JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2021MS002959

Special Section:
Machine learning application to Earth system modeling

Deep Learning Based Cloud Cover Parameterization for ICON

Arthur Grundner[1,2], Tom Beucler[3], Pierre Gentine[2], Fernando Iglesias-Suarez[1], Marco A. Giorgetta[4], and Veronika Eyring[1,5]

[1]Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR), Institut für Physik der Atmosphäre, Oberpfaffenhofen, Germany, [2]Center for Learning the Earth with Artificial Intelligence and Physics (LEAP), Columbia University, New York, NY, USA, [3]Institute of Earth Surface Dynamics, University of Lausanne, Lausanne, Switzerland, [4]Max Planck Institute for Meteorology, Hamburg, Germany, [5]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

Abstract A promising approach to improve cloud parameterizations within climate models and thus climate projections is to use deep learning in combination with training data from storm-resolving model (SRM) simulations. The ICOsahedral Non-hydrostatic (ICON) modeling framework permits simulations ranging from numerical weather prediction to climate projections, making it an ideal target to develop neural network (NN) based parameterizations for sub-grid scale processes. Within the ICON framework, we train NN based cloud cover parameterizations with coarse-grained data based on realistic regional and global ICON SRM simulations. We set up three different types of NNs that differ in the degree of vertical locality they assume for diagnosing cloud cover from coarse-grained atmospheric state variables. The NNs accurately estimate sub-grid scale cloud cover from coarse-grained data that has similar geographical characteristics as their training data. Additionally, globally trained NNs can reproduce sub-grid scale cloud cover of the regional SRM simulation. Using the game-theory based interpretability library SHapley Additive exPlanations, we identify an overemphasis on specific humidity and cloud ice as the reason why our column-based NN cannot perfectly generalize from the global to the regional coarse-grained SRM data. The interpretability tool also helps visualize similarities and differences in feature importance between regionally and globally trained column-based NNs, and reveals a local relationship between their cloud cover predictions and the thermodynamic environment. Our results show the potential of deep learning to derive accurate yet interpretable cloud cover parameterizations from global SRMs, and suggest that neighborhood-based models may be a good compromise between accuracy and generalizability.

Plain Language Summary Climate models, such as the ICOsahedral Non-hydrostatic climate model, operate on low-resolution grids, making it computationally feasible to use them for climate projections. However, physical processes –especially those associated with clouds– that happen on a sub-grid scale (inside a grid box) cannot be resolved, yet they are critical for the climate. In this study, we train neural networks that return the cloudy fraction of a grid box knowing only low-resolution grid-box averaged variables (such as temperature, pressure, etc.) as the climate model sees them. We find that the neural networks can reproduce the sub-grid scale cloud fraction on data sets similar to the one they were trained on. The networks trained on global data also prove to be applicable on regional data coming from a model simulation with an entirely different setup. Since neural networks are often described as black boxes that are therefore difficult to trust, we peek inside the black box to reveal what input features the neural networks have learned to focus on and in what respect the networks differ. Overall, the neural networks prove to be accurate methods of reproducing sub-grid scale cloudiness and could improve climate model projections when implemented in a climate model.

Key Points:

- Neural networks can accurately learn sub-grid scale cloud cover from realistic regional and global storm-resolving simulations
- Three neural network types account for different degrees of vertical locality and differentiate between cloud volume and cloud area fraction
- Using a game theory based library we find that the neural networks tend to learn local mappings and are able to explain model errors

Supporting Information:

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

Correspondence to:

A. Grundner,
Arthur.Grundner@dlr.de

Citation:

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*, e2021MS002959. https://doi.org/10.1029/2021MS002959

Received 20 DEC 2021
Accepted 5 DEC 2022

Author Contributions:

Conceptualization: Arthur Grundner, Pierre Gentine, Fernando Iglesias-Suarez, Veronika Eyring
Data curation: Arthur Grundner, Fernando Iglesias-Suarez, Marco A. Giorgetta

1. Introduction

Clouds play a key role in the climate system. They regulate the hydrologic cycle and have a substantial influence on Earth's radiative budget (Allen & Ingram, 2002). Yet, in climate models with horizontal resolutions commonly on the order of 100 km, clouds are sub-grid scale phenomena, that is, they cannot be directly resolved but need to be "parameterized." These parameterizations are a major cause of uncertainties in climate model projections (e.g., Randall et al., 2003; Schneider et al., 2017) and effective climate sensitivity (Meehl et al., 2020; Schlund et al., 2020).

The long-standing deficiencies in cloud parameterizations have motivated the development of high-resolution global cloud-resolving climate models (Klocke et al., 2017; Stevens, Satoh, et al., 2019) with the ultimate goal of

Formal analysis: Arthur Grundner, Tom Beucler, Pierre Gentine
Funding acquisition: Veronika Eyring
Investigation: Arthur Grundner, Tom Beucler, Pierre Gentine
Methodology: Arthur Grundner, Tom Beucler, Pierre Gentine, Fernando Iglesias-Suarez, Marco A. Giorgetta, Veronika Eyring
Project Administration: Pierre Gentine, Veronika Eyring
Resources: Fernando Iglesias-Suarez, Marco A. Giorgetta, Veronika Eyring
Software: Arthur Grundner
Supervision: Tom Beucler, Pierre Gentine, Fernando Iglesias-Suarez, Veronika Eyring
Validation: Arthur Grundner, Tom Beucler
Visualization: Arthur Grundner
Writing – original draft: Arthur Grundner, Tom Beucler
Writing – review & editing: Arthur Grundner, Tom Beucler, Pierre Gentine, Fernando Iglesias-Suarez, Marco A. Giorgetta, Veronika Eyring

explicitly resolving clouds and convection. Yet, these simulations are extremely computationally demanding and cannot be run on climate timescales for multiple decades or for ensembles. Deep learning for the parameterization of sub-grid scale processes has been identified as a promising approach to improve parameterizations in climate models and to reduce uncertainties in climate projections (Eyring et al., 2021; Gentine et al., 2021).

In the atmospheric component of the state-of-the-art ICOsahedral Non-hydrostatic (ICON) climate model (ICON-A), clouds result from an interplay of different parameterization schemes (Giorgetta et al., 2018). In it, the cloud cover scheme takes an integral role. Its cloud cover directly influences the tendencies and hence statistics of cloud liquid water, cloud ice, and water vapor through the microphysics scheme (Lohmann & Roeckner, 1996; Pincus & Stevens, 2013), and the energy balance through the radiation scheme.

Our goal is to develop a machine learning based parameterization that can replace ICON's semi-empirical cloud cover scheme. We cover the background of these two fields in Section 1.1 by reviewing (a) the existing cloud cover scheme in ICON (in Section 1.1.1) and (b) the machine learning based parameterizations (in Section 1.1.2) before defining the scope of our study in Section 1.2.

1.1. Background

1.1.1. Existing Cloud Cover Scheme in ICON

Cloud cover is estimated as a diagnostics in ICON, which is based on the local amount of relative humidity (RH), and a semi-empirical relationship devised by Sundqvist et al. (1989) and further adapted by Xu and Krueger (1991) (see Lohmann and Roeckner (1996)) and Mauritsen et al. (2019). In this scheme, cloud cover exists whenever RH exceeds a specified lower bound (the *critical RH threshold*), which depends solely on atmospheric and surface pressure.

RH-based cloud cover schemes have some notable drawbacks. First of all, knowing RH does not fully determine cloud cover. For instance Walcek (1994) had shown, that with an RH of 80% and a pressure between 800 and 730 hPa, the probability of observing any amount of cloud cover can be nearly uniform. In addition, no clear critical RH threshold seems to exist. Furthermore, even though they influence cloud characteristics, RH-based schemes do not directly differentiate between local dynamical conditions (e.g., whether the grid column undergoes deep convection; A. Tompkins, 2005). The ICON-A cloud cover scheme also does not account for vertical sub-grid scale cloud cover variability. An exception to this is the recent adaptation to artificially increase RH in regions below subsidence inversions to incorporate thin marine stratocumuli (Mauritsen et al., 2019).

Finally, most cloud schemes are based on local thermodynamic variables, yet rapid advection (e.g., updrafts) could lead to non-locality in the relationship. Overall, the formation and dissipation of clouds is still poorly understood (Stensrud, 2009). Therefore, physics-based cloud parameterizations have to build on incomplete knowledge and are prone to inaccuracies. They usually also contain tuning parameters. In the ICON-A cloud cover scheme these are the RH for 100% cloud cover, the asymptotic critical RH in the upper troposphere, the critical RH at the surface, and the shape factor. These parameters have to be adjusted following the primary goal of a well balanced top-of-the-atmosphere energy budget (Giorgetta et al., 2018).

1.1.2. Machine Learning Based Parameterizations

The field of machine learning based parameterizations is growing and can loosely be classified into two groups: The first group consists of studies about machine learning based parameterizations that emulate and speed up existing parameterizations. In Beucler et al. (2020), Gentine et al. (2018), Han et al. (2020), Mooers et al. (2020), and Wang et al. (2022) these existing parameterizations were superparameterizations, that is, embedded two-dimensional cloud-resolving models (Khairoutdinov et al., 2005). For instance, in a pioneering study by Rasp et al. (2018), a neural network (NN) was successfully trained to estimate sub-grid scale convective effects by learning from the output of the superparameterized Community Atmosphere Model in an idealized aquaplanet setting. Other notable members of this group, that focused on emulating more traditional parameterizations, are Chevallier et al. (2000), Chantry et al. (2021), Gettelman et al. (2021), Krasnopolsky et al. (2005), and Seifert and Rasp (2020). The second group consists of studies about machine learning based parameterizations that learn from three-dimensional, high-resolution data. In most of those studies, the high-resolution data was coarse-grained to the low-resolution grid of the climate model. The first proof of concept was established by Krasnopolsky et al. (2013) who trained a very small NN on coarse-grained regional data. Later, Brenowitz and

Bretherton (2018), Brenowitz and Bretherton (2019), Brenowitz et al. (2020), Yuval and O'Gorman (2020), and Yuval et al. (2021) adapted this approach. However, in contrast to our study, they worked with idealized aqua-planet simulations and coarse-graining limited to the horizontal dimension.

While some of these studies were conducted in a purely "offline" fashion, that is, decoupled from the dynamics of the climate model, Brenowitz and Bretherton (2019), Brenowitz et al. (2020), Chantry et al. (2021), Gettelman et al. (2021), Krasnopolsky et al. (2005), Ott et al. (2020), Rasp et al. (2018), Wang et al. (2022), Yuval and O'Gorman (2020), and Yuval et al. (2021) also achieved stable online simulations in specific setups.

Recent research has suggested that emulating sub-grid scale physics on a process-by-process level may lead to more stable machine learning powered climate simulations (Yuval et al., 2021). It may also facilitate interpretability and targeted studies of the interaction between large-scale (thermo)dynamics and cloudiness.

1.2. Machine Learning Based Cloud Cover Parameterization

In the context of these new advances, our study is the first machine learning based approach specifically focused on the parameterization of cloud cover.

Our novel approach to a cloud cover parameterization is based on the idea of training a supervised deep learning scheme to estimate cloud cover from the thermodynamical state, using coarse-grained high-resolution data. We allow for vertical sub-grid scale cloud cover variability by learning the fraction of a grid volume that is cloudy ("cloud volume fraction"; Brooks et al., 2005). Cloud volume fraction is the preferable measure of cloud cover, for instance in ICON's microphysics scheme where in-cloud condensation and evaporation rates are multiplied by the volume fraction of the grid box that is cloudy (Lohmann & Roeckner, 1996). In Section 4.2, we also introduce NNs that predict the horizontally projected amount of cloudiness inside a grid cell ("cloud area fraction"). The reason is that we still require cloud area fraction as a parameter for the (ICON's two-stream) radiation scheme (Pincus & Stevens, 2013) to evaluate whether radiation penetrates through a cloud or not.

The ICON modeling framework is used in realistic conditions on a variety of timescales and resolutions (Zängl et al., 2015). It thus allows us to work with data from high-resolution ICON simulations to train machine learning based parameterizations fit for the low-resolution ICON climate model. Observations, on the other hand, are temporally and spatially sparse and would thus constitute less adequate training data (Rasp et al., 2018). The basis of our training data form new storm-resolving ICON simulations from the Next Generation Remote Sensing for Validation Studies (NARVAL) flight campaigns (Stevens, Ament, et al., 2019) and the Quasi-Biennial Oscillation in a Changing Climate (QUBICC) project (Giorgetta et al., 2022), with horizontal resolutions of 2.5 and 5 km respectively. At these resolutions one can generally consider deep convection to be resolved (Vergara-Temprado et al., 2020), and therefore these simulations forego the use of convective parameterizations. Hohenegger et al. (2020) systematically compared 27 different statistics in ICON simulations with resolutions ranging from 2.5 to 80 km. They concluded that simulations with explicit convection at resolutions of 5 km or finer may indeed be used to simulate the climate. Stevens et al. (2020) have shown that the NARVAL simulations can more accurately represent clouds and precipitation than simulations with an active convective parameterization.

We train NNs on coarse-grained data from these high-resolution simulations. Here, two commonly used ICON-A grids (with horizontal resolutions of 80 and 160 km) are the target grids we coarse-grain to. ICON uses an icosahedral grid in the horizontal and a terrain-following height grid in the vertical. On these grids, more sophisticated and partly new methods of coarse-graining are required than on simpler regular grid types. As our machine learning algorithm we choose NNs, which are able to incorporate this wealth of data to –in principle– approximate any type of nonlinear function (Gentine et al., 2018; Hornik, 1991). While being generally fast at inference time, NNs also have computational advantages over alternative machine learning based approaches such as random forests (Yuval et al., 2021). Hence, an NN-powered parameterization of cloud cover could accelerate and improve the representation of cloud-scale processes (from radiative feedbacks to precipitation statistics).

In this study, we focus on developing an *offline* (i.e., without coupling to the dynamical core), ML-based cloud cover parameterization for ICON. While offline skill does not always guarantee *online* performance once the NN is coupled back to the dynamical core (Gagne et al., 2020), Ott et al. (2020) showed that offline skill generally correlated with the stability (although not necessarily the accuracy) of online simulations. Several time-consuming tasks are required to achieve operational online skill, such as ensuring excellent extrapolation skills to different

distributions of state variables for stable simulations (across climate-regimes). Then, a re-calibration of the coarse-resolution climate model against the observed state of the atmosphere (top-of-the-atmosphere radiative fluxes, global mean surface temperature, clouds, precipitation, wind fields, etc., Giorgetta et al. (2018)) is most likely necessary, for example, since there are too few (low-level) clouds in the ICON model, and other tunable parameters are currently calibrated to compensate for that fact (Crueger et al., 2018). After all, the performance of a (ML-based) cloud cover parameterization always depends on the accuracy of its inputs, which in turn are affected by other parameterizations in an online setting (e.g., cloud ice/water mixing ratios and specific humidity are modified by ICON's microphysics scheme). Finally, these tasks depend on the correct implementation of the Python-trained NNs into climate model source code (typically written in Fortran). To keep this study tractable, we therefore chose to leave the online implementation for future work; taking the first necessary step of demonstrating a robust offline parameterization, we focus on five questions.

The first key question that we tackle in this study is whether we can train an NN based cloud cover parameterization that is able to emulate high-resolution cloudiness. We then ask the following subquestions: For the sake of generalizability and computational efficiency should we keep the parameterization as local as possible? Or shall we consider non-local effects for improved accuracy? Can we apply this parameterization universally or is it tied to the regions and climatic conditions over which it was trained upon? And can we extract useful physical information from the NN after it has been trained, gaining insight into the interaction between the large-scale (thermo)dynamic state and convective-scale cloudiness?

We first introduce the training data (Section 2.1) and the NNs (Section 3), before evaluating regionally (Section 4.1) and globally (Section 4.2) trained networks in their training regime, studying their generalization capability (Section 4.3) and interpreting their predictions (Sections 4.4 and 4.5).

2. Data

2.1. ICON High-Resolution Simulations

The training data consists of coarse-grained data from two distinct ICON storm-resolving model (SRM) simulations. Both simulations provide hourly model output.

The first simulation is a limited-area ICON simulation over the tropical Atlantic and parts of South America and Africa (10°S–20°N, 68°W–15°E). The simulation ran for a bit over 2 months (December 2013 and August 2016) in conjunction with the NARVAL (NARVALI and NARVALII) expeditions (Klocke et al., 2017; Stevens, Ament, et al., 2019). The model was initialized at 0 UTC every day and ran for 36 hr. We use the output from the model runs with a native resolution of ≈2.5 km. NARVAL data also exists with a higher resolution of ≈1.2 km, but it covers a significantly smaller domain (in 4°S–18°N, 64°W–42°W). The native vertical grid extends up to 30 km on 75 vertical layers.

The second simulation is a global ICON simulation that ran as part of the QUBICC project. Currently there is a set of hindcast simulations available of which we chose three to work with (hc2, hc3, hc4). Each simulation covers 1 month (November 2004, April 2005, and November 2005). While the horizontal resolution (≈5 km) is lower than in NARVAL, the vertical grid extends higher (up to 83 km) on a finer grid (191 layers).

The two simulations used different collections of parameterization schemes. While the NARVAL simulations were set up to run with ICON's numerical weather prediction physics package (Prill et al., 2019), the QUBICC simulations used the so-called Sapphire physics, developed for SRM simulations and based on ICON's ECHAM physics package as described in Giorgetta et al. (2022). An overview of the specifically chosen parameterization schemes can be found in Table S1 in Supporting Information S1. By virtue of their high resolution, both simulations dispensed with parameterizations for convection and orographic/non-orographic gravity wave drag. For microphysics they used the same single-moment scheme, which predicts rain, snow, and graupel in addition to water vapor, liquid water, and ice (Doms et al., 2011; Seifert, 2008). Different schemes were used for the vertical diffusion by turbulent fluxes (NARVAL: Raschendorfer (2001), QUBICC: Mauritsen et al. (2007)), for the radiative transfer (NARVAL: Barker et al. (2003); Mlawer et al. (1997), QUBICC: Pincus et al. (2019)), and the land component (NARVAL: Schrodin and Heise (2001); Schulz et al. (2015), QUBICC: Raddatz et al. (2007)). The simulations also differed in their cloud cover schemes. The QUBICC simulation assumed to resolve cloud-scale motions, diagnosing a fully cloudy grid cell whenever the cloud condensate ratio exceeds a small threshold and

a cloud-free grid cell otherwise. The cloud cover scheme used in NARVAL alternatively produces fractional cloud cover with a diagnostic statistical scheme that combines information from convection, turbulence, and microphysics.

In ICON terminology, the NARVAL simulations ran on an R2B10 and the QUBICC simulations on an R2B9 (horizontal) grid. Generally speaking, an RnBk grid is a refined spherical icosahedron. The refinement is performed by (a) dividing its triangle edges into n parts, creating new triangles by connecting the new edge points and by (b) completing k subsequent edge bisections while once more connecting the new edge points after each bisection (Giorgetta et al., 2018). In between these refinement steps, the position of each vertex is slightly modified using a method called spring dynamics, which improves the numerical stability of differential operators (Tomita et al., 2001; Zängl et al., 2015).

A key limitation of the data lies in a temporal mismatch between some model output variables from one common time step. This is caused by the sequential processing of some parameterization schemes in the ICON model (Giorgetta et al., 2018). For instance, the cloud cover scheme diagnoses cloud cover before the microphysics scheme alters the cloud condensate mixing ratio, which has led to ≈7% of the cloudy grid cells in our data to be condensate-free. However, this mismatch should not exceed the fast physics time step in the model, which was set to 40 s in the QUBICC and to 24 s in the NARVAL simulations. Another limitation of our QUBICC data is that the mixing length in the vertical diffusion scheme was mistakenly set to 1000 m instead of 150 m, causing unrealistically strong vertical diffusion in some situations.

2.2. Coarse-Graining

We now use both NARVAL and QUBICC data to derive training data for our machine learning based cloud cover parameterization.

This requires coarse-graining the data horizontally and vertically to the low-resolution ICON-A grid since we cannot a priori assume that the same (cloud cover) parameterization will work across a very wide range of spatial resolutions. Our goal is to mimic typical inputs of our cloud cover parameterization, which are the large-scale state variables of ICON-A. We design our coarse-graining methodology to best estimate grid-scale mean values, which we use as proxies for the large-scale state variables. Figure 1 shows an example of horizontal and vertical coarse-graining of cloud cover snapshots from the QUBICC and the NARVAL data set.

We coarse-grain the simulation variables from R2B9 and R2B10 grids to the default R2B4 grid of Giorgetta et al. (2018) with a resolution of ≈160 km. To demonstrate the robustness of our machine learning algorithms across typical ICON-A resolutions, we additionally coarse-grain to the low-resolution R2B5 grid used in Hohenegger et al. (2020) with a resolution of ≈80 km. Afterward, we vertically coarse-grain the data to 27 terrain-following sigma height layers, up to a height of 21 km because no clouds were found above that height. The technical aspects of our coarse-graining methodology can be found in Appendix A. We now turn toward the specifics of the NNs.

3. Neural Networks

3.1. Setup

We set up three general types of NNs of increasing representation power. Each NN follows its own assumption as to how (vertically) local the problem of diagnosing cloud cover is. Choosing three different NN architectures allows us to design a vertically local (cell-based), a non-local (column-based), and an intermediate (neighborhood-based) model type.

The (*grid-*)*cell-based model* only takes data from the same grid cell level and potentially some surface variables into account. In that sense, the traditional cloud cover parameterization in ICON-A, being a function of local RH, pressure, and surface pressure, is similarly a cell-based parameterization (with the exception of including the lapse rate in certain situations). Such a local model is very versatile and can be implemented in models with varying vertical grids.

The *neighborhood-based model* has variables as its input that come from the same grid cell and from the ones above and below, and also includes some surface variables. The atmospheric and dynamical conditions in the

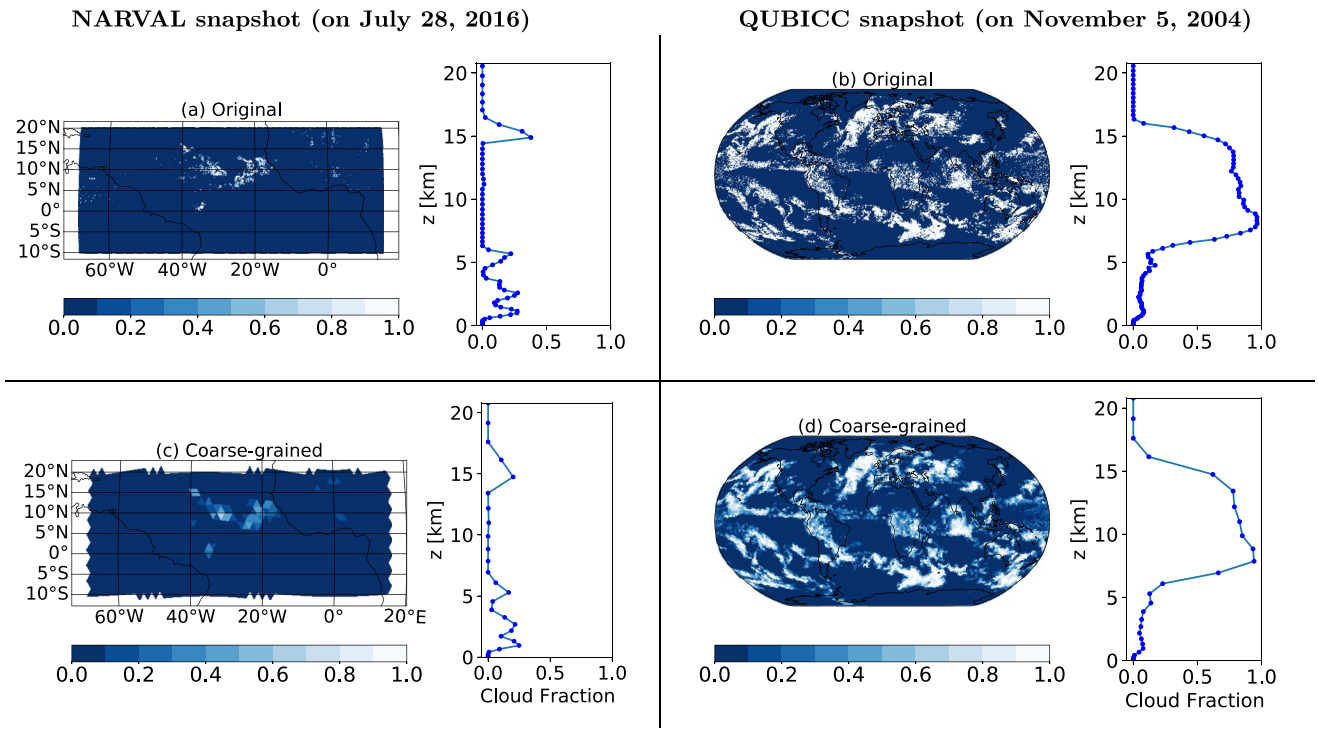


Figure 1. Illustration of coarse-graining using the example of cloud fraction. Here we show distinct snapshots of the horizontal fields (on a single layer) and vertical profiles (from a single column) from the high-resolution NARVAL and Quasi-Biennial Oscillation in a Changing Climate (QUBICC) simulations (top row) and the corresponding coarse-grained horizontal fields and vertical profiles (bottom row). We coarse-grain the NARVAL/QUBICC data sets horizontally from 2.5 km/5 km to 160 km/80 km and vertically from 66/87 to 27 layers up to a height of 21 km. Final coarse-grained grid boxes constitute the training data for the machine learning models.

close spatial neighborhood of the grid cell most likely have a significant influence on cloudiness as well. A grid column undergoing deep convection for instance is very likely to have different cloud characteristics than a grid cell in a frontal stratus cloud (A. Tompkins, 2005). Furthermore, strong subsidence inversions that lead to thin stratocumuli cannot be detected by looking at the same grid cell only. As an example, this dependence of cloudiness on the surroundings has been actualized in A. M. Tompkins (2002). In their study, the sub-grid distribution of total water is described as a function of horizontal and vertical turbulent fluctuations, effects of convective detrainment and microphysical processes.

The *column-based model* operates on the entire grid column at once, and therefore has as many output nodes as there are vertical layers. In a column-based approach we do not have to make any a priori assumptions as to how many grid cells from above and below a given grid cell should be taken into account. Furthermore, surface variables are naturally included in the set of predictors. Coefficients of a multiple linear model fitted to the data suggest that the parameterization of cloud cover is a non-local problem, further motivating the use of a column-based model (see Figure S1 in Supporting Information S1). The input-output architecture of these three NN types is illustrated in Figure S2 in Supporting Information S1.

We specify three NNs to be trained on the (coarse-grained) NARVAL R2B4 data and three networks to be trained with (coarse-grained) QUBICC R2B5 data. Using data that is coarse-grained to different resolutions allows us to demonstrate the applicability of the approach across resolutions. The primary goal of the NNs trained on NARVAL R2B4 data is to show the ability to reproduce SRM cloud cover from coarse-grained variables, whereas for the globally-trained QUBICC R2B5 NNs it is a versatile applicability and more grid-independence. In this context, the largest differences between the R2B4- and R2B5 models exist in the specification of the neighborhood-based models:

The set of predictors for the neighborhood-based R2B5 model contains data from the current grid cell and its immediate neighbors (above and below it). On the layer closest to the surface this requires padding to create data from "below." The vertical thickness of grid cells decreases with decreasing altitude. Therefore, we assume a

Table 1
Overview of the NNs and Their Input Features

	NN Type	Land	Lake	Cor.	T_s	z_g	q_v	q_c	q_i	T	p	ρ	u	v	clc_{t-1}
N1	Cell-based	✓	–	–	–	✓	✓	–	✓	✓	✓	–	–	–	–
N2	Column-based	–	✓	–	–	✓	✓	✓	✓	✓	✓	✓	–	–	–
N3	Neighborhood-based	–	✓	–	–	✓	✓	✓	✓	✓	✓	✓	–	–	✓
Q1	Cell-based	✓	–	✓	–	✓	✓	✓	✓	✓	✓	–	✓	✓	–
Q2	Column-based	✓	–	–	–	✓	✓	✓	✓	✓	✓	–	–	–	–
Q3	Neighborhood-based	–	–	✓	✓	✓	✓	✓	✓	✓	✓	–	✓	✓	–

Note. Models N1–N3 are trained on NARVAL R2B4 and models Q1-Q3 on QUBICC R2B5 data. 2D variables (fraction of land/lake, Coriolis parameter and surface temperature) are listed in the first three columns. More information on the choices and meaning of the features can be found in the SI.

layer separation of 0 for this artificial layer below, allowing us to fill it with values from the layer closest to the surface.

The neighborhood-based R2B4 model considers two grid cells above and two below. We did not extend the padding to create another artificial layer, but trained a unique network per vertical layer. This allows for maximum flexibility, discarding input features that are non-existent or constant on a layer-wise basis. Additionally, the R2B4 model has cloud cover from the previous model output time step (1 hr) in its set of predictors.

An overview of the NNs and their input parameters can be found in Table 1. The input parameters were mostly motivated by the existing cloud cover parameterizations in ICON-A and the Tompkins Scheme (A. M. Tompkins, 2002). All NNs have a common core set of input features. Choosing varying additional features allows us to study their influence. However, we found that none of these additional features have a crucial impact on a model's performance. We generally chose as few input parameters as possible to avoid extrapolation situations outside of the training set as much as possible. By doing so, we hope to maximize the generalization capability of the NNs.

3.2. Training

In this section we explain the training methodology and the corresponding tuning of the models' and the optimizer's hyperparameters (e.g., model depth, activation functions, initial learning rate). These hyperparameters have a large impact on the potential quality of the NN. The importance of hyperparameter tuning for NN parameterizations was pointed out in Ott et al. (2020), and Yuval et al. (2021) proposed its particular need in a real-geography setting.

The choice of hyperparameters for an NN depends on the amount and nature of the training data which in turn depends strongly on the setup. A column-based model in an R2B4 setup trained on NARVAL data can be trained with no more than $1.7 \cdot 10^6$ data samples, using all available data. In contrast, a cell-based model in an R2B5 setup trained on QUBICC data can learn from maximally $4.6 \cdot 10^9$ data samples. Table S2 in Supporting Information S1 shows the amount of available training data for every setup. Mainly the coarse-grained QUBICC data had to be (further) preprocessed to (a) reduce the size of the data set, (b) scale the cloud cover target to a common range, (c) normalize the training data, and (d) combat the class imbalance of having a relatively large number of cloud-free grid cells in the training data. Steps c) and (d) were also necessary for the coarse-grained NARVAL data. The more balanced ratio between cloudy and cloud-free grid cells (which encourages the neural networks to correctly recognize cloudy cells) for (d) was achieved by randomly sub-sampling from the cloud-free grid cells. More details on the preprocessing can be found in the Supporting Information S1.

To train the NARVAL R2B4 networks we follow conventional machine learning practices and split the (coarse-grained and preprocessed) R2B4 data into randomly sampled disjoint training, validation and test sets (78%/8%/20% of the data). By randomly splitting the data, we ensure (with a high probability) that the model will see every weather event present in the training data, with the caveat that strongly correlated samples could be distributed across the three subsets. In contrast, for the QUBICC R2B5 models, we focus on universal applicability. We therefore use a temporally coherent three-fold cross-validation split (illustrated in Figure S3 in Supporting

Table 2
Hyperparameters of the NNs and the Optimizer

	Models N1–N3 and Q2	Models Q1 and Q3
Hidden layers	2	3
Units per hidden layer	256	64
Activation fct. for each layer	ReLU → ReLU → linear	tanh → leaky ReLU ($\alpha = 0.2$) → tanh → linear
L1, L2 reg. coef. for each layer	None	L1: $4.7 \cdot 10^{-3}$, L2: $8.7 \cdot 10^{-3}$
Batch Normalization	None	After the second hidden layer
Optimizer	N1–N3: Nadam, Q2: Adam	Q1: Adam, Q3: Adadelta
↪ Initial learning rate	10^{-3}	$4.3 \cdot 10^{-4}$
↪ Batch size	N1–N3: 32, Q2: 128	1,028
↪ Maximal number of epochs	N1–N3: 70, Q2: 40	Q1: 30, Q3: 50

Information S1). Every fold covers roughly 15 days to make generalization to the validation folds more challenging. We choose 15 days to stay above weather-timescales (so that for instance the same frontal system does not appear in the training and validation folds) and to mitigate temporal auto-correlation between training and validation samples. The validation folds of each split are equally difficult to generalize to, since a part of every month is always included in the training folds. The three-fold split itself lowers the risk of coincidentally working with one validation set that is very conducive to the NN.

After tuning the hyperparameters using the Bayesian optimization algorithm within the SHERPA package (Hertel et al., 2020) we found that a common architecture was optimal for the models N1-N3 and Q2. We list the space of hyperparameters we explored in the SI. For models Q1 and Q3 we had more training data. To counteract the increase in training time, we increased the batch size to keep a similar amount of iterations per training epoch. After renewed hyperparameter tuning we found a different architecture for models Q1 and Q3. The final choice of hyperparameters for the NNs is shown in Table 2. The relatively small size of the NNs (which is comparable to those of Brenowitz and Bretherton (2019)) helps against overfitting the training data and allows for faster training of the networks. By performing systematic optimization of hyperparameters we also found that these networks are already able to capture the functional complexity of the problem.

4. Results

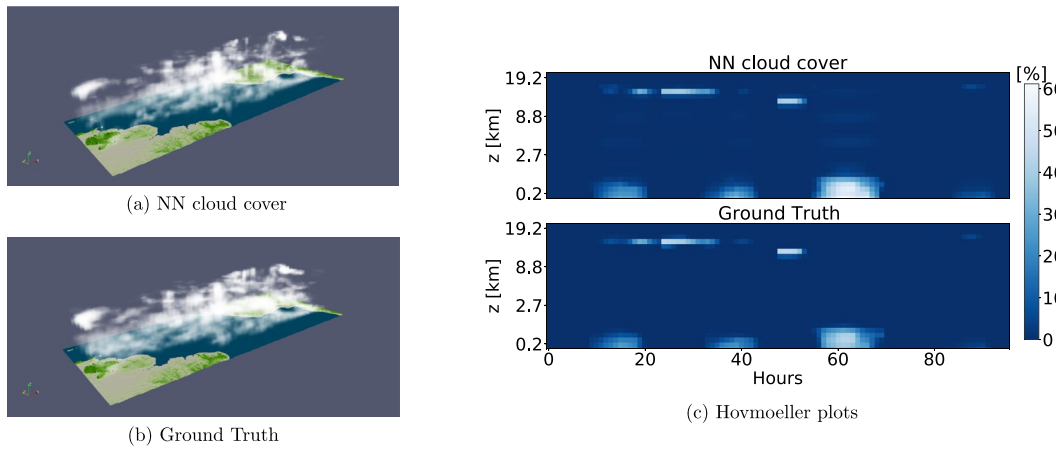
4.1. Regional Setting (NARVAL)

In this section we show the results of the NNs trained and evaluated on the coarse-grained and preprocessed NARVAL R2B4 data (see Supporting Information S1 for more details on the preprocessing). For these regionally-trained NNs we define cloud cover as a cloud volume fraction.

The snapshots and Hovmoeller plots of Figure 2 provide visual evidence concerning the capability of the (here column-based) NN to reproduce NARVAL cloud scenes. The ground truth consists of the coarse-grained NARVAL cloud cover fields, which the NN reconstructs while only having access to the set of coarse-grained input features. In the Hovmoeller plots we trace the temporal evolution of cloudiness throughout 4 days in a randomly chosen grid column of the NARVAL region. Given the large-scale data from the grid column, the NN is able to deduce the presence of all six distinct lower- and upper-level clouds.

The models' mean-squared errors (MSEs) (shown in Table 3) represent the absolute average squared mismatch per grid cell in percent between the predicted and the true cloud cover. For a given data set $X = \{X_i\}_{i=1}^{N}$, where for each of the samples X_i the true cloud cover is given by Y_i and the predicted cloud cover by $\hat{Y}_i$, the MSE is defined by

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \hat{Y}_i\right)^2. \quad (1)$$



(a) NN cloud cover

(b) Ground Truth

(c) Hovmoeller plots

Figure 2. The column-based neural network (NN) trained and evaluated on the coarse-grained NARVAL R2B4 data. Panels (a and b) show cloud cover snapshots with (a) displaying the cloud scene as it is estimated by the NN and (b) the reference cloud scene from the coarse-grained NARVAL data. Note that some columns over land could not be vertically interpolated due to overlapping topography and are therefore missing in (a). The upper plot of panel (c) shows the cloud cover predictions of 1 August – 4 August 2016 by the NN in some arbitrary location within the NARVAL region. The plot below depicts the data's actual (coarse-grained) cloud cover. The vertical axis shows average heights of selected vertical layers.

As opposed to Figure 2, the MSEs provide more statistically tangible information. The column-based model (which has the largest number of learnable parameters) and the neighborhood-based model (which consists of a unique NN per vertical layer) have lower MSEs than the cell-based model. More trainable parameters allow for the model to adjust better to the ground truth. We also found that by adding more input features (RH, liquid water content, lapse rate and surface pressure) to the cell-based model, we can further decrease its MSE to $\approx 5\,(\%)^2$. On the flip side, every additional input feature bears the risk of impeding the versatile applicability of the model and reducing its capacity to generalize to unseen conditions. By training multiple models of the same type, we verified these MSEs to be robust (varying by $\pm 0.12\ (\%)^2$). The MSEs for the neighborhood-based model are averaged over all NNs (i.e., one per vertical layer), while the upper-most two layers are left out due to the rare presence of clouds at these altitudes.

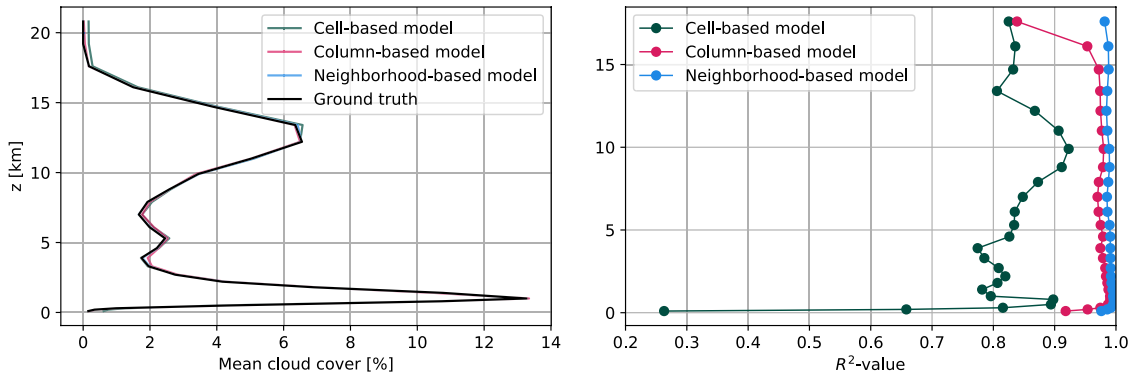
Our data is temporally and spatially correlated. As a consequence, our division into random subsets for training, validation, and testing leads to very similar MSEs on the respective subsets. And the error on the training set is only slightly smaller than on the validation and test sets.

With MSEs being below 16 $(\%)^2$, Table 3 shows that the NNs are able to diagnose cloud cover better than our baseline models (with the exception of the cell-based random forest). These baseline models are fitted to the same normalized data sets as the respective NNs. As our first baseline we evaluate a constant output model, which outputs the average cloud cover. The constant output model's MSE thus also represents the variance of cloud cover in the data. Small differences in the preprocessing of the data for each model type lead to differences in the MSEs of the zero and constant output model. The (multiple) linear model is trained on the data using the ordinary least squares method. For the random forests, we use the default implementation of the RandomForestRegressor in scikit-learn, adjusting the number and the maximum depth of the trees so that the training duration is similar to the NNs. Further adjustments of these two hyperparameters that would further increase or decrease the training durations either reach computational limits or show no decrease in validation loss. While the cell-based random forest actually achieves a lower MSE than the NN, its $\approx 10^5$ larger size (400 GB) makes it impractical to manage. When forced to have a similar storage requirement using the two hyperparameters mentioned above, its MSE (26.22) becomes larger than that of the NN.

We implemented the Sundqvist scheme as it is described in Giorgetta et al. (2018). It is a simplified version of the currently implemented (mainly

Table 3
Mean-Squared Errors (in $(\%)^2$) of NARVAL and Baseline Models Evaluated on the Coarse-Grained and Preprocessed NARVAL Data

		Type		
		Cell-based	Column-based	Neighborhood-based
Neural networks	Training set	15.16	1.64	0.84
	Validation set	15.18	1.78	1.00
	Test set	15.19	1.78	1.01
Baseline models	Constant output model	109.63	92.23	86.48
	Best linear model	81.71	18.56	4.79
	Random forest	10.40	6.15	1.73
	Sundqvist scheme	51.14	–	–



(a) Cloud cover profiles (b) Coefficients of determination (best value: 1)

Figure 3. Evaluation of the NARVAL R2B4 models on the coarse-grained and preprocessed NARVAL R2B4 data. The three cloud cover maxima of panel (a) are located roughly at 1 km, 5.3 and 12.2 km. The maximal absolute discrepancy between the averaged neural network predictions and the ground truth for a given vertical layer is less than 0.5%. In panel (b), the two upper-most layers are not shown.

cell-based) ICON-A cloud cover parameterization, because it does not include an adjustment for cloud cover in regions below subsidence inversions over the ocean (see Mauritsen et al. (2019)). We fitted the Sundqvist scheme to the data by doing a grid search over a space of tuning parameters around the values used in the ICON-A model. The grid search yielded a better set of tuning parameters than those found by implementing the scheme as a layer in TensorFlow and optimizing the tuning parameters using gradient descent. To still allow for a differentiation between grid cells over land and ocean, we found optimal sets of tuning parameters for cells that are mainly over land ($\{r_{sat}, r_{0,top}, r_{0,surf}, n\} = \{1.12, 0.3, 0.92, 0.8\}$) and for cells that are mainly over the sea ($\{r_{\text{sat}}, r_{0,\text{top}}, r_{0,\text{surf}}, n\} = \{1.07, 0.42, 0.9, 1.1\}$).

Figure 3a shows that the mean vertical profiles of cloud cover predicted by the NNs closely align with the "Ground truth" profile of coarse-grained cloud cover. The profiles feature three maxima that can be attributed to the three modes of tropical convection: shallow, congestus, and deep. Note that in contrast to Müller (2019), we do find a clear peak for deep convective clouds in the coarse-grained NARVAL and NARVALII data, which could be due to differences in how we define cloudy grid cells (using the cloud cover model output rather than a boolean based on the total cloud condensate mass mixing ratio exceeding 0.1 g/kg).

In Figure 3b we show the coefficient of determination/R^2-value profiles for the different models. For a given vertical layer l, the R^2-value is defined by

$$R_l^2 = 1 - \frac{mse_l}{\text{var}_l}. \tag{2}$$

For a given vertical layer l, mse_l is the MSE between a given model's prediction and the true cloud cover and var_l the variance of cloud cover. Clearly, (a) $R_l^2 \leq 1$, (b) $R_l^2 = 1$ implies $mse_l = 0$, and (c) if $R_l^2 \leq 0$, then a function always yielding the cloud cover mean on layer l would outperform the model in question.

We see that the neighborhood- and column-based models generally have R^2-values exceeding 0.9, or equivalently $mse_l \leq 0.1 \cdot \text{var}_l$. The somewhat lower reproduction skill for the cell-based model concurs with the MSEs found in Table 3. The models exhibit strongly negative R^2-values above 19 km and are therefore not shown in the figure, that is, on these layers a constant-output model would be more accurate than the NNs. The reason for this is that there are almost no clouds above 19 km; the variance of cloud cover is not greater than 10^{-4} $(\%)^2$. Nevertheless, the neighborhood-based model with its unique NN per vertical layer is still able to learn a reasonable mapping at 19.2 km, achieving an R^2-value of 0.93. Altogether, we found the mean cloud cover statistics to be independent of how the NNs were initialized prior to training.

Table 4
Mean-Squared Errors (in (%)^2) of the NNs Trained With a 3-Fold Cross-Validation Split on the Coarse-Grained and Preprocessed Quasi-Biennial Oscillation in a Changing Climate Data

		Type		
		Cell-based	Column-based	Neighborhood-based
Neural networks	Cloud volume fraction	32.77 (28.98)	8.14 (8.03)	25.07 (20.46)
	Cloud area fraction	87.98 (80.96)	20.07 (19.79)	52.19 (46.61)
Baseline models	Constant output model	684.51	431.28	558.28
	Best linear model	401.47	97.81	297.63
	Random forest	25.90	161.98	54.74
	Sundqvist scheme	474.12	–	–

Note. Due to computational reasons, only 1% of the data (i.e., $\approx 10^7$ samples) was used to compute the MSE of the Sundqvist scheme. We only show the MSEs of the models with the lowest loss on their respective validation folds. Here, the neighborhood-based models comprise one model per split, evaluated on all layers. In parentheses we compute the losses after bounding the model output to the [0, 100]% interval. The baseline models are trained and evaluated on coarse-grained and preprocessed QUBICC cloud volume fraction data.

4.2. Global Setting (QUBICC)

Having studied the performance of our regionally trained NNs, we now shift the focus to the NNs trained and evaluated on the coarse-grained and preprocessed global QUBICC R2B5 data set. Changing the region as well as the resolution of the training data allows us to conduct studies across these domains in Section 4.4.

Table 4 shows the performance of the cloud volume and cloud area fraction NNs on their validation folds. For each model type and each of the three cross-validation splits we trained one NN and then selected the NN that has the lowest MSE on the entire QUBICC data set. Generally, this is also the NN with the lowest loss on its validation set. When comparing Table 4 with Table 3, we find that QUBICC(-trained) NNs exhibit larger MSEs than NARVAL(-trained) NNs. Causes for the higher MSEs can be attributed to the data now stemming from the entire globe and the higher stochasticity present in the higher resolution R2B5 data. Both of these reasons allow for a larger range of outputs for similar inputs, inevitably increasing the MSE of our deterministic model. Nevertheless, with the exception of the cell-based random forest, we are still well below the MSEs given by our baseline models. However, as in Section 4.1, the cell-based random forest requires much more (factor of $\approx 10^6$) memory, and a random forest of similar size to the NN has a larger MSE (85.86). The parameters for the Sundqvist scheme were again found using separate grid searches for grid cells that are mainly over land ($\{r_{\text{sat}}, r_{0,\text{top}}, r_{0,\text{surf}}, n\} = \{1.1, 0.2, 0.85, 1.62\}$) and for grid cells that are mainly over sea ($\{r_{\text{sat}}, r_{0,\text{top}}, r_{0,\text{surf}}, n\} = \{1, 0.34, 0.95, 1.35\}$). In a similar vein, estimating cloud area fraction is a more challenging task than estimating cloud volume fraction. Depending on whether a cloud primarily spans horizontally or vertically, practically any value of cloud area fraction can be attained in a sufficiently humid grid cell. This could explain the increased MSEs of the cloud area fraction models.

In Table 4 we also include bounded losses in parentheses. That means that the NN's cloud cover predictions that are smaller than 0% are set to 0% before its MSE is computed. And likewise, predictions greater than 100% are set to 100%. The difference between these two types of losses is relatively small. We can deduce that the NNs usually stay within the desired range of [0, 100]% without being forced to do so. On average, 76.4% of the predictions of all our QUBICC-trained neural networks in their respective validation sets lie within the [0, 100]%, and 95% of the predictions lie within the slightly larger [−1, 100]% range.

In Figure 4 we show that the local cell-based model—the model type with the largest MSE—is still able to reproduce the mean cloudiness statistics of the validation sets that it did not have access to during training. These validation sets each consist of the union of two blocks of 15 days, which is sufficiently temporally displaced from the training data to be above weather timescales. We can see that the validation set bias of the model corresponding to the third split is larger than that of the first two splits. The model from the second split has the overall best performance on the QUBICC data set and is therefore analyzed further in Section 4.3.

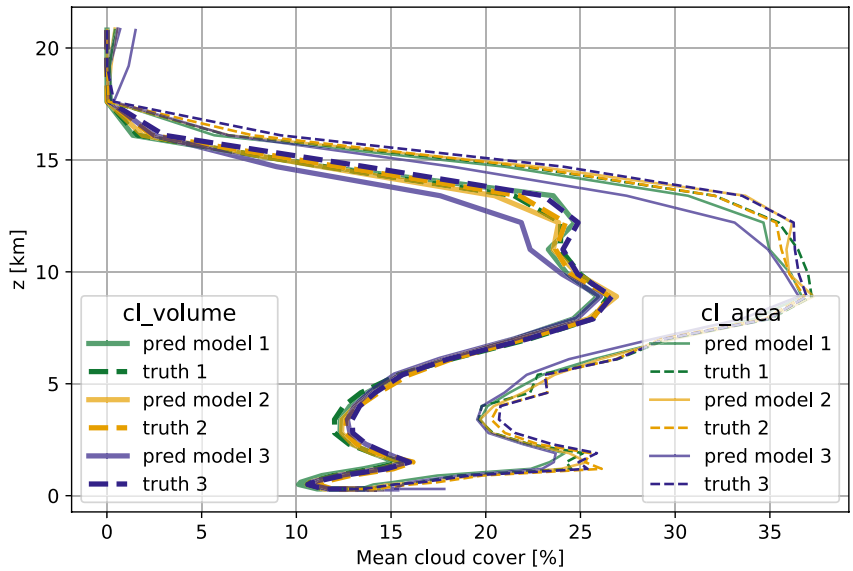


Figure 4. The cell-based cloud volume and cloud area fraction models of the 3-fold cross-validation split evaluated on their respective validation sets. The validation losses of the models from split 2 are given in Table 4.

Despite the challenging setting, Figures 5a and 5c show that the models are very well able to reproduce the average profiles of cloud volume and cloud area fraction of the global data set. The same holds true for the ability to capture the variance in time and the horizontal for a given vertical layer, which is conveyed by the R^2-values being usually well above 0.8 for all layers below 15 km. As in Figure 3, layers above 19 km had to be omitted in the R^2-plots. When it comes to reconstructing the QUBICC cloudiness, the column-based model with its large amount of adaptable parameters is able to outperform the other two model types.

After introducing and successfully evaluating both regionally and globally trained networks on their training regimes, we investigate the extent to which we can apply these NNs.

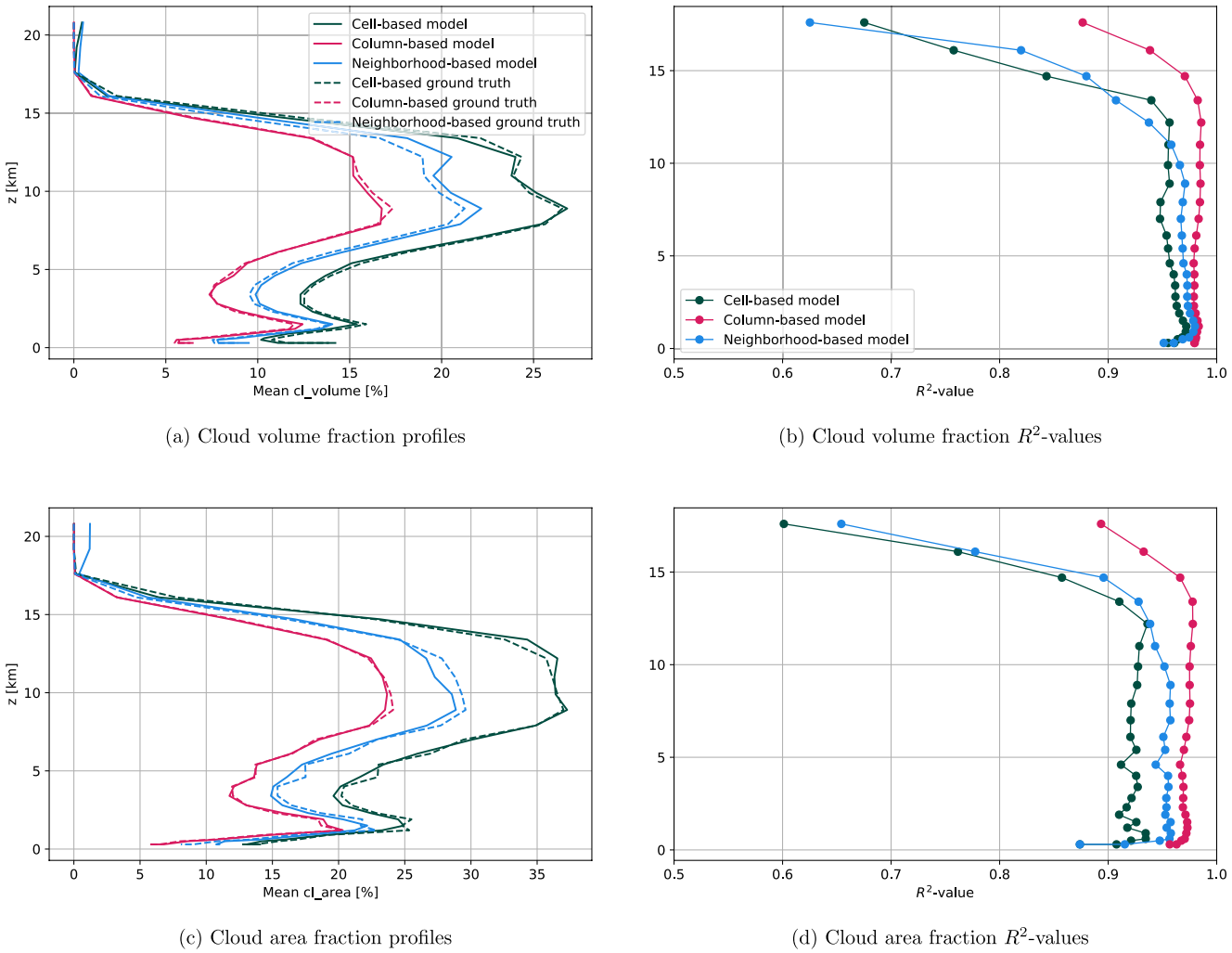
4.3. Generalization Capability

In this section we demonstrate that our globally-trained QUBICC networks can successfully be used to predict cloud cover on the distinct regional NARVAL data set. Furthermore, we show that, with the input features we chose for our NNs, achieving the converse, that is, applying regionally-trained networks on the global data set, is out of reach.

We note that, beside the regional extent, the QUBICC data covers a different timeframe and was simulated with a different physics package and on a coarser resolution (5 km) than the NARVAL data (2.5 km). As opposed to NARVAL's fractional cloudiness scheme, the QUBICC cloud cover scheme diagnosed only entirely cloudy or non-cloudy cells. These differences make the application of NNs trained on one data set to the other data set non-trivial.

4.3.1. From Global to Regional

We first study the capability of QUBICC-trained models to generalize to the NARVAL data (see Figure 6). We see that the models estimate cloud volume and cloud area fraction quite accurately. This is the case despite the significant differences between QUBICC's and NARVAL's mean vertical profiles of cloud cover. We generally recognize a decrease of R^2-value (by $\approx$0.2) when compared to the models' performance on its training data (Figure 5). A certain decrease was to be expected with the departure from the training regime. But as the R^2-values on average still exceed 0.7, we find that the models can be applied successfully to the NARVAL data. In comparison, the Sundqvist scheme we tuned on the QUBICC R2B5 data has a layer-wise averaged R^2-value of $-0.54/0.29$ for cloud volume/area fraction on the NARVAL data, but only if we discard the surface-closest layer.



(a) Cloud volume fraction profiles

(b) Cloud volume fraction R^2-values

(c) Cloud area fraction profiles

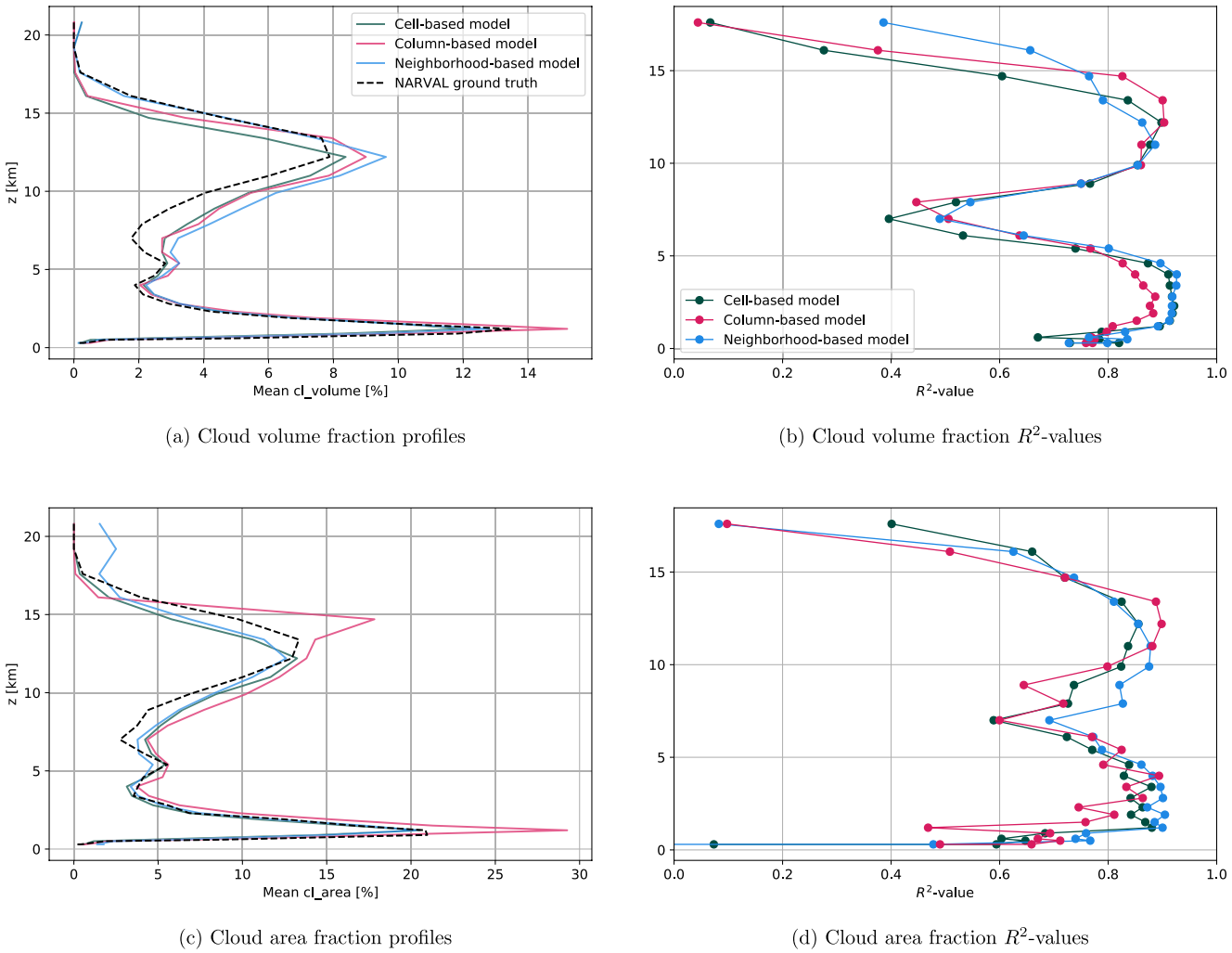
(d) Cloud area fraction R^2-values

Figure 5. Evaluation of Quasi-Biennial Oscillation in a Changing Climate (QUBICC) cloud volume and cloud area models on coarse-grained and preprocessed QUBICC R2B5 data. The layer-wise averaged R^2-values of the cell-, column-, and neighborhood-based models shown in (b) are (0.94, 0.98, 0.94) and in (d) are (0.90, 0.97, 0.93). The ground truth profiles do not match due to differences in preprocessing, especially in how many cloud-free cells were removed from the respective data sets (see Supporting Information S1 for more details). The column-based ground truth profile represents the true QUBICC cloud cover profiles since its data was not altered by preprocessing.

However, there is a significant bias affecting all three NN types, namely consistent overprediction of both cloud volume and cloud area fraction between 6 and 9 km. In this altitude range, this is visible in all four plots, either through the mismatch in mean cloud cover or the dip in R^2-value. This behavior will be further investigated in Section 4.5. Another minor bias is a slightly poorer generalization of the column-based model to the NARVAL data (see e.g., Figure 6c). We can understand this as a sign of overfitting if we also take into account that the column-based model showed a higher skill on the training data than the other two model types.

4.3.2. From Regional to Global

We have seen that the NNs are able to reproduce the cloud cover distribution of the storm-resolving NARVAL simulation, limited to its tropical region. We coarse-grain the QUBICC data to the same R2B4 grid resolution that the NARVAL NNs were trained with. This helps us to investigate to what extent the NNs can actually generalize to out-of-training regimes. We focus on the tropics first, extending the evaluation from the NARVAL region (68W–15 E, 10S–20 N) to the entire tropical band (23.4S–23.4 N). Note that the QUBICC data shows a much stronger presence of deep convection and a weaker presence of shallow and congestus-type convection. Nevertheless, the NNs are able to reproduce the general structure of the mean cloud cover profile, in particular the peak due to deep convection. The flattened peak of shallow convection is most accurately represented



(a) Cloud volume fraction profiles

(b) Cloud volume fraction R^2-values

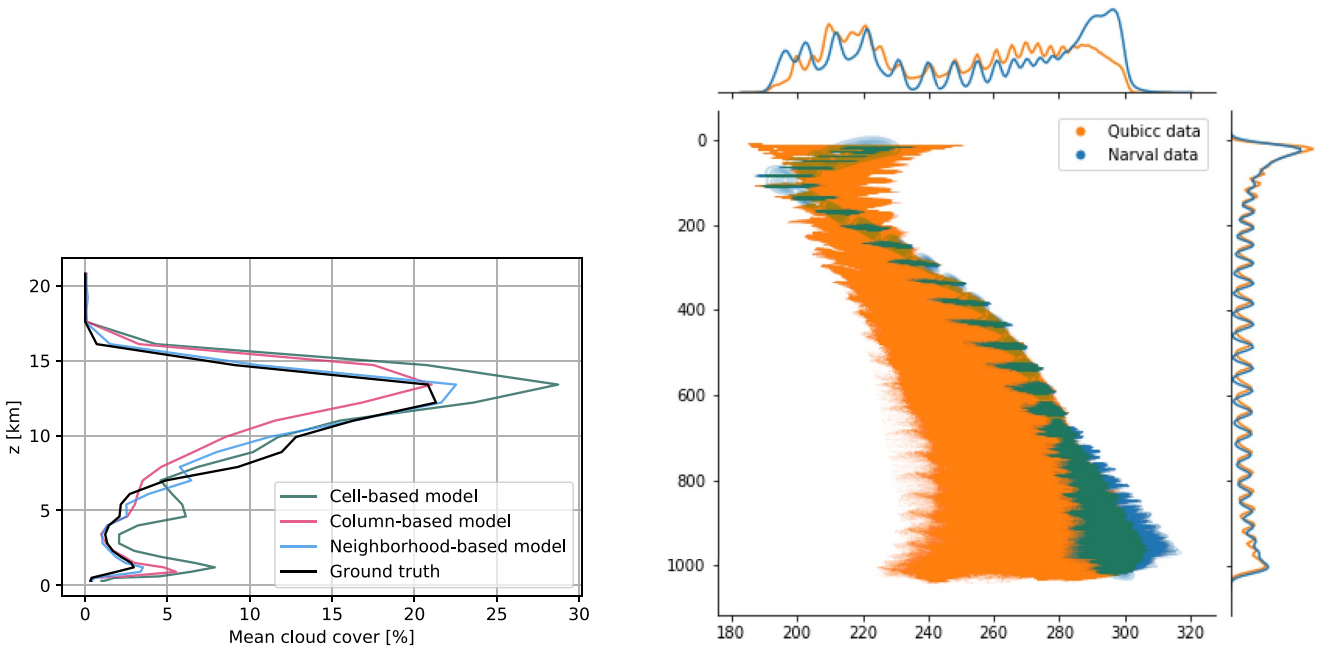
(c) Cloud area fraction profiles

(d) Cloud area fraction R^2-values

Figure 6. Evaluation of Quasi-Biennial Oscillation in a Changing Climate R2B5 cloud volume and cloud area models on NARVAL R2B5 data. The layer-wise averaged R^2-values of the cell-, column-, and neighborhood-based models shown in (b) are (0.74, 0.74, 0.79) and in (d) are (0.72, 0.71, 0.72).

by the neighborhood-based model, while the weakened congestus-type convection is reproduced by both the neighborhood- and the column-based models.

However, the NNs are not able to generalize to the entire globe. To show this, we use two column-based models as an example. Looking at Figure S4 in Supporting Information S1, we can see that they are unable to reproduce mean cloudiness statistics over the region covering the Southern Ocean and Antarctica. In addition, models with the same architecture produce entirely different cloudiness profiles. In this polar region, the NNs are evidently forced to extrapolate to out-of-training regimes and are thus unable to produce correct or consistent predictions. Let us look exclusively at the univariate distributions of the QUBICC input features (those for temperature and pressure are plotted on the margins of Figure 7b). Then we can see that their values are usually covered by the distribution of the NARVAL training data. Only their joint distribution reveals that a large number of QUBICC samples exhibit combinations of pressure and temperature that were not present in the training data. For instance, temperatures as cold as 240K never occur in tandem with pressure values as high as 1,000 hPa in the tropical training regime of the NARVAL data. This circumstance is particularly challenging for the neighborhood- and column-based models. This is because the input nodes in these two NARVAL model types correspond to specific vertical layers. So the NNs have to extrapolate when facing (during training) unseen input feature values on any vertical layer, such as in our example cold temperatures on a vertical layer located at around 1,000 hPa.



(a) Cloud cover profiles: Tropics

(b) Joint and marginal distributions of temperature and pressure

Figure 7. Panel (a): Evaluation of NARVAL R2B4 models (NARVAL region: 68W–15 E, 10S–20 N) on Quasi-Biennial Oscillation in a Changing Climate (QUBICC) R2B4 data over the tropical zone (23.4S–23.4 N). We plot the means over 10 days (20 November–29 November 2004). Different NNs of the same type produce consistent mean vertical cloudiness profiles (±1%). The layer-wise averaged R^2-values below 15 km of the cell-, column-, and neighborhood-based models are (−0.88, 0.29, 0.67), and within the upper troposphere (between 6 and 12 km) they are (0.72, 0.62, 0.84). Panel (b): Joint distribution of temperature and pressure in NARVAL R2B4 and QUBICC data. On the margins we see the univariate distributions of temperature and pressure. The jagged structure emerges from the underlying coarse vertical grid.

In this section, we demonstrated that the QUBICC NNs can be used on NARVAL data, while in our setup the converse is not feasible. This begs the question: In which way do these NNs differ and have they actually learned a meaningful dependence of cloud cover on the thermodynamic environment?

4.4. Understanding the Relationship of Predicted Cloud Cover to Its Thermodynamic Environment

In this section, our goal is to dig into the NNs and understand which input features drive the cloud cover predictions. We furthermore want to uncover similarities and differences between the NARVAL- and QUBICC-trained NNs that help understand differences in their generalization capability.

NNs are not inherently interpretable, that is, we cannot readily infer how the input features impacted a given prediction by simply looking at the networks' weights and biases. Instead, we need to use an *attribution method* that uses an explanation method built on top of the NN (Ancona et al., 2019). Within the class of attribution methods, few are adapted for regression problems. A common choice (see e.g., Brenowitz et al. (2020)) is to use gradient-based attribution methods. However, these methods may not fairly account for all inputs when explaining a model's prediction (Ancona et al., 2019). Additionally, gradient-based approaches can be strongly affected by noisy gradients (Ancona et al., 2019) and generally fail when a model is "saturated," that is, when changes in the input do not lead to changes in the output (Shrikumar et al., 2017).

Instead we approximate Shapley values for every prediction using the SHAP (SHapley Additive exPlanations) package (Lundberg & Lee, 2017). The computation of Shapley values is solidly founded in game theory and the Shapley values alone satisfy three "desirable" properties (Lundberg & Lee, 2017). Shapley values quantify the influence of how an input feature moves a specific model prediction away from its *base value*, defined as the expected output. The base value is usually an approximation of the average model output on the training data set. With Shapley values, the difference of the predicted output and the base value is fairly distributed among the input features (Molnar, 2020). A convenient property is that one can recover this difference by summing over the Shapley values ("efficiency property").

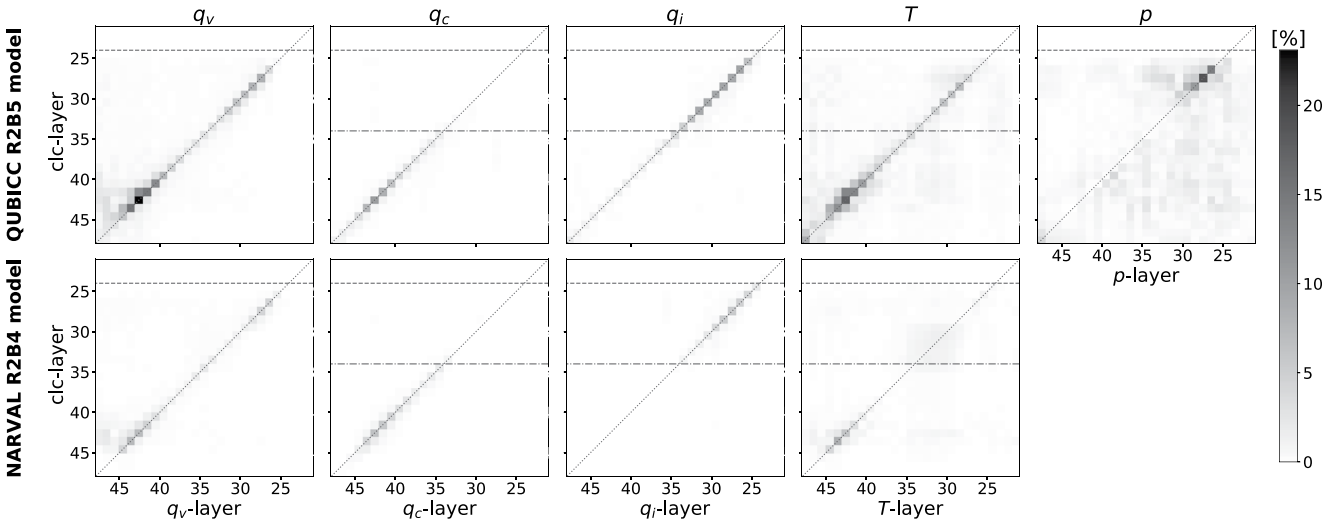


Figure 8. Average absolute SHAP values of the Quasi-Biennial Oscillation in a Changing Climate (QUBICC) R2B5 and the NARVAL R2B4 column-based models when applied to the same, sufficiently large subset of the NARVAL R2B5 data. We use the conventional ICON-A numbering of vertical layers from layer 21 (at a height of ≈20.8 km) decreasing in height to layer 47, which coincides with Earth's surface. The dashed line shows the tropopause, here at ≈15 km, the dash dotted line shows the freezing level (i.e., where temperatures are on average below 0 degrees C), here at ≈5 km. Tests with four different seeds show that the pixel values are robust (the absolute values never differ by more than 0.55%). The input features that are not shown exhibit smaller absolute SHAP values ($\rho < 1.8\%$, $p < 1.5\%$, $z_g < 0.7\%$, *land/lake* <0.1%) everywhere and are thus omitted.

The DeepExplainer within the SHAP package is able to efficiently compute approximations of Shapley values for deep NNs (Lundberg & Lee, 2017). SHAP also comes with various visualization methods, which allow us to aggregate local sample-based interpretations to form global model interpretations.

We now show how we use SHAP to compare the way NARVAL (R2B4)- and QUBICC (R2B5)-trained networks arrive at good predictions. We focus on the column-based (cloud volume fraction) models. These are uniquely able to uncover important non-local effects, have the largest number of input features to take into account and have on average the lowest MSEs in their training regimes (taking into account both Tables 3 and 4).

We collect local explanations on a sufficiently large subset of the NARVAL R2B5 data. For this, we compute the base values by taking the average model predictions on subsets of the respective training data sets. A necessary condition for the base value is that it approximates the expected NN output (on the entire training set) well. We found that $\approx 10^4$ QUBICC samples are sufficient for the average NN prediction to converge. Therefore, we used this size for the random subsets of the QUBICC and of the smaller NARVAL training set as well. We showed that on the NARVAL R2B5 data set, the QUBICC models are able to reconstruct the mean vertical profile with high R^2-values (Figure 6). Impressively, the column-based version of our NARVAL R2B4 models also makes successful predictions on the NARVAL R2B5 data set (with an average R^2-value of 0.93; Figure S5 in Supporting Information S1) despite the doubling of the horizontal resolution.

The size of the subset of NARVAL R2B5 data ($\approx 10^4$ samples) is chosen to be sufficiently large to yield robust estimates of average absolute Shapley values. Averaging the absolute Shapley values over many input samples measures the general importance of each input feature on the output. An input feature with a large average absolute Shapley value contributes strongly to a change in the model output. It on average increases or decreases the model output by precisely this value.

The absolute SHAP values (Figure 8) suggest that both models learned a remarkably local mapping, with a clear emphasis on the diagonal (especially above the boundary layer). That means that the prediction at a given vertical layer mostly depends on the inputs at the same location. The models have learned to act like our cell- or neighborhood-based models without human intervention.

The input features have a larger influence in the QUBICC model than they do in the NARVAL model. We can also see this phenomenon, if we use a similar base value for both models (see Figure S6 in Supporting Information S1). This is most likely due to the fact that the QUBICC model was exposed to a wide variety of climatic

conditions across the entire globe during training, resulting in a greater variance in cloud cover. The NN is thus used to deviate from the average cloud cover, putting more emphasis on its input features, and consequently causing larger Shapley values.

Both models take into account that in the boundary layer the supply of moisture q_v from below in combination with temperature anomalies that could drive convective lifting influence the sub-grid distribution of cloud condensates and henceforth cloud cover. Such a non-local mixing due to updrafts presents limitations for purely local parameterizations. In the boundary layer (which we define to be at below 1 km), temperature T and specific humidity q_v are found to be the most important variables (having the largest sum of absolute SHAP values) for the NNs. Higher in the troposphere, the local amount of moisture has a significant impact on cloud cover. Specific cloud liquid water content q_c is a major predictor of cloud cover below the freezing level, while specific cloud ice content q_i is a major predictor of cloud cover above the freezing level. In contrast to the global QUBICC model, the tropical NARVAL model only considers the impact of q_i at sufficiently high altitudes, which allow for the formation of cloud ice. The QUBICC model also learned to place more emphasis on T and q_v in the lower troposphere and pressure p in the higher troposphere than the NARVAL model.

Generally, the most important variables above the boundary layer and below the freezing level are temperature T (for the QUBICC model) and cloud water q_c (for the NARVAL model). Above the freezing level, the QUBICC model emphasizes pressure p most, while the NARVAL model learns a similar impact of T, q_i, and p (not shown). Due to the Clausius-Clapeyron relation, RH depends most strongly on temperature. Taking into account that throughout the troposphere RH is the best single indicator for cloud cover (Walcek, 1994), this is a likely explanation for the models' large emphasis on temperature.

After using SHAP to illustrate which features drive the (column-based) NN predictions, we use the same approach to understand the source of a specific generalization error of the QUBICC NNs (Figure 6).

4.5. Understanding Model Errors

In this section, our goal is to understand the source of flawed NN predictions. We want to analyze what type of dependence on which input features is most responsible for erroneous predictions. This type of analysis reveals differences in the (NN-learned) characteristics of the training data set and a data set to which an NN is applied to.

In the evaluation of the QUBICC (R2B5) cloud volume fraction models on NARVAL R2B5 data (Figure 6) we have seen a pronounced dip in performance ($R^2 \leq 0.8$ for all models) on a range of altitudes between 6 and 9 km. The dip was accompanied by an overestimation of cloud cover (relative error >15%). We specifically focus on explaining the bias at 7 km. The vertical layer that corresponds to this altitude is the 32nd ICON-A layer. On layer 32, the R^2-values are minimal ($R^2 \leq 0.5$ for all models) making it arguably the largest tropospheric generalization error of the models. However, the method we employ here can be used to understand other generalization errors as well.

The NARVAL (R2B4) models are perfectly able to make predictions on NARVAL R2B5 data on layer 32 (Figure S5 in Supporting Information S1), making it a suitable benchmark model. As in the previous section we use SHAP on the column-based models. In order to be able to compare Shapley values corresponding to certain features individually, we follow the strategy outlined in Appendix B.

Figure 9a shows the influence of each input feature from the entire grid column on the average model output on layer 32. We find that the QUBICC model bias is driven by q_v and q_i. Compared to the NARVAL model, the QUBICC model clearly overestimates the impact of these two variables. This impact is dampened somewhat by a net decreasing effect of p and T on the cloud cover predictions. In the NARVAL model the impact of these features is much less pronounced. The reason is probably once again that the model has not learned the need for deviating much from the base value in its tropical training regime.

When investigating the vertical profile of Shapley values in Figures 9b and 9c we find that the local values have the largest effect on cloud cover. This local importance is also corroborated by Figure 8. We can zoom in and look at the more precise conditionally-averaged functional dependence of clc_32 on these local q_i_32 and q_v_32 variables (Figures 9d and 9e). We find the two functions to be very similar, albeit differing in their slope. The QUBICC model quickly increases cloud cover with increasing values of q_i_32 and q_v_32. The QUBICC model's

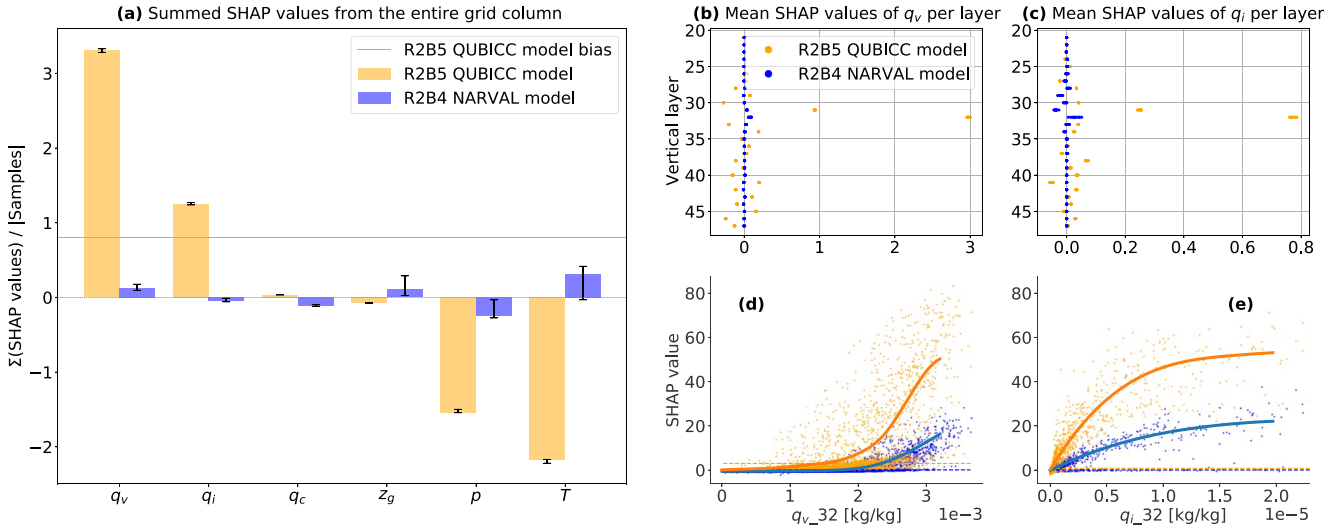


Figure 9. SHAP/Shapley value statistics per input feature for cloud cover predictions on vertical layer 32 (at ≈ 7 km) of the column-based models with a focus on q_v and q_i in (b–e). Input features the models have not in common are neglected. As in Figure 8, the Shapley values for both models are computed on the same sets of 10^4 random NARVAL R2B5 samples (using 10 different seeds). (a): The sum of average SHAP values over all vertical layers. The black lines show the range of values (min/max). The absolute Quasi-Biennial Oscillation in a Changing Climate R2B5 model bias (of 0.95%) on layer 32 (cf. Figure 6a) can approximately be recovered by summing over all orange values (which yields 0.81%). (b, c): The vertical profiles of SHAP values for q_v and q_i for all 10 seeds. In the SHAP dependence plots (d, e) we zoom in on the features with the largest SHAP values (q_i and q_v of layer 32). (d, e): Each dot corresponds to one NARVAL R2B5 sample. The lines show smoothed conditional expectations computed over all seeds. The dashed lines show the average SHAP value of the input features q_v and q_i on layer 32 whose values can also be found in (b and c).

large emphasis on q_i_32 could be a relict from the cloud cover scheme in the native QUBICC data. This scheme had set cloud cover to 100%, whenever the cloud condensate ratio had exceeded a given threshold.

5. Summary

In this study we develop the first machine learning based parameterization for cloud cover based on the ICON model and deep NNs. We train the NNs with coarse-grained data from regional and global SRM simulations with real geography. We demonstrate that in their training regime, the NNs are able to learn the sub-grid scale cloud cover from large-scale variables (Figures 3 and 5). Additionally we show that our globally trained NNs can also be successfully applied to data originating from a regional simulation that differs in many respects (e.g., its physics package, horizontal/vertical resolution, and time frame; Figure 6). Using SHAP we compare regionally and globally trained NNs to understand the relationship between predicted cloud cover and its thermodynamic environment and vertical structure (Figure 8). We are able to uncover that specific humidity and cloud ice are the drivers of one NN's largest tropospheric generalization error (Figure 9).

We implement three different types of NNs in order to assess the degree of (vertical) locality and the amount of information they need when it comes to the task of diagnosing cloud cover. We find that by enforcing more locality, the performance of the NN suffers on its training set (Figures 3 and 5). However, the more local cell- and neighborhood-based NNs show slightly fewer signs of overfitting the training data (Figure 6). Generally we find that none of the three types clearly outperforms the other two types and that the potentially non-local model in actuality also mostly learned to disregard non-local effects (Figure 8). Overall, the neighborhood-based model trained on the global QUBICC data (Q3) is most likely the preferable model. It has a good accuracy on the training data, the lowest generalization error on the NARVAL data, is low-dimensional, easy to implement and cross-model compatible. The last point refers to the fact that (unlike the column-based model) it is not tied to the vertical grid it was trained on.

Furthermore, the NNs are trained to differentiate between cloud volume and cloud area fraction, which are distinct interpretations of cloud cover. We found cloud area fraction to be a somewhat more difficult value to predict. The shape of a cloud, which determines its cloud area fraction, is harder to extract from grid-scale averaged thermodynamic variables. We agree with Brooks et al. (2005) that a distinction between these two concepts

of cloud cover would be expedient inside a general circulation model for two reasons: First, both interpretations are used in the microphysics and radiation schemes. Second, depending on the interpretation, cloud cover can differ significantly (Figure A2).

The natural next step will be to implement and evaluate the machine learning based parameterization for cloud cover in the ICON model. In such an ICON-ML model, the machine learning based parameterization would substitute the traditional cloud cover parameterization. The NN predictions for cloud area and cloud volume fraction would be used as parameters for the radiation and microphysics parameterizations, depending on which interpretation is most appropriate in each case. Preliminary online simulations covering one QUBICC month (not shown) demonstrate the potential of our neighborhood-based NN parameterization as it is (a) able to process its input variables from the coarse-scale distributions while (b) pushing the statistics of, for example, the cloud water mixing ratio, to that of the (coarse-grained) high-resolution statistics as desired. However, as we discussed in the introduction, more work is required to create an ICON-ML model that produces accurate and robust results.

The presence of condensate-free clouds in the training data shows inaccuracies that are present both in the NARVAL and the QUBICC training data. These could have been avoided by introducing targeted multiple calls to the same parameterization scheme in the high-resolution model that generated the data. However, we emphasize that the machine learning approach is general enough that if the data were generated more carefully then our approach would still work.

Our regionally-trained networks are not able to generalize to the entire globe. Similar difficulties might arise when applying our globally-trained networks to a very different climate (Rasp et al., 2018). In practice, this would require us to filter out data samples which the NN cannot process in a meaningful way. Alternatively, one could train the NNs with climate-invariant features only, eliminating the need of ever extrapolating to out-of-training distributions (Beucler et al., 2021). By additionally using causal discovery methods to guide their selection, one would most likely arrive at a more rigorous and physically consistent set of input features (Nowack et al., 2020; Runge et al., 2019). Another useful modification to our NNs would be to add a method that allows us to estimate the uncertainty associated with a prediction, for example, either by adding dropout (Gal & Ghahramani, 2016) or by implementing the NNs as Bayesian NNs.

From a climate science perspective, instead of diagnosing cloud cover from large-scale variables directly, one could also train an NN to output parameters specifying distributions for sub-grid scale temperature and moisture. Cloud cover could then be derived from these distributions (see *statistical cloud cover schemes* in e.g., Stensrud (2009); A. M. Tompkins (2002)). By reusing the distributions for other parameterizations as well, we could increase the consistency among cloud parameterizations. However, this approach would require us to make assumptions concerning the general form of these distributions (Larson, 2017) and we leave this for future work.

Overall, this study demonstrated the potential of deep learning combined with high-resolution data for developing parameterizations of cloud cover.

Appendix A: Coarse-Graining Methodology

Our goal is to best estimate grid-scale mean values. Ideally, we would derive the large-scale grid-scale mean $\bar{S}$ of a given variable S by integrating over the grid cell volume $V \subseteq \mathbb{R}^3$. In practice, we compute a weighted sum over the values $S_{i,j}$ of all high-resolution grid cells H. Here, i is the horizontal and j is the vertical index of a high-resolution grid cell. We define the weights $\alpha_{i,j} \in [0, 1]$ as the fraction of V that a high-resolution grid cell indexed by (i, j) fills. This is a basic discretization of the integral.

To make this term easier to compute in practice, we introduce another approximation. Instead of computing $\alpha_{i,j}$ directly, we split it into the fraction of the horizontal area of V (denoted by $\gamma_i \in [0, 1]$) *times* the fraction of the vertical thickness of V (denoted by $\beta_j \in [0, 1]$) that the high-resolution grid cell indexed by (i, j) fills. We first compute the weights γ_i and the weighted sum over the horizontal indices i (horizontal coarse-graining). Only afterward do we compute the weights β_j and the weighted sum over the vertical indices j (vertical coarse-graining).

Note that this is indeed an approximation. The geometric heights and vertical thicknesses of grid cells in H on a specific vertical layer j do not need to match exactly. These slight differences are lost when horizontally coarse-graining to fewer grid boxes. Therefore, the second approximation is an approximation because we (a)

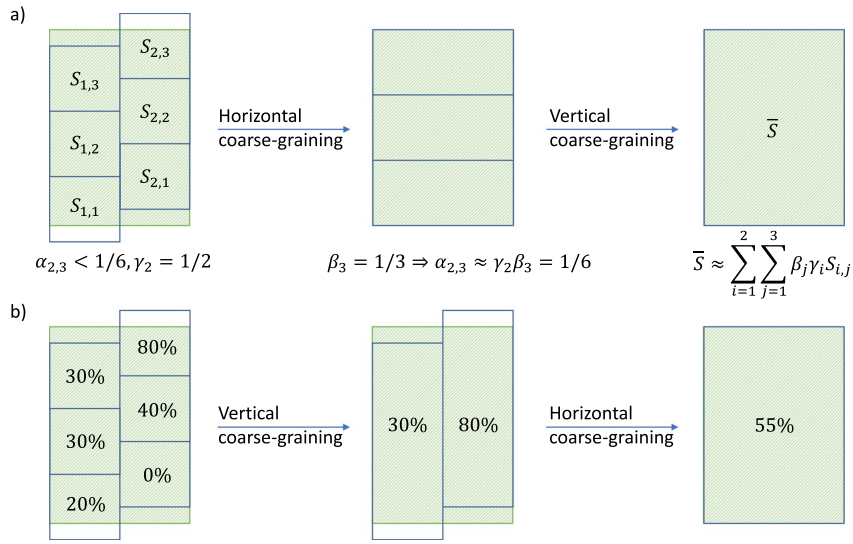


Figure A1. Sketch of our general coarse-graining methodology in panel (a) and for cloud area fraction in panel (b). We picture a vertical slice through two grid columns. For simplicity we assume that the grid boxes all have the same depth. The greenly hatched area depicts a coarse-scale grid box V. Panel (a): Due to our approximation the weight $\alpha_{2,3}$ for the value in grid box $S_{2,3}$ is 1/6 and therefore larger than it were without the sequential horizontal and vertical coarse-graining steps. Panel (b): In the vertical range of V we vertically coarse-grain cloud cover values according to a maximum overlap assumption before we coarse-grain in the horizontal.

compute the vertical overlap β_j *after* we horizontally coarse-grain the grid cells and (b) work on a terrain-following height grid which allows for vertical layers of varying heights over mountaineous land areas. Over ocean areas, where the height levels have no horizontal gradient, this simplification in the computation of the weights has no disadvantage.

In short, let $\alpha_{i,j}$, β_j, $\gamma_i \in [0, 1]$ be the weights describing the amount of overlap in volume/vertical/horizontal between the high-resolution grid cells and the low-resolution grid cell. We then calculate the large-scale grid-scale mean as the weighted sum of high-resolution variables

$$\bar{S} \equiv \frac{1}{|V|} \int_V S dx \approx \sum_{(i,j)\in H} \alpha_{i,j} S_{i,j} \approx \sum_{(i,j)\in H} \beta_j \gamma_i S_{i,j}. \tag{A1}$$

We also illustrate our approach in panel a) of Figure A1.

The use of spring dynamics in between model grid refinement steps allows for the presence of fractional horizontal overlap γ_i. As our method for horizontal coarse-graining we choose the first order conservative remapping from the Climate Data Operators package (Schulzweida, 2019), which is able to handle fractional overlap and the irregular ICON grid to coarse-grain to and from.

There are locations where the low-resolution grid cells that are closest to Earth's surface extend significantly further downwards than the high-resolution grid cells. This is due to topography that can only be seen at fine scales and makes it difficult to endue these low-resolution grid cells with a meaningful average computed from the high-resolution cells. We therefore omit these grid cells during coarse-graining. This issue is present only in scattered, isolated grid cells over land and it affects a small fraction of all grid cells (0.2%) and columns (4.7%). So it does not pertain entire regions, which would decrease the scope and quality of the data set. While horizontally coarse-graining NARVAL data, we analogously omit low-resolution grid cells that are not located entirely inside the NARVAL region.

To derive the cloud area fraction C we cannot start by coarse-graining horizontally. We first need to utilize the high-resolution information on whether the fractional cloud cover on vertically consecutive layers of a low-resolution grid column overlaps or not. Therefore, we first vertically coarse-grain cloud cover to a grid that would—after subsequently horizontally coarse-graining—resemble the ICON-A grid as much as possible. For the first step, we assume maximum overlap as the level separation of vertical layers is relatively small. We

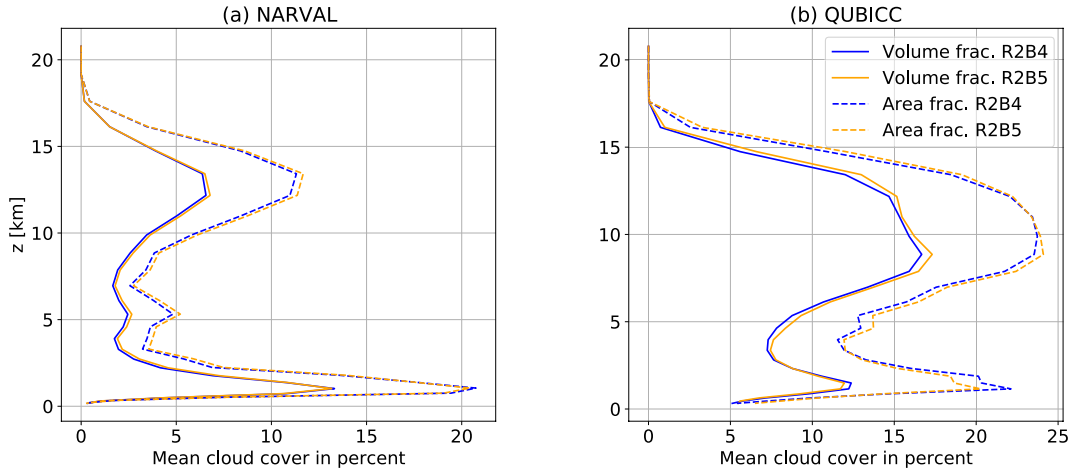


Figure A2. Comparison of the coarse-grained mean cloud volume and mean cloud area fraction profiles for (a) NARVAL and (b) QUBICC. In a given grid cell, the cloud volume fraction is never greater than the cloud area fraction. Close to the surface, the grid cell thickness and thus also the vertical sub-grid variability of clouds is small. There it follows that the cloud area fraction is approximately equal to the cloud volume fraction.

thus calculate the coarse-grained cloud area fraction $\overline{C}$ as the sum of the vertically maximal cloud cover values $\max_j\{C_{i,j}\}$ weighted by the horizontal grid cell overlap fractions γ_i

$$\overline{C} = \sum_{(i,j)\in H} \gamma_i \max_j\{C_{i,j}\}\,. \tag{A2}$$

Equation A2 is exemplified in panel b) of Figure A1. For QUBICC grid cells, which are always either fully cloudy or cloud-free, we can directly interpret Equation A2 as returning the fraction of high-resolution horizontal grid points that are covered by a cloud of any non-zero vertical extent within a coarse vertical cell. Due to the fractional cloudiness and the maximum overlap assumption, this link is less direct for the NARVAL data. Figure A2 illustrates the different mean vertical profiles of cloud volume fraction and cloud area fraction. Considerable differences in their coarse-grained vertical profiles (differing absolutely by almost 10% on some layers) corroborate the need to distinguish these two concepts of cloud cover.

Appendix B: Comparing Two Neural Networks Using Attribution Methods

We use SHAP to compare two neural networks and to decompose model errors. However, our error decomposition framework can be used with any attribution method (Layer-wise Relevance Propagation, Local Interpretable Model-agnostic Explanations, Integrated Gradients, etc., Samek et al. (2019)) which fulfills the property that the attributed feature importances sum up to the predicted model output (possibly shifted by a constant value).

For a given NN h, data sample X and input feature i, the SHAP package computes the corresponding Shapley value $\phi_{h,X,i}$. Shapley values satisfy the so-called efficiency property for every sample, which means that they sum up to the difference between the model output and its *base value* (the expected model output)

$$\sum_{i\in I} \phi_{h,X,i} = h(X) - \mathbb{E}[h(X)], \tag{B1}$$

where $I \subseteq \mathbb{N}$ consists of the features' indices. A Shapley value $\phi_{f,X,i}$ can thus be interpreted as the amount by which an input feature i contributes to the deviation of f's prediction from the base value. Shapley values are constructed so that $f(X) - \mathbb{E}[f(X)]$ is fairly distributed among the features.

Let f be the QUBICC R2B5 and g the NARVAL R2B4 NN. Their base values $B_f := \mathbb{E}[f(X)]$ and $B_g := \mathbb{E}[g(X)]$ are computed as the average prediction of f and g on a subset of their respective training data sets (the so-called *background data set*). By repeatedly drawing an appropriate sample from the training set of f, we can construct its background data set such that $B_f = B_g$. Plugging f and g into (Equation B1) we get

$$\sum_{i\in I}\phi_{f,X,i}-\sum_{j\in J}\phi_{g,X,j}=f(X)-g(X)+B_f-B_g=f(X)-g(X), \tag{B2}$$

where $I, J \subseteq \mathbb{N}$. Let S be a random subset of the NARVAL R2B5 data and the overline $\bar{\cdot}$ denote the average over all samples in S. The size of S is chosen to be large enough such that (a) $\overline{f}$ and $\overline{g}$ are good approximations of the predicted averages of f and g on the entire NARVAL R2B5 data set (as shown in Figure 6a and Figure S5a in Supporting Information S1) and (b) the mean Shapley values are robustly estimated.

The sum of Shapley values corresponding to input features that are present in only one model (such as ρ) are in our case very small (absolute value < 0.08) and thus negligible. Hence, by averaging over (Equation B2) we can approximate the mismatch between the average outputs of f and g by the sum of the difference of averaged Shapley values corresponding to features that f and g have in common

$$\begin{aligned}\overline{f}-\overline{g} &= \sum_{i\in I\cap J}\left(\overline{\phi_{f,X,i}}-\overline{\phi_{g,X,i}}\right)+\sum_{i\in I\setminus J}\overline{\phi_{f,X,i}}-\sum_{i\in J\setminus I}\overline{\phi_{g,X,i}}\\ &\approx \sum_{i\in I\cap J}\left(\overline{\phi_{f,X,i}}-\overline{\phi_{g,X,i}}\right).\end{aligned} \tag{B3}$$

So by comparing $\overline{\phi_{f,X,i}}$ and $\overline{\phi_{g,X,i}}$ for all common features $i \in I \cap J$ individually, we can explain which input features contribute to the difference between $\overline{f}$ and $\overline{g}$. Having ensured that S satisfies (a) and (b), we can generalize (Equation B3) to the entire NARVAL R2B5 data set.

Data Availability Statement

The neural network and analysis code can be found at https://github.com/agrundner24/iconml_clc and is preserved at DOI: https://doi.org/10.5281/zenodo.5788873. Primary data used in this work is archived by the Max Planck Institute for Meteorology (contact: marco.giorgetta@mpimet.mpg.de). The coarse-grained model output used for training the neural networks amounts to several TB. An extract from the training data is made available in the GitHub repository. The software code for the ICON model is available from https://code.mpimet.mpg.de/projects/iconpublic.

Acknowledgments
We thank the associate editor, two anonymous reviewers, and an internal reviewer for their insightful and constructive comments that helped to improve the manuscript. Funding for this study was provided by the European Research Council (ERC) Synergy Grant "Understanding and Modelling the Earth System with Machine Learning (USMILE)" under the Horizon 2020 research and innovation programme (Grant agreement No. 855187). Beucler acknowledges funding from the Columbia University sub-award 1 (PG010560-01). Gentine acknowledges funding from the NSF Science and Technology Center, Center for Learning the Earth with Artificial Intelligence and Physics (LEAP) (Award 2019625). We thank the Max Planck Institute for Meteorology for providing access to the NARVAL simulation data. Further we acknowledge PRACE for awarding us access to Piz Daint at ETH Zurich/CSCS, Switzerland, which made the QUBICC simulations possible (ID 2019215178). This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID bd1083 and bd1179. Open Access funding enabled and organized by Projekt DEAL.

References

Allen, M. R., & Ingram, W. J. (2002). Constraints on future changes in climate and the hydrologic cycle. *Nature*, *419*(6903), 228–232. https://doi.org/10.1038/nature01092

Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-based attribution methods. In *Explainable AI: Interpreting, explaining and visualizing deep learning* (pp. 169–191). Springer. https://doi.org/10.1007/978-3-030-28954-6_9

Barker, H. W., Stephens, G., Partain, P., Bergman, J., Bonnel, B., Campana, K., et al. (2003). Assessing 1d atmospheric solar radiative transfer models: Interpretation and handling of unresolved clouds. *Journal of Climate*, *16*(16), 2676–2699. https://doi.org/10.1175/1520-0442(2003)016<2676:adasrt>2.0.co;2

Beucler, T., Pritchard, M., Gentine, P., & Rasp, S. (2020). Towards physically-consistent, data-driven models of convection. In *Igarss 2020-2020 ieee international geoscience and remote sensing symposium*. https://doi.org/10.1109/IGARSS39084.2020.9324569

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., et al. (2021). Climate-invariant machine learning. arXiv preprint arXiv:2112.08440.

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parameterizations 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

Brooks, M. E., Hogan, R. J., & Illingworth, A. J. (2005). Parameterizing the difference in cloud fraction defined by area and by volume as observed with radar and lidar. *Journal of the Atmospheric Sciences*, *62*(7), 2248–2260. https://doi.org/10.1175/JAS3467.1

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., & Palmer, T. (2021). Machine learning emulation of gravity wave drag in numerical weather forecasting. *Journal of Advances in Modeling Earth Systems*, *13*(7), e2021MS002477. https://doi.org/10.1029/2021MS002477

Chevallier, F., Morcrette, J.-J., Cheruy, F., & Scott, N. A. (2000). Use of a neural-network-based long-wave radiative-transfer scheme in the ECMWF atmospheric model. *Quarterly Journal of the Royal Meteorological Society*, *126*(563), 761–776. https://doi.org/10.1002/qj.49712656318

Crueger, T., Giorgetta, M. A., Brokopf, R., Esch, M., Fiedler, S., Hohenegger, C., et al. (2018). Icon-a, the atmosphere component of the icon Earth system model: II. Model evaluation. *Journal of Advances in Modeling Earth Systems*, *10*(7), 1638–1662. https://doi.org/10.1029/2017ms001233

Doms, G., Förstner, J., Heise, E., Herzog, H., Mironov, D., Raschendorfer, M., et al. (2011). *A description of the nonhydrostatic regional Cosmo model, part II: Physical parameterization*. Deutscher Wetterdienst.

Eyring, V., Mishra, V., Griffith, G. P., Chen, L., Keenan, T., Turetsky, M. R., et al. (2021). 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

Gagne, D. J., Christensen, H. M., Subramanian, A. C., & Monahan, A. H. (2020). Machine learning for stochastic parameterization: Generative adversarial networks in the Lorenz'96 model. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001896. https://doi.org/10.1029/2019ms001896

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd international conference on machine learning*.

Gentine, P., Eyring, V., & Beucler, T. (2021). Deep learning for the parametrization of subgrid processes in climate models. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 307–314.

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., Gagne, D. J., Chen, C.-C., Christensen, M. W., Lebo, Z. J., Morrison, H., & Gantos, G. (2021). Machine learning the warm rain process. *Journal of Advances in Modeling Earth Systems*, *13*(2), e2020MS002268. https://doi.org/10.1029/2020ms002268

Giorgetta, M. A., Crueger, T., Brokopf, R., Esch, M., Fiedler, S., Hohenegger, C., 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), 1638–1662. https://doi.org/10.1029/2017ms001233

Giorgetta, M. A., Sawyer, W., Lapillonne, X., Adamidis, P., Alexeev, D., Clément, V., et al. (2022). The ICON-a model for direct QBO simulations on GPUs (version icon-cscs:baf28a514). *Geoscientific Model Development*, *15*(18), 6985–7016. https://doi.org/10.5194/gmd-15-6985-2022

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

Hertel, L., Collado, J., Sadowski, P., Ott, J., & Baldi, P. (2020). Sherpa: Robust hyperparameter optimization for machine learning. *SoftwareX*, *12*, 100591. https://doi.org/10.1016/j.softx.2020.100591

Hohenegger, C., Kornblueh, L., Klocke, D., Becker, T., Cioni, G., Engels, J. F., et al. (2020). Climate statistics in global simulations of the atmosphere, from 80 to 2.5 km grid spacing. *Journal of the Meteorological Society of Japan*, *98*(1), 73–91. https://doi.org/10.2151/jmsj.2020-005

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, *4*(2), 251–257. https://doi.org/10.1016/0893-6080(91)90009-t

Khairoutdinov, M., Randall, D., & DeMott, C. (2005). Simulations of the atmospheric general circulation using a cloud-resolving model as a superparameterization of physical processes. *Journal of the Atmospheric Sciences*, *62*(7), 2136–2154. https://doi.org/10.1175/JAS3453.1

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

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370–1383. https://doi.org/10.1175/MWR2923.1

Larson, V. E. (2017). Clubb-silhs: A parameterization of subgrid variability in the atmosphere. arXiv preprint arXiv:1711.03675.

Lohmann, U., & Roeckner, E. (1996). Design and performance of a new cloud microphysics scheme developed for the ECHAM general circulation model. *Climate Dynamics*, *12*(8), 557–572. https://doi.org/10.1007/BF00207939

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *31st conference on neural information processing systems*.

Mauritsen, T., Bader, J., Becker, T., Behrens, J., Bittner, M., Brokopf, R., et al. (2019). Developments in the MPI-M Earth system model version 1.2 (MPI-ESM1.2) and its response to increasing CO_2. *Journal of Advances in Modeling Earth Systems*, *11*(4), 998–1038. https://doi.org/10.1029/2018ms001400

Mauritsen, T., Svensson, G., Zilitinkevich, S. S., Esau, I., Enger, L., & Grisogono, B. (2007). A total turbulent energy closure model for neutrally and stably stratified atmospheric boundary layers. *Journal of the Atmospheric Sciences*, *64*(11), 4113–4126. https://doi.org/10.1175/2007jas2294.1

Meehl, G., Senior, C., Eyring, V., Flato, G., Lamarque, J., Stouffer, R., et al. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the CMIP6 Earth system models. *Science Advances*, *6*(26). https://doi.org/10.1126/sciadv.aba1981

Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J., & Clough, S. A. (1997). Radiative transfer for inhomogeneous atmospheres: RRTM, a validated correlated-k model for the longwave. *Journal of Geophysical Research*, *102*(D14), 16663–16682. https://doi.org/10.1029/97jd00237

Molnar, C. (2020). *Interpretable machine learning*. Lulu.com.

Mooers, G., Tuyls, J., Mandt, S., Pritchard, M., & Beucler, T. (2020). Generative modeling of atmospheric convection. *Proceedings of the 10th international conference on climate informatics*. https://doi.org/10.1145/3429309.3429324

Müller, S. (2019). *Convectively generated gravity waves and convective aggregation in numerical models of tropical dynamics*. (Doctoral dissertation), Universität Hamburg Hamburg. https://doi.org/10.17617/2.3025587

Nowack, P., Runge, J., Eyring, V., & Haigh, J. D. (2020). Causal networks for climate model evaluation and constrained projections. *Nature Communications*, *11*(1), 1415. https://doi.org/10.1038/s41467-020-15195-y

Ott, J., Pritchard, M., Best, N., Linstead, E., Curcic, M., & Baldi, P. (2020). *A fortran-keras deep learning bridge for scientific computing* (Vol. 2020, pp. 1–13). Scientific Programming. https://doi.org/10.1155/2020/8888811

Pincus, R., Mlawer, E. J., & Delamere, J. S. (2019). Balancing accuracy, efficiency, and flexibility in radiation calculations for dynamical models. *Journal of Advances in Modeling Earth Systems*, *11*(10), 3074–3089. https://doi.org/10.1029/2019MS001621

Pincus, R., & Stevens, B. (2013). Paths to accuracy for radiation parameterizations in atmospheric models. *Journal of Advances in Modeling Earth Systems*, *5*(2), 225–233. https://doi.org/10.1002/jame.20027

Prill, F., Reinert, D., Rieger, D., Zängl, G., Schröter, J., Förstner, J., et al. (2019). Icon tutorial: Nwp mode and icon-art [Computer software manual]. ICON. Retrieved from https://code.mpimet.mpg.de/attachments/19568/ICON_tutorial_2019.pdf

Raddatz, T., Reick, C., Knorr, W., Kattge, J., Roeckner, E., Schnur, R., et al. (2007). Will the tropical land biosphere dominate the climate–carbon cycle feedback during the twenty-first century? *Climate Dynamics*, *29*(6), 565–574. https://doi.org/10.1007/s00382-007-0247-8

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

Raschendorfer, M. (2001). The new turbulence parameterization of lm. *COSMO newsletter*, *1*, 89–97.

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

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., et al. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(1), 2553. https://doi.org/10.1038/s41467-019-10105-3

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, explaining and visualizing deep learning* (Vol. 11700). Springer Nature.

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., Schar, 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

Schrodin, R., & Heise, E. (2001). *The multi-layer version of the dwd soil model terra_lm*. DWD.

Schulz, J.-P., Vogel, G., Becker, C., Kothe, S., & Ahrens, B. (2015). Evaluation of the ground heat flux simulated by a multi-layer land surface scheme using high-quality observations at grass land and bare soil. In *Egu general assembly conference abstracts* (p. 6549).

Schulzweida, U. (2019). Cdo user guide. https://doi.org/10.5281/zenodo.3539275

Seifert, A. (2008). A revised cloud microphysical parameterization for COSMO-LME. *COSMO Newsletter*, *7*, 25–28.

Seifert, A., & Rasp, S. (2020). Potential and limitations of machine learning for modeling warm-rain cloud microphysical processes. *Journal of Advances in Modeling Earth Systems*, *12*(12). https://doi.org/10.1029/2020ms002301

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning* (pp. 3145–3153).

Stensrud, D. J. (2009). *Parameterization schemes: Keys to understanding numerical weather prediction models*. Cambridge University Press.

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*, *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. (2019). 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., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., et al. (2019). Dyamond: The dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1), 61. https://doi.org/10.1186/s40645-019-0304-z

Sundqvist, H., Berge, E., & Kristjánsson, J. E. (1989). Condensation and cloud parameterization studies with a mesoscale numerical weather prediction model. *Monthly Weather Review*, *117*(8), 1641–1657. https://doi.org/10.1175/1520-0493(1989)117<1641:cacpsw>2.0.co;2

Tomita, H., Tsugawa, M., Satoh, M., & Goto, K. (2001). Shallow water model on a modified icosahedral geodesic grid by using spring dynamics. *Journal of Computational Physics*, *174*(2), 579–613. https://doi.org/10.1006/jcph.2001.6897

Tompkins, A. (2005). The parametrization of cloud cover. In *ECMWF moist processes lecture note series tech* (Vol. 25). Memo.

Tompkins, A. M. (2002). A prognostic parameterization for the subgrid-scale variability of water vapor and clouds in large-scale models and its use to diagnose cloud cover. *Journal of the Atmospheric Sciences*, *59*(12), 1917–1942. https://doi.org/10.1175/1520-0469(2002)059⟨1917:APPFTS⟩2.0.CO;2

Vergara-Temprado, J., Ban, N., Panosetti, D., Schlemmer, L., & Schär, C. (2020). Climate models permit convection at much coarser resolutions than previously considered. *Journal of Climate*, *33*(5), 1915–1933. https://doi.org/10.1175/jcli-d-19-0286.1

Walcek, C. J. (1994). Cloud cover and its relationship to relative humidity during a springtime midlatitude cyclone. *Monthly Weather Review*, *122*(6), 1021–1035. https://doi.org/10.1175/1520-0493(1994)122⟨1021:CCAIRT⟩2.0.CO;2

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

Xu, K.-M., & Krueger, S. K. (1991). Evaluation of cloudiness parameterizations using a cumulus ensemble method. *Monthly Weather Review*, *119*(2), 342–367. https://doi.org/10.1175/1520-0493(1991)119⟨0342:EOCPUA⟩2.0.CO;2

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), 3295. https://doi.org/10.1038/s41467-020-17142-3

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). 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