decided to follow Nordeng (1994); Baba and Giorgetta (2020) and calculated the fractional detrainment as

$$\delta = -\frac{1}{\sigma}\frac{\partial \sigma}{\partial z}, \quad (3)$$

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 & 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 for example, 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.

2.2. Coarse-Graining

The coarse-graining was done first in the horizontal and afterward 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. 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 data set.

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 afterward 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 & Houze, 1997). A 3 hr window is just about short enough to resolve the life cycle of such clouds and still allow a minimal smoothing of higher frequency variability.

2.3. 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 & Charn, 2015) are met:

$$q_l + q_i > 0.01g\,kg^1, \quad w > 0, \quad B \propto \theta_v - \overline{\theta_v} > 0, \tag{4}$$

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 conditions (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 & 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 1. 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.

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. 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 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.

2.4. 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:

$$P' = \ln\left(1 + \frac{P}{1kg\ m^{-2}h^{-}1}\right). \tag{5}$$

The reason for this is that precipitation intensities are typically represented by a heavily skewed (gamma) distribution (Martinez-Villalobos & 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 & Thuerey, 2021). Additionally, it is well known that coarse GCMs have a bias toward low intensity precipitation events (Moseley et al., 2016; Rasp et al., 2018). The rescaling should help mitigate some of this problem.

3. Machine Learning Models

As mentioned in the introduction, ML-based convection parameterizations have been developed using different kinds of methods. These include RFs (Limon & Jablonowski, 2023; O'Gorman & Dwyer, 2018; Yuval & 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 (CNNs) (Han et al., 2020, 2023), Residual Neural Networks (ResNets) (Wang et al., 2022), Generative Adversarial Networks (Nadiga et al., 2022), and Variational Encoders (VAEs)/Variational Auto Encoder Decoders (VEDs) (Behrens et al., 2022; Mooers et al., 2021). 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: Multilayer Perceptron, CNN, Residual Neural Network (He et al., 2016), and a CNN 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 S3 in Supporting Information S1). 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. (2022), 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 & Ba, 2014). For the implementation of all deep learning models we relied on the Pytorch library (Paszke et al., 2019).

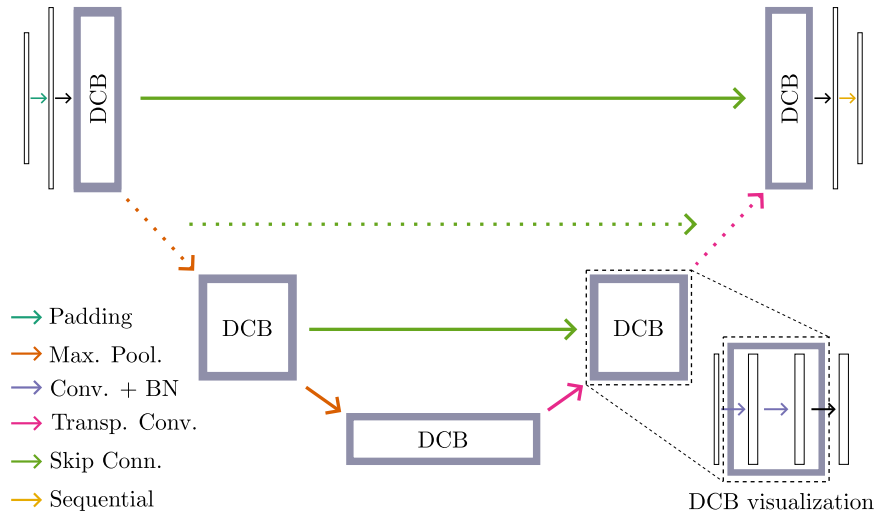

Figure 4. Visualization of the used U-Net architecture. The abbreviations DCB, Conv., Transp. Conv., and BN stand for double convolution block, convolutional layer, transpose convolutional layer, and batch normalization layer, respectively. The dotted lines mark the possibility for more blocks depending 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.

Furthermore, we decided to use a U-Net architecture, see Figure 4. 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 4) 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 & Kennard, 1970) regression. Additionally, we used three tree-based models. These include Random Forests (RF) (Breiman, 2001), Extra Trees (ET) (Geurts et al., 2006), and Gradient Boosted Trees (GBT) (Friedman, 2002). Further information about the different ML models can be found in Section S2 in Supporting Information S1.

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 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 hyperparameter optimization (HPO) on a subset of the data from five random days ($\sim 1.6 \cdot 10^5$ samples) because

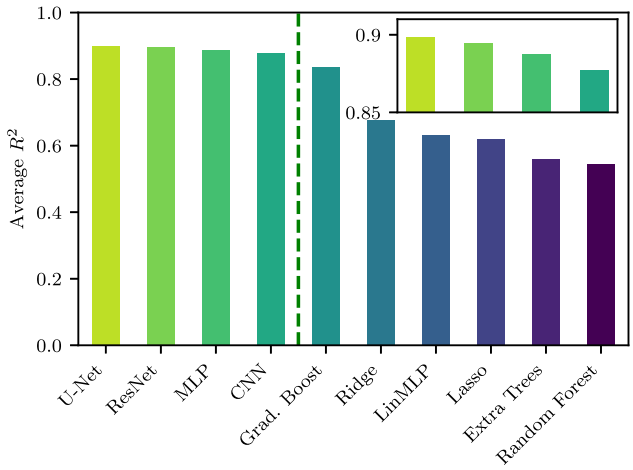

Figure 5. Coefficient of determination (R^2) on a test set for different types of models. All models were hyperparameter-optimized, and the best models were then trained on the whole data set. The deep learning methods are displayed on the left of the green dashed line and the non-deep learning methods 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.

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 S3 in Supporting Information S1.

4. 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 4.1. Afterward, in Section 4.2, 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 4.3.

The architecture of the best performing model, the U-Net, is first introduced in Section 4.1 and the ablation, improving online stability, is described in Section 4.2.

4.1. 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 8. 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 S3 in Supporting Information S1).

Figure 5 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 & Dwyer, 2018; Yuval & 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 & 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 GBT (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.

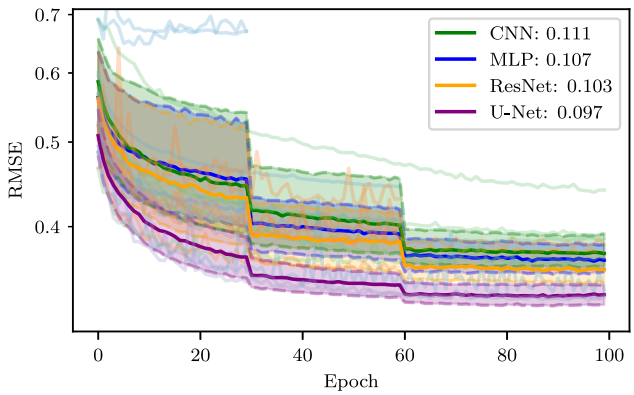

Figure 6. Root mean squared error during HPO on the validation set of the four different deep learning methods. The straight thick lines correspond to the median of the HPO ensemble, the shaded areas are drawn in between the first and third quartile. Additionally, 10 realizations for each DL method are 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).

The deep learning models outperform the other methods but, for example, 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 5 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 5,000 ensemble members in total. The resulting median/upper/lower quartile profiles can be seen in Figure 6. 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 S3 in Supporting Information S1. A visualization of the HPO and the training and validation process in general is shown in Figure S8 in Supporting Information S1. We notice that the U-Net has a consistently lower error than the 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 S1 in Supporting Information S1) and Figure S6 in Supporting Information S1, 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 6) 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 5. We first compare the U-Net and GBT flux predictions with the true values for $F_u^{\mathrm{sg}}, F_v^{\mathrm{sg}}, F_h^{\mathrm{sg}}, F_{q_c}^{\mathrm{sg}}$ over all levels. The results can be seen in Figure 7, and a corresponding plot showing the distribution for the remaining tracer subgrid fluxes can be seen in Figure S1 in Supporting Information S1. 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 for example, 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 & Deser, 2008). We assume that, therefore, 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 $F_{q_v}^{\mathrm{sg}}$. 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 5 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 8, 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.

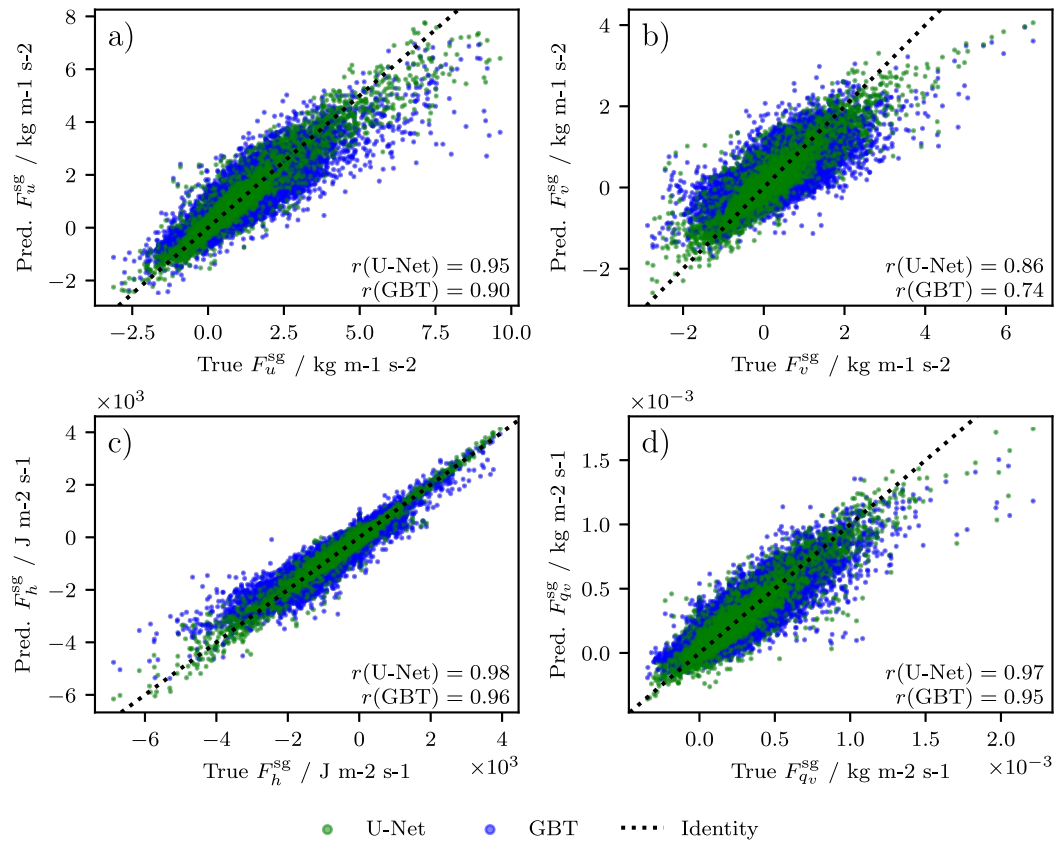

Figure 7. Scatter plot for the subgrid fluxes of (a) zonal, (b) meridional momentum, (c) liquid/ice water static energy, and (d) specific humidity. Data for the U-Net is shown in green, for the GBT in blue, and the diagonal is marked by a dotted line. The Pearson correlation coefficient r between the true and the predicted subgrid flux is noted in the lower right corner of each plot for both U-Net and GBT.

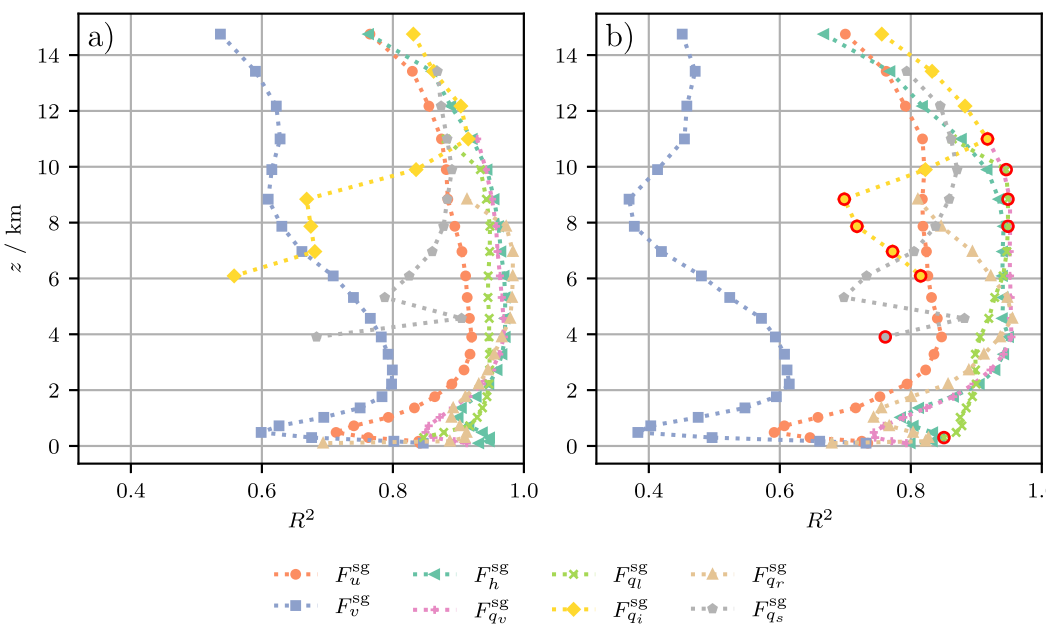

Figure 8. Average R^2 profile for all subgrid flux variables for (a) the U-Net and (b) GBT model. Data points where the GBT model actually has a higher R^2 than the U-Net are additionally marked by a red circle.

This method filters all levels which show significantly less variation compared to all other levels. Looking at Figure 8 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.

Comparing the plots in Figure 8, the two models show similar patterns as seen, for example, for the F_v^{sg} curve, but the GBT curves are mostly shifted toward 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 & 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 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, for example, 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 S2 in Supporting Information S1. 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 hr^{-1}, 19.75 mm hr^{-1}, and 16.83 mm hr^{-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 S3 in Supporting Information S1) we notice that both models have a lower error in the region of the ITCZ and an increase in error toward higher latitudes. This reflects the difference in the abundance of training data as seen in Figure 1.

4.2. 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 SHapley Additive exPlanations (SHAP) (Lundberg & 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 & Lee, 2017) as an efficient explainer for deep neural networks, and the `TreeExplainer`/`KernelExplainer` class for decision tree-based models such as GBT.

Figure 9a 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-

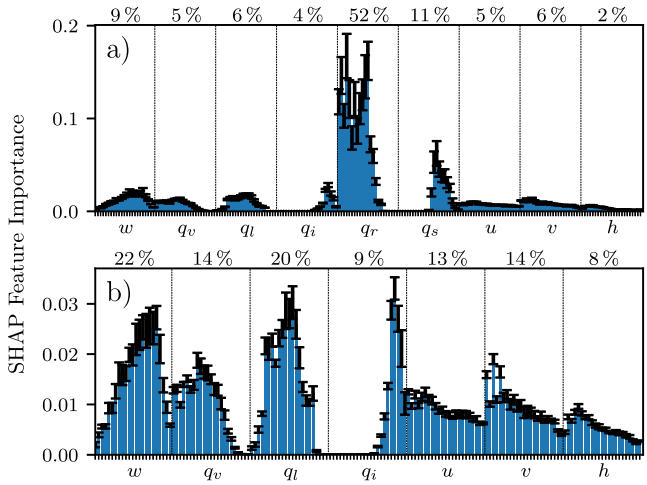

Figure 9. Feature importances (i.e., the mean absolute values of the calculated SHAP values) of input variables for (a) the full U-Net model, and (b) the ablated (without qr, qs) U-Net model. The mean feature importance is visualized by the height of the bar, and the standard deviation over five different 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.

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 S4 in Supporting Information S1). 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 9. 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 variables seen in Figure 9 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 9 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

$$\hat{x}_{ij} = \frac{x'_{ij}}{\max_k\left(|x'_{ik}|\right)}. \tag{6}$$

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

$$S_i = \langle \hat{x}_{ij} \cdot shap(x_{ij}, y) \rangle_j. \tag{7}$$

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.

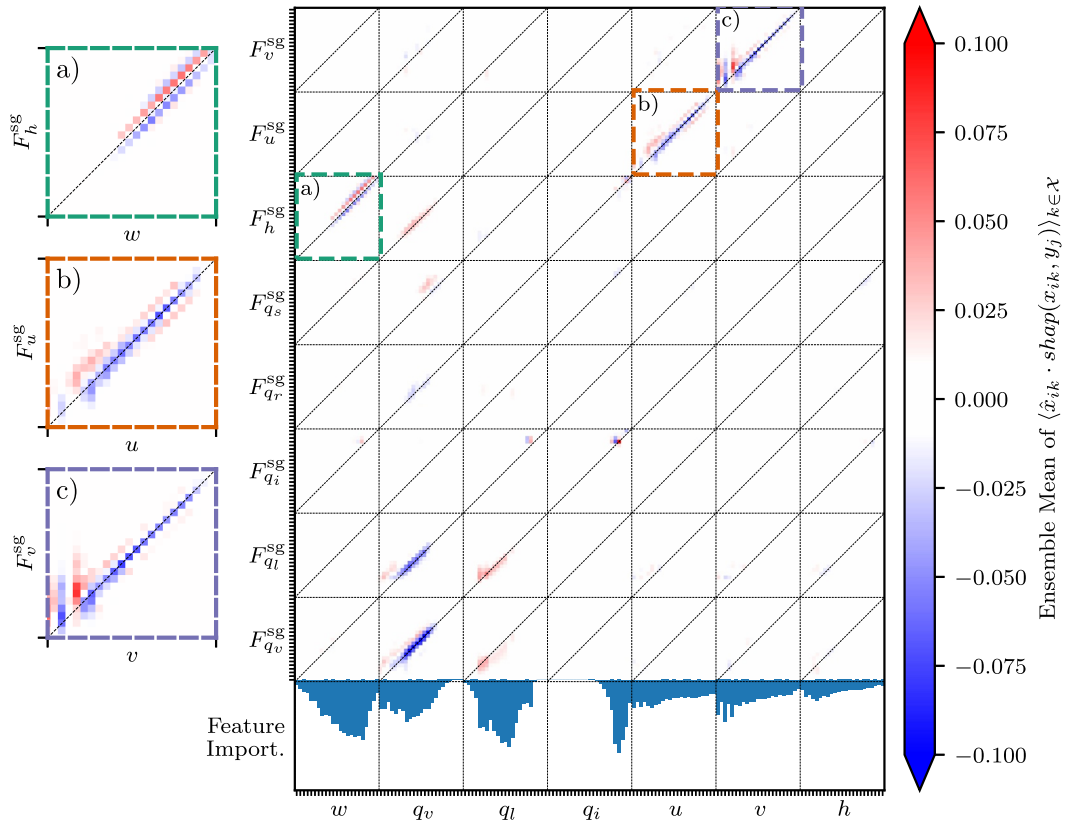

Figure 10. Ensemble mean of weighted SHAP values aggregated according to Equation 7 for the U-Net. The variables q_r, q_s were ablated. The height level for each variable is increasing from left to right/from bottom to top. The feature importance depicted in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes. Insets (a–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.

Applying this method to the trained U-Net model gives the matrix visualized in Figure 10. 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 $F_{q_v}^{sg}/F_{q_l}^{sg}$ 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 vapor. 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 data set 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_h^{sg}) 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 10. Inset (a) shows the influence of w on F_h^{sg}. The main pattern is in the upper layers where we can 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 & 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 ∼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 11. 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 4 for convective conditions in a grid cell. Therefore, it is reasonable that the network learns to pay attention to these inputs.

Since the weighted SHAP values displayed in Figure 10 consistently show vastly different patterns than in Figure 11, 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 & 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 10, 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.

4.3. 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

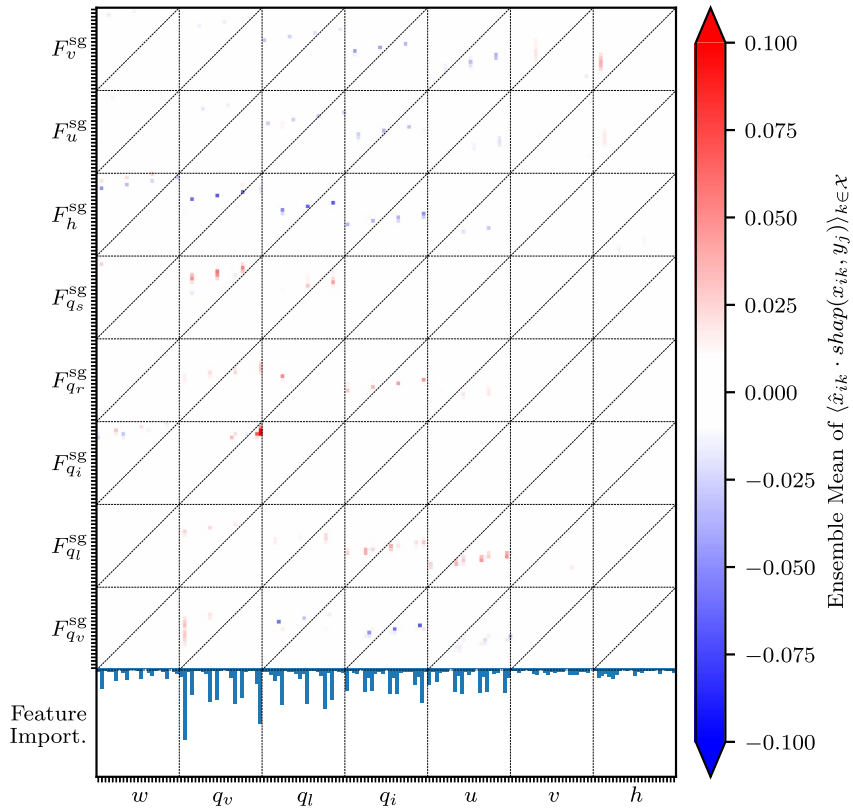

Figure 11. Ensemble mean of weighted SHAP values aggregated according to Equation 7 for the GBT model. The variables q_r, q_s were ablated. The feature importance depicted in the lower part of the figure shows the mean absolute SHAP values averaged over all target fluxes.

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 for example, changes in the training data set (Rasp, 2020) or the inclusion of variables on specific levels and the choice of the loss function (Brenowitz & Bretherton, 2018, 2019). Trial and error is often used to find stable schemes among the offline trained parameterizations (Wang et al., 2022). 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 & 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., 2022). A technical advantage of training on these data sets 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 for example, adjust for current compensating biases in the interplay of various parameterization schemes (Grundner et al., 2023). 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 for example, temperature, humidity, and also wind patterns differ considerably in the extratropics. Configurations 3 and 4 are applied closer to their training data set domain, that is, 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 1. 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.

We initialized a simulation from interpolated Integrated Forecasting System analysis data for the 01.01.1979 and ran the ICON model for 1 month. After this initialization phase we wrote out initial conditions for each day at 0000 UTC. With these initial conditions we started 10 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 2, 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 & 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 12 and for the spatial distribution of mean precipitation in Figure S7 in Supporting Information S1. In Figure 12, the distribution of precipitation over the first 2 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 2.

The spatial distribution (monthly means) of precipitation over the region where we have a high-resolution reference can be seen in Figure S7 in Supporting Information S1. 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

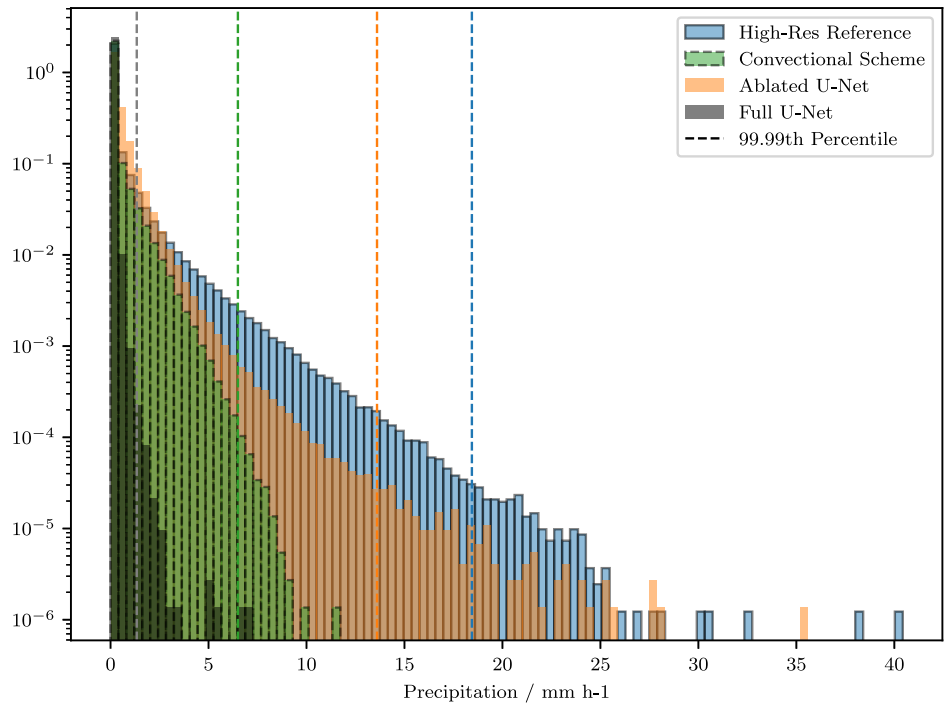

Figure 12. The precipitation distributions of the first 2 weeks over the tropics for the three simulations starting on 01.02.1979 for the full U-Net (configuration 4) in gray/dark green, the conventional cumulus scheme in green, and the ablated U-Net (configuration 3) in orange. Also, the precipitation 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.

precipitation. The conventional scheme shows a heavy land bias for the mean precipitation and shows too low precipitation values.

Figure 12 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 Figure 9 and Figure S4 in Supporting Information S1), 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 hr^{-1} for the NARVAL data, 13.07 mm hr^{-1} for the ablated U-Net, 6.67 mm hr^{-1} for the reference simulation, and only 1.34 mm hr^{-1} for the full U-Net.

Looking at the stability of the coupled simulations, Figure 13 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 (configuration 2) very quickly become unstable, after about 6–18 hr. 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 13. 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 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

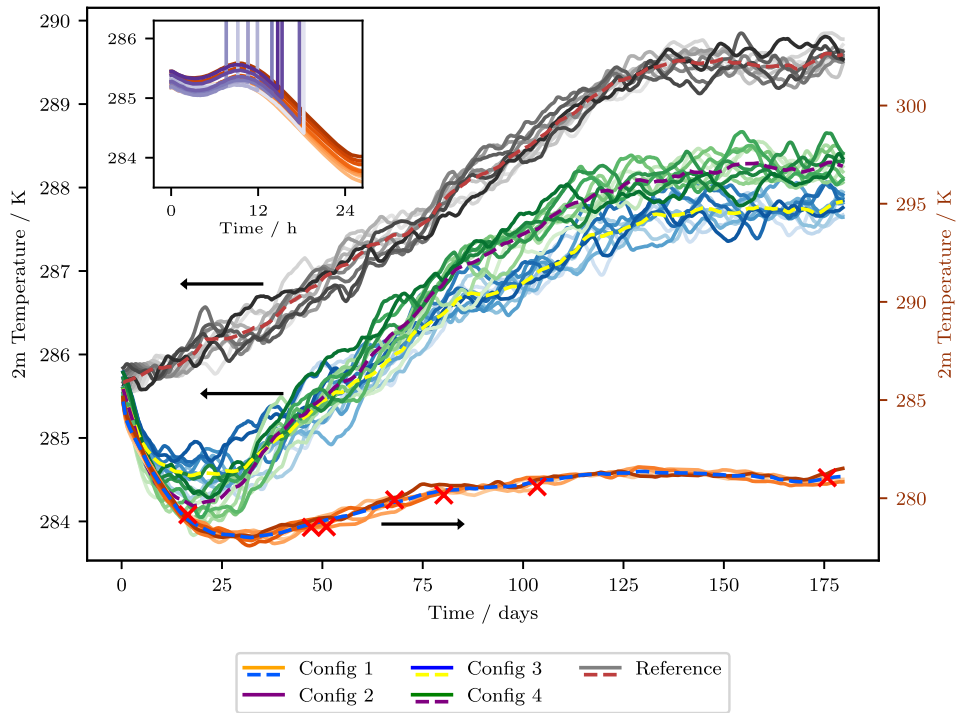

Figure 13. The stability of the ablated versus the full U-Net in form of a time series of the global mean air temperature on 2 m height over 180 days. For each defined configuration, the 10 realizations are drawn in orange, purple, blue and green colors, respectively. Solely for the full U-Net coupled globally (Configuration 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 2 m 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 hr 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 6 min as opposed to the more stable simulations with an output frequency of 6 hr in the main plot. For all of the data except the inset, a rolling mean over 24 hr 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.

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 (2 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 12 and heat is therefore transported less efficiently to higher levels. Comparing these two fully stable simulations to the reference simulation mean in gray, 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 14. 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.

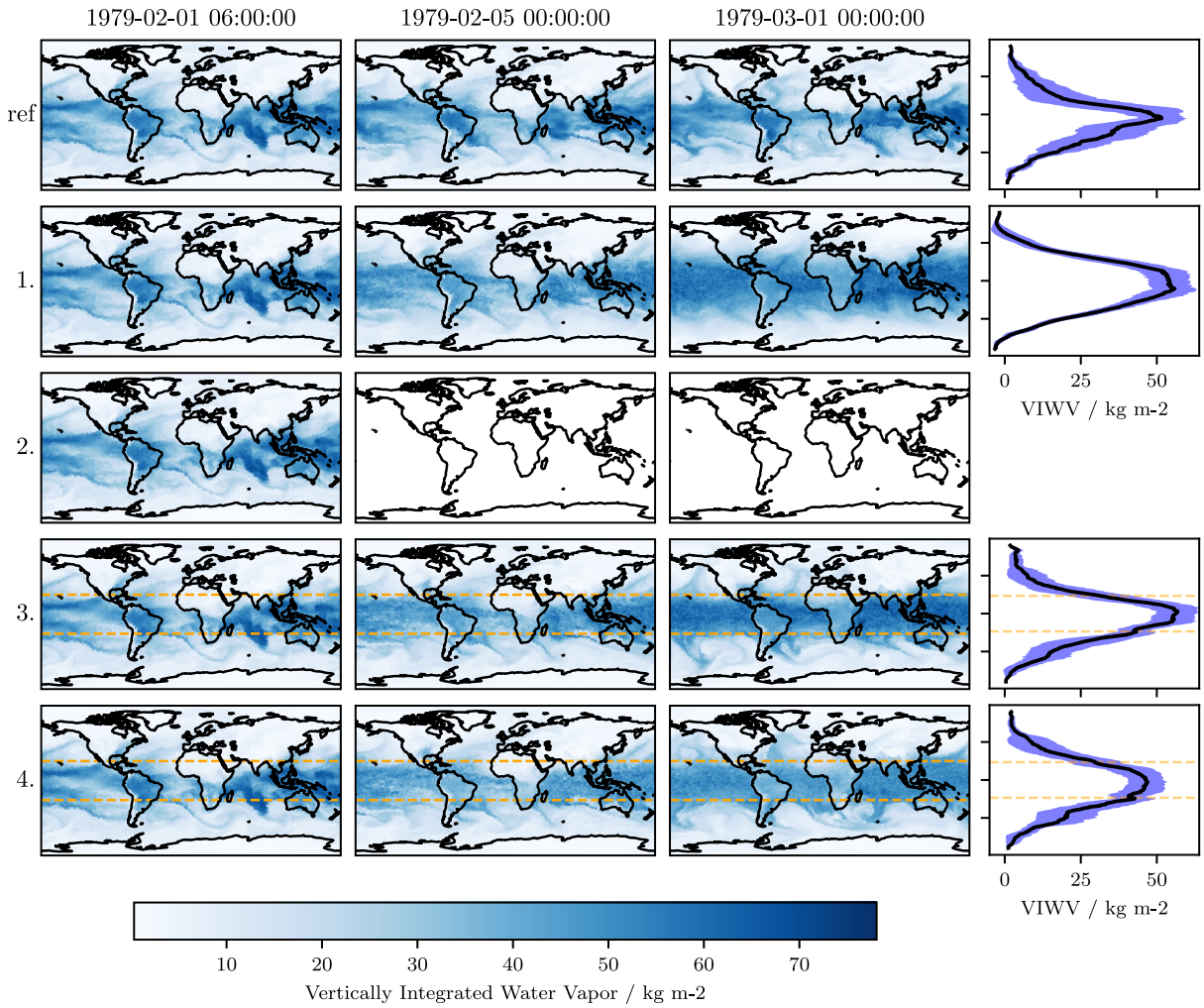

Figure 14. The vertically integrated water vapor for three simulation snapshot with convection parameterized with the conventional physical cumulus convection scheme of the ICON model as a reference (ref), (1) by the ablated U-Net, (2) by the full U-Net, (3) by the ablated U-Net applied only in the tropics, (4) by the full U-Net applied only in the tropics. For row (3, 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.

For the coupled full U-Net applied at all latitudes/longitudes we can only see one snapshot after 6 hr in Figure 14 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 14 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 8) 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, for example, 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 14 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 S4 in Supporting Information S1), leads to model blow-ups rather quickly when coupled globally, outside the training domain.

5. 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 6), 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, for example, Yuval and O'Gorman (2020); Yuval and O'Gorman (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 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 for example, 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 & 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 S4 in Supporting Information S1), 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 S5 in Supporting Information S1. 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, for example, 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, for example, 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 12) in contrast to the conventional parameterization and the full U-Net. The stable simulations are showing for example, 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 4.3. 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 for example, 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, for example, radiation. By doing this, we will avoid distributional shifts between the coarse-grained data set and the coarse simulations.

Data Availability Statement

The code is published under https://github.com/EyringMLClimateGroup/heuer23_ml_convection_parameterization and preserved (Heuer, 2024). The simulation data used to train and evaluate the machine learning algorithms amounts to several TB and can be reconstructed with the scripts provided in the GitHub repository. Access to the NARVAL data set was provided by the German Climate Computing Center (DKRZ) The software code for the ICON model is available from https://code.mpimet.mpg.de/projects/iconpublic.

Acknowledgments
Funding for this study was provided by the European Research Council (ERC) Synergy Grant "Understanding and Modeling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant 855187). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1179. The authors gratefully acknowledge 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). Furthermore, we thank the authors of Klocke et al. (2017) for creating and providing the high-resolution simulations of the tropical Atlantic used in this study. Open Access funding enabled and organized by Projekt DEAL.

References

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

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

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., & 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. https://doi.org/10.1175/1520-0469(1974)031<0674:ioacce>2.0.co;2

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. https://doi.org/10.1029/2019ms001732

Balestriero, R., Pesenti, J., & LeCun, Y. (2021). Learning in high dimension always amounts to extrapolation. *arXiv:2110.09485*. https://doi.org/10.48550/arXiv.2110.09485

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. https://doi.org/10.1029/2022MS003130

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., 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., et al. (2024). Climate-invariant machine learning. *Science Advances*, *10*(6), eadj7250. https://doi.org/10.1126/sciadv.adj7250

Bock, L., Lauer, A., Schlund, M., Barreiro, M., Bellouin, N., Jones, C., et al. (2020). Quantifying progress across different CMIP phases with the ESMValTool. *Journal of Geophysical Research: Atmospheres*, *125*(21). https://doi.org/10.1029/2019jd032321

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

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parametrizations of convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. https://doi.org/10.1175/jas-d-20-0082.1

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289–6298. https://doi.org/10.1029/2018gl078510

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

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., et al. (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

Cambridge-ICCS. (2024). Ftorch - A library for coupling (py)torch machine learning models to Fortran. Retrieved from https://github.com/Cambridge-ICCS/FTorch

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/10.1073/pnas.2026290118

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

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

Collins, W. D., Rasch, P. J., Boville, B. A., Hack, J. J., McCaa, J. R., Williamson, D. L., et al. (2006). The formulation and atmospheric simulation of the community atmosphere model version 3 (CAM3). *Journal of Climate*, *19*(11), 2144–2161. https://doi.org/10.1175/JCLI3760.1

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. (European Conference on Severe Storms 2002). https://doi.org/10.1016/S0169-8095(03)00047-4

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 CO_2. *Geophysical Research Letters*, *49*(8), e2022GL098174. https://doi.org/10.1029/2022GL098174

Eyring, V., Gillett, N., Rao, K. A., Barimalala, R., Parrillo, M. B., Bellouin, N., et al. (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 et al. (Eds.), (pp. 423–552). Cambridge University Press. https://doi.org/10.1017/9781009157896.005

Eyring, V., Mishra, V., Griffith, G. P., Chen, L., Keenan, T., Turetsky, M. R., et al. (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

Forster, P., Storelvmo, T., Armour, K., Collins, W., Dufresne, J.-L., Frame, D., et al. (2021). The earth's energy budget, climate feedbacks, and climate sensitivity [Book Section]. In V. Masson-Delmotte et al. (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., et al. (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. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. (Nonlinear Methods and Data Mining). https://doi.org/10.1016/S0167-9473(01)00065-2

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

Gettelman, A., Salby, M. L., & Sassi, F. (2002). Distribution and influence of convection in the tropical tropopause region. *Journal of Geophysical Research*, *107*(D10). ACL 6-1-ACL 6-12. https://doi.org/10.1029/2001JD001048

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*(3), 459. https://doi.org/10.3390/rs14030459

Giorgetta, M. A., Brokopf, R., Crueger, T., Esch, M., Fiedler, S., Helmert, J., et al. (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

Grundner, A., Beucler, T., Gentine, P., & Eyring, V. (2023). Data-driven equation discovery of a cloud cover parameterization. *arXiv e-prints*. arXiv:2304.08063. https://doi.org/10.48550/arXiv.2304.08063

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

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

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*.

Heuer, H. (2024). EyringMLClimateGroup/heuer23_ml_convection_parameterization: Version 1.0 [Software]. *Zenodo*. https://doi.org/10.5281/zenodo.12773936

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. https://doi.org/10.1080/00401706.1970.10488634

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

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. https://doi.org/10.1029/2023JD039202

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

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), A183. https://doi.org/10.17815/jlsrf-7-183

Kalthoff, N., Adler, B., Barthlott, C., Corsmeier, U., Mobbs, S., Crewell, S., et al. (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

Kim, Y., Eckermann, S. D., & Chun, H. (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*, *79*(4), 973–987. https://doi.org/10.1175/jas-d-21-0197.1

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. https://doi.org/10.1038/s41561-017-0005-4

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*, 1–13. https://doi.org/10.1155/2013/485913

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

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). https://doi.org/10.1029/2004gl021024

Kwa, A., Clark, S. K., Henn, B., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., et al. (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

LeMone, M. A. (1983). Momentum transport by a line of cumulonimbus. *Journal of the Atmospheric Sciences*, *40*(7), 1815–1834. https://doi.org/10.1175/1520-0469(1983)040⟨1815:MTBALO⟩2.0.CO;2

Li, L., Jamieson, K., Rostamizadeh, A., Gonina, E., Hardt, M., Recht, B., & Talwalkar, A. (2018). A system for massively parallel hyperparameter tuning. *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

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* (Vol. 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., & Wetherald, R. T. (1967). Thermal equilibrium of the atmosphere with a given distribution of relative humidity. *Journal of the Atmospheric Sciences*, *24*(3), 241–259. https://doi.org/10.1175/1520-0469(1967)024⟨0241:TEOTAW⟩2.0.CO;2

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

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

Mooers, G., Pritchard, M., Beucler, T., Srivastava, P., Mangipudi, H., Peng, L., et al. (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

Mooers, G., Tuyls, J., Mandt, S., Pritchard, M., & Beucler, T. G. (2021). Generative modeling of atmospheric convection. In *Proceedings of the 10th international conference on climate informatics* (pp. 98–105). Association for Computing Machinery. https://doi.org/10.1145/3429309.3429324

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

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

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. https://doi.org/10.1029/2018MS001351

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (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., et al. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Phillips, N. A. (1956). The general circulation of the atmosphere: A numerical experiment. *Quarterly Journal of the Royal Meteorological Society*, *82*(352), 123–164. https://doi.org/10.1002/qj.49708235202

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

Ramadhan, A., Marshall, J., Souza, A., Lee, X. K., Piterbarg, U., Hillier, A., et al. (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. https://doi.org/10.1175/bams-84-11-1547

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

Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russell, T., et al. (2023). Weatherbench 2: A benchmark for the next generation of data-driven global weather models. *arXiv*.

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. https://doi.org/10.1073/pnas.1810286115

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. https://doi.org/10.1029/2020ms002405

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

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 [Conference Proceedings]. In *International conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Rotunno, R., Klemp, J. B., & Weisman, M. L. (1988). A theory for strong, long-lived squall lines. *Journal of the Atmospheric Sciences*, *45*(3), 463–485. https://doi.org/10.1175/1520-0469(1988)045⟨0463:ATFSLL⟩2.0.CO;2

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. https://doi.org/10.1007/s40641-019-00131-0

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

Schneider, T., Teixeira, J., Bretherton, C. 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

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

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

Stephens, G. L., L'Ecuyer, T., Forbes, R., Gettelmen, A., Golaz, J.-C., Bodas-Salcedo, A., et al. (2010). Dreary state of precipitation in global models. *Journal of Geophysical Research*, *115*(D24). https://doi.org/10.1029/2010JD014532

Stevens, B., Acquistapace, C., Hansen, A., Heinze, R., Klinger, C., Klocke, D., 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. https://doi.org/10.2151/jmsj.2020-021

Stevens, B., Ament, F., Bony, S., Crewell, S., Ewald, F., Gross, S., et al. (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. https://doi.org/10.1175/bams-d-18-0198.1

Stevens, B., Giorgetta, M., Esch, M., Mauritsen, T., Crueger, T., Rast, S., et al. (2013). Atmospheric component of the MPI-M earth system model: ECHAM6. *Journal of Advances in Modeling Earth Systems*, *5*(2), 146–172. https://doi.org/10.1002/jame.20015

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., et al. (2019b). DYAMOND: The dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1), 1–17. https://doi.org/10.1186/s40645-019-0304-z

Stull, R. B. (1988). An introduction to boundary layer meteorology. https://doi.org/10.1007/978-94-009-3027-8

Tibshirani, R. (2018). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tiedtke, M. (1989). A comprehensive mass flux scheme for cumulus parameterization in large-scale models. *Monthly Weather Review*, *117*(8), 1779–1800. https://doi.org/10.1175/1520-0493(1989)117<1779:acmfsf>2.0.co;2

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

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). 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

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., et al. (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

Yu, F., & Koltun, V. (2015). Multi-scale context aggregation by dilated convolutions. *arXiv:1511.07122*. https://doi.org/10.48550/arXiv.1511.07122

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

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., 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. https://doi.org/10.1029/2020gl091363

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

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.

References From the Supporting Information

Alsabti, K., Ranka, S., & Singh, V. (1998). CLOUDS: A decision tree classifier for large datasets. In *Proceedings of the fourth international conference on knowledge discovery and data mining* (pp. 2–8). AAAI Press.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.