Deep Learning Based Cloud Cover Parameterization for ICON

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

[1]Deutsches Zentrum für Luft- und Raumfahrt (DLR)
[2]University of Lausanne
[3]Columbia University
[4]Max Planck Institute for Meteorology
[5]Deutsches Zentrum für Luft- und Raumfahrt

November 22, 2022

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.

Deep Learning Based Cloud Cover Parameterization for ICON

Arthur Grundner[1,2], Tom Beucler[3], Fernando Iglesias-Suarez[1], Pierre Gentine[2], 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]Columbia University, Center for Learning the Earth with Artificial intelligence And Physics (LEAP), New York, NY 10027, USA
[3]University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland
[4]Max Planck Institute for Meteorology, Hamburg, Germany
[5]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

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

Corresponding author: Arthur Grundner, `Arthur.Grundner@dlr.de`

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 ICON 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.

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, i.e. they cannot be directly resolved but need to be "parameterized". It turns out that insufficiencies in cloud parameterizations are a major cause of the uncertainty of climate projections (e.g. Eyring et al., 2021; Randall et al., 2003; Schneider et al., 2017). This uncertainty in climate projections has not decreased in the last 40 years (Meehl et al., 2020).

These 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 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 estimate constitutes an important parameter for the radiation scheme and influences the tendencies of cloud liquid water, cloud ice, and water vapor in the microphysics' scheme (Lohmann & Roeckner, 1996; Pincus & Stevens, 2013). Cloud cover is estimated as a diagnostics 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 can only exist 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 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).

Our novel approach to a cloud cover parameterization is based on the idea of training a supervised deep learning scheme to estimate the coarse-grained cloud cover using coarse-grained high-resolution thermodynamical variables as inputs. 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, with horizontal resolutions of 2.5 km 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 km 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 neural networks (NNs) on coarse-grained data from these high-resolution simulations. Here, two commonly used ICON-A grids (with horizontal resolutions of 80 km 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, a NN-powered parameterization of cloud cover could accelerate and improve the representation of cloud-scale processes (from radiative feedbacks to precipitation statistics).

The field of machine learning based parameterizations is growing and ranges from radiation (Chevallier et al., 2000; Krasnopolsky et al., 2005), convection (Beucler, Pritchard, Gentine, & Rasp, 2020; Gentine et al., 2018; Mooers et al., 2020; Rasp et al., 2018) and microphysics (Gettelman et al., 2021; Seifert & Rasp, 2020) to nonorographic gravity waves (Chantry et al., 2021). For instance, in a pioneering study by Rasp et al. (2018), a 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. Often, the effects of multiple sub-grid scale processes are learned (Brenowitz & Bretherton, 2018, 2019; Brenowitz et al., 2020; Han et al., 2020; Krasnopolsky et al., 2013; Yuval & O'Gorman, 2020; Yuval et al., 2021). 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. In the context of these new advances, our study is the first machine learning based approach specifically focused on the parameterization of cloud cover. Some of these other studies also use coarse-grained high-resolution data as training data. 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, 2019); Brenowitz et al. (2020); Yuval and O'Gorman (2020); Yuval et al. (2021) adapted this approach. However, in contrast to our study, they worked with idealized aquaplanet simulations and coarse-graining limited to the horizontal dimension.

The first key question that we want to tackle in this study is whether we can train a NN based cloud cover parameterization that is able to emulate high-resolution cloudiness. We then want to 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 (Sec. 2) and the NNs (Sec. 3), before evaluating regionally (Sec. 4.1) and globally (Sec. 4.2) trained networks in their training regime, studying their generalization capability (Sec. 4.3) and peeking inside the black box (Sec. 4.4, 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 two 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 hours. 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 one 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 80 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 NWP 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 (Giorgetta et al., 2018). An overview of the specifically chosen parameterization schemes can be found in Table 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 (Mauritsen et al., 2007; Raschendorfer, 2001), for the radiative transfer (Barker et al., 2003; Mlawer et al., 1997; Pincus et al., 2019), and the land component (Raddatz et al., 2007; Schrodin & Heise, 2001; Schulz et al., 2015). 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 **i)** dividing its triangle edges into n parts, creating new triangles by connecting the new edge points and by **ii)** completing k subsequent edge bisections while once more connecting the new edge points after each bisection. 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).

2.2 Coarse-Graining Methodology

We can 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. 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.

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 resolutions, we additionally coarse-grain to the low-resolution R2B5 grid used in Hohenegger et al. (2020) with a resolution of ≈ 80 km. Afterwards, 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.

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 afterwards 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 **i)** compute the vertical overlap β_j *after* we horizontally coarse-grain the grid cells and **ii)** work on a terrain-following height grid which allows for vertical layers of varying heights over mountainous 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 $\alpha_{i,j}, \beta_j, \gamma_i \in [0,1]$ 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

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

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 CDO package (Schulzweida, 2019), which is able to handle fractional overlap and the irregular ICON grid to coarse-grain to and from. Figure 1 shows an example of horizontal and vertical coarse-graining of cloud cover snapshots from the QUBICC and the NARVAL data set.

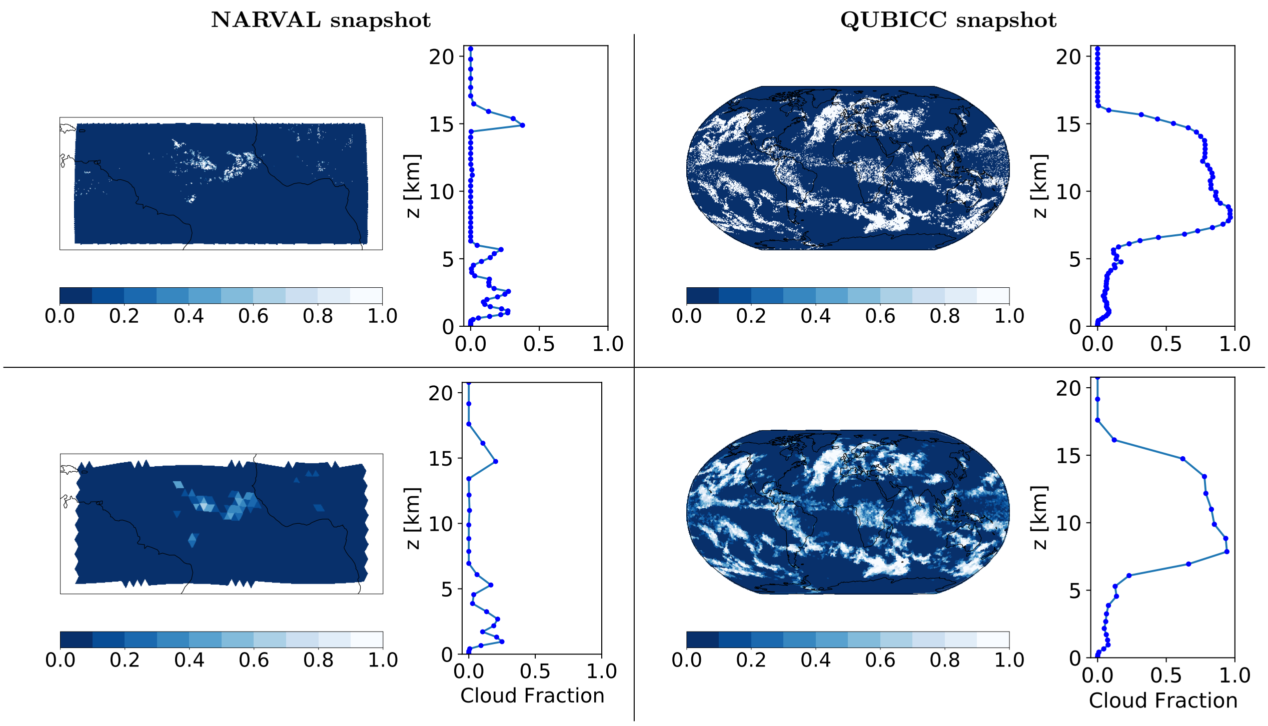


Figure 1. Illustration of coarse-graining using the example of cloud fraction. Here we show snapshots of the horizontal fields (on a single layer) and vertical profiles (from a single column) from the high-resolution NARVAL and QUBICC simulations (top row) and the corresponding coarse-grained horizontal 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.

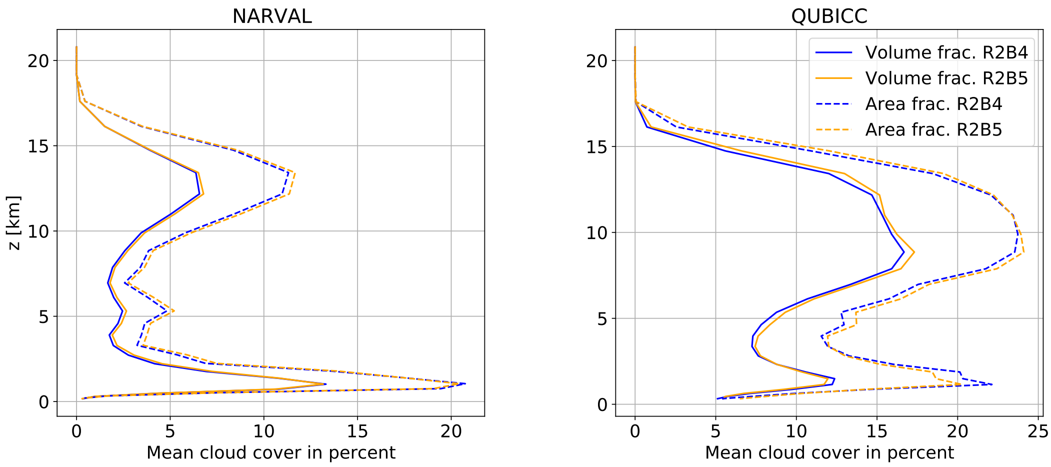


Figure 2. Comparison of the coarse-grained mean cloud volume and mean cloud area fraction profiles for NARVAL (left) and QUBICC (right). The cloud volume fraction is generally never greater than the cloud area fraction. Close to the surface, the grid cell thickness and thus also the vertical sub-grid variability of clouds is small. There it follows that the cloud area fraction is approximately equal to the cloud volume fraction.

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. While horizontally coarse-graining NARVAL data, we analogously omit low-resolution grid cells that are not located entirely inside the NARVAL region.

To derive 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 assumed maximum overlap as the level separation of vertical layers is relatively small. We 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

$$\overline{C} = \sum_{(i,j)\in H} \gamma_i \max_j\{C_{i,j}\}. \qquad (2)$$

For QUBICC grid cells, which are always either fully cloudy or cloud-free, we can directly interpret equation (2) 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 2 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.

Having introduced and coarse-grained the training data, we can now turn towards 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 relative humidity, 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, including some surface variables. Local atmospheric and dynamical conditions most likely have a significant influence on cloudiness. 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, 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). The input-output architecture of these three NN types is illustrated in Figure S2.

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 largest differences between the R2B4- and R2B5 models exist in the neighborhood-based models:

The set of predictors for the neighborhood-based R2B5 model contains data from the current grid cell and its 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 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. Although 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 hour) in its set of predictors.

Table 1. Overview of the NNs and their input features. 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 shaded in purple. More information on the choices and meaning of the features can be found in the SI.

	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			✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	

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 little 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 a NN depends on the amount and nature of the training data. For relatively few training samples it is computationally admissible to train the networks using a small batch size. This becomes computationally prohibitive when working with a large amount of data, because it would require too many iterations to process the entire data set. On our GPU, doubling the batch size halves the duration to process the data set. The amount of training data 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 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) avoid faulty input samples, d) normalize the training data, and e) combat the class imbalance of having a relatively large number of cloud-free grid cells in the training data. Steps d) and e) were also necessary for the coarse-grained NARVAL data. The more balanced ratio between cloudy and cloud-free grid cells for e) was achieved by randomly sub-sampling from the cloud-free grid cells.

To train the NARVAL R2B4 networks we 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. For the QUBICC

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: $4.7 \cdot 10^{-2}$, L2: $8.7 \cdot 10^{-2}$
Batch Normalization	None	After the second hidden layer
Optimizer	Nadam/Adam	Adam/Adadelta
↪ Initial learning rate	10^{-3}	$4.3 \cdot 10^{-4}$
↪ Batch size	32/128	1028
↪ Maximal number of epochs	70/40	$30 - 50$

R2B5 models, on the other hand, the focus is on a more universal applicability. We therefore use a temporally coherent three-fold cross-validation split (illustrated in Figure S3). 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 we found that a common architecture was optimal for the models N1-N3 and Q2 of Table 1. The training data for models Q1 and Q3 was more abundant and necessitated an increase of the batch size during optimization. This in turn required an adjustment of the architecture. 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.

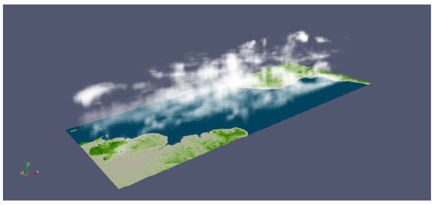
4 Results

4.1 Regional Setting (NARVAL)

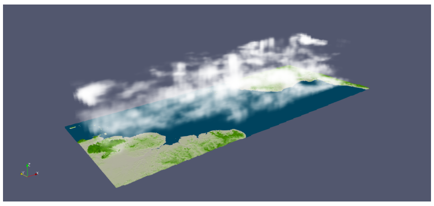
In this section we show the results of the NNs trained and evaluated on the coarse-grained and preprocessed NARVAL R2B4 data (see SI for more details on the preprocessing). For these regionally-trained NNs we define cloud cover as a cloud volume fraction.

The snapshots and Hovmoeller plots of Figure 3 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 four 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.

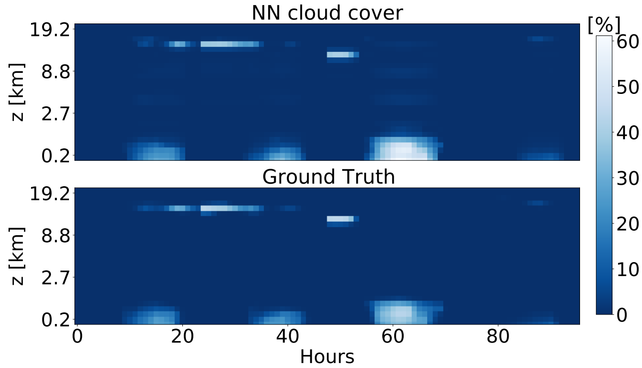
The models' mean-squared errors (MSEs) (shown in Table 3) represent the absolute average squared mismatch per grid cell in percent between the predicted and the true cloud cover. As opposed to Figure 3, the MSEs provide more statistically tangible



(a) NN cloud cover



(b) Ground Truth



(c) Hovmoeller plots

Figure 3. The column-based 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 August 1 - August 4, 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.

Table 3. MSEs (in $(\%)^2$) of NARVAL and baseline models evaluated on the coarse-grained and preprocessed NARVAL data

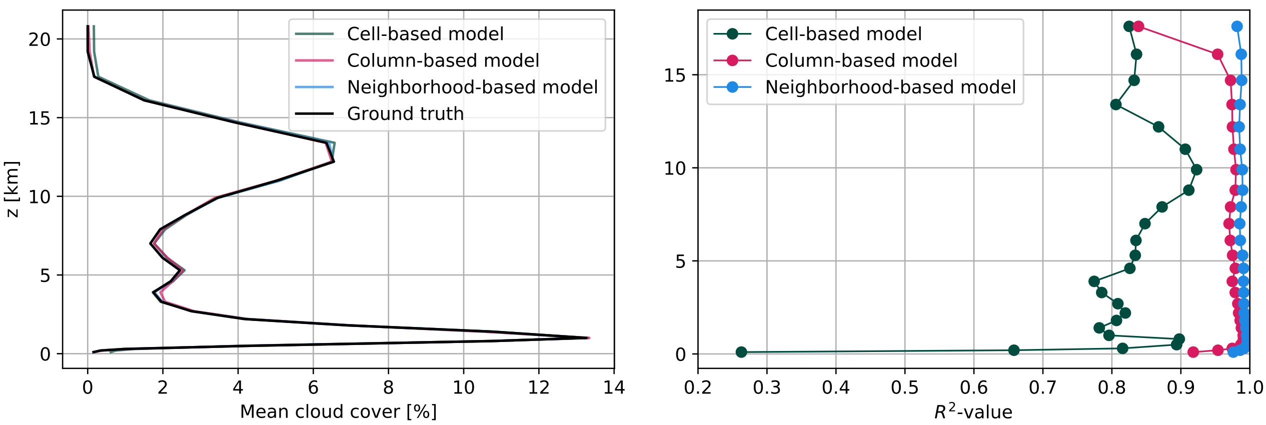
		Type		
		Cell-based	Column-based	Neighborhood-based
Our models	Training set	15.16	1.64	0.84
	Validation set	15.18	1.78	1.00
	Test set	15.19	1.78	1.01
Baseline models	Untrained NN	131.07	105.97	113.34
	Zero output model	129.62	113.91	113.37
	Constant output model	109.63	92.23	86.48
	Best linear model	81.71	18.56	4.79
	Simple Sundqvist scheme	85.19		

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 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. These baseline models are fitted to the same normalized data sets as the respective NNs. As our first baseline we evaluate an untrained NN, which is a NN with random weights and biases. Second, we fit a zero output model, which always yields 0, and 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 and can thus attain the lowest MSE of our baseline models. The simple Sundqvist scheme is a simplified version of the (mainly cell-based) ICON-A cloud cover parameterization. We simplify it by assuming a constant surface pressure of 1013.25 hPa and no adjustment for cloud cover in regions below subsidence inversions.

By isolating vertical layers we can better illustrate the distribution of actual and inferred cloud cover in the troposphere. The averaged vertical profile of cloud cover features three maxima (depicted in Figure 4a). These can be attributed to the three modes of tropical convection (shallow, congestus, and deep). The model-based cloud cover profiles closely align with the actual cloud cover profile. In contrast to Müller (2019), we find a clear peak for deep convective clouds in the coarse-grained NARVAL (and particularly also in the NARVALII) data. However, the author defined grid cells to be cloudy whenever the total cloud condensate mass mixing ratio exceeded 0.1g/kg and not based on the cloud cover model output field.



(a) Cloud cover profiles

(b) Coefficients of determination (best value: 1)

Figure 4. 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 km and 12.2 km. The maximal absolute discrepancy between the averaged NN 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.

In Figure 4b 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

$$R_l^2 = 1 - \frac{mse_l}{var_l}. \tag{3}$$

For a given vertical layer l, mse_l is the mean-squared error between a given model's prediction and the true cloud cover and var_l the variance of cloud cover. Clearly, i) $R_l^2 \leq 1$, ii) $R_l^2 = 1$ implies $mse_l = 0$, and iii) 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 $mse_l \leq 0.1 \cdot var_l$. 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, i.e. 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.

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. MSEs (in $(\%)^2$) of the models trained with a 3-fold cross validation split on the coarse-grained and preprocessed QUBICC data. For each type we highlight the chosen model in bold. 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.

	Cloud volume fraction		**Cloud area fraction**	
	Training loss	Validation loss	Training loss	Validation loss
Cell-based				
Split 1	33.40 (29.72)	33.79 (30.11)	87.58 (81.59)	88.38 (82.37)
Split 2	33.22 (29.43)	32.77 (28.98)	88.14 (81.21)	87.98 (80.96)
Split 3	39.60 (36.15)	40.94 (37.48)	88.83 (82.66)	90.03 (83.87)
Column-based				
Split 1	8.01 (7.84)	8.13 (7.98)	19.85 (19.60)	21.36 (21.05)
Split 2	7.89 (7.72)	8.14 (8.03)	20.31 (20.01)	20.07 (19.79)
Split 3	7.95 (7.83)	9.51 (8.80)	20.27 (19.91)	96.44 (20.58)
Neighborhood-based				
Split 1	26.27 (22.40)	25.43 (21.56)	53.96 (46.88)	55.51 (48.44)
Split 2	28.39 (24.13)	27.28 (23.04)	54.49 (47.71)	53.12 (46.28)
Split 3	24.73 (20.12)	25.07 (20.46)	51.77 (46.18)	52.19 (46.61)

Table 4 gives an overview of the performance of all model types trained and evaluated on each of the data splits. 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

Table 5. MSEs (in $(\%)^2$) of cloud volume fraction baseline models trained and evaluated on coarse-grained and preprocessed QUBICC data

	Type		
	Cell-based	Column-based	Neighborhood-based
Untrained NN	913.91	471.17	699.21
Zero output model	923.94	537.24	692.95
Constant output model	684.51	431.28	558.28
Best linear model	401.47	97.81	297.63
Simple Sundqvist model	773.56		

Due to computational reasons, only 0.001% *of the data (i.e.* $\approx 10^4$ *samples) was used to compute the MSE of the simple Sundqvist model.*

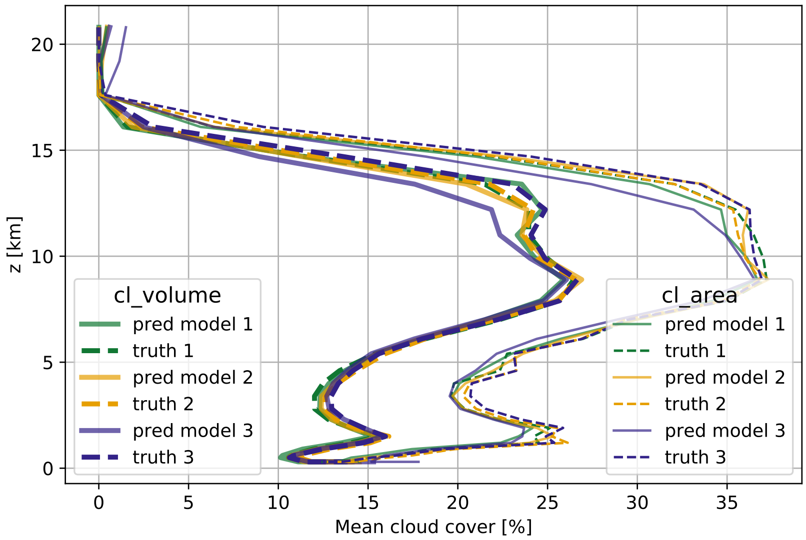
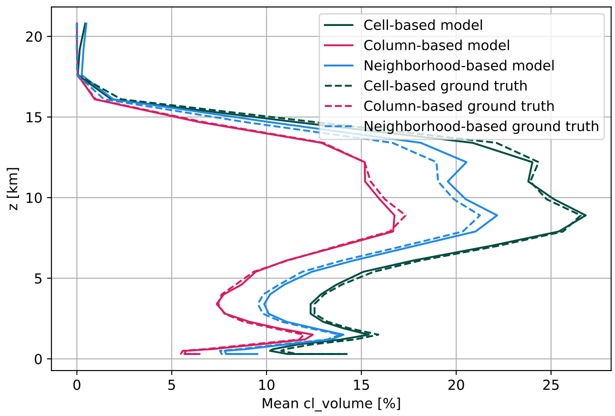


Figure 5. The cell-based cloud volume and cloud area fraction models of the 3-fold cross-validation split evaluated on their respective validation sets.

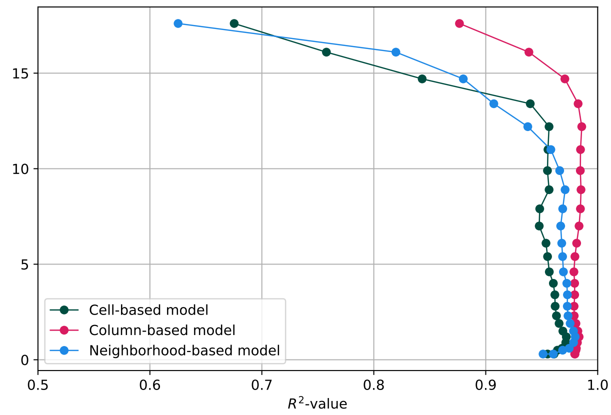
the higher stochasticity present in the higher resolution R2B5 data. Both of these reasons allow for a larger range of outputs for similar inputs, inevitably increasing the MSE of our deterministic model. Nevertheless, we are still well below the MSEs given by our baseline models in Table 5.

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.

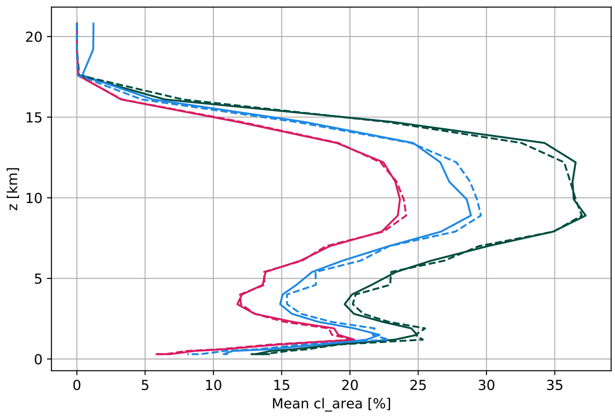
In Table 4 we also include bounded losses in parentheses. That means that the NN's cloud cover predictions, which are smaller than 0% are set to be 0%, before its MSE is computed. And likewise, predictions greater than 100% are set to be 100%. The difference between these two types of losses is relatively small. We can deduce that the NNs (with the surprising exception of the column-based NN for cloud area fraction from the third split) usually stay within the desired range of $[0, 100]\%$ without being forced to do



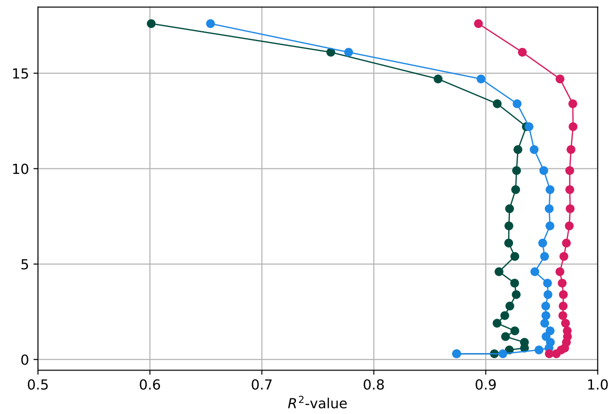
(a) Cloud volume fraction profiles



(b) Cloud volume fraction R^2-values



(c) Cloud area fraction profiles



(d) Cloud area fraction R^2-values

Figure 6. Evaluation of QUBICC cloud volume and cloud area models on coarse-grained and preprocessed QUBICC R2B5 data. The average 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 SI for more details). The column-based ground truth profile represents the true QUBICC cloud cover profiles since its data was not altered by preprocessing.

so. In bold we highlight the splits on which we trained the models that produce the lowest error on the entire data set. The corresponding models are used in all subsequent figures.

In Figure 5 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. This makes the validation loss already indicative of the performance of the models on data outside of their immediate training distribution. 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. This suggests that the first two splits provide more stratified and thus more suitable sets of training data.

Despite the challenging setting, Figures 6a and 6c 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 4, 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, i.e. 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.

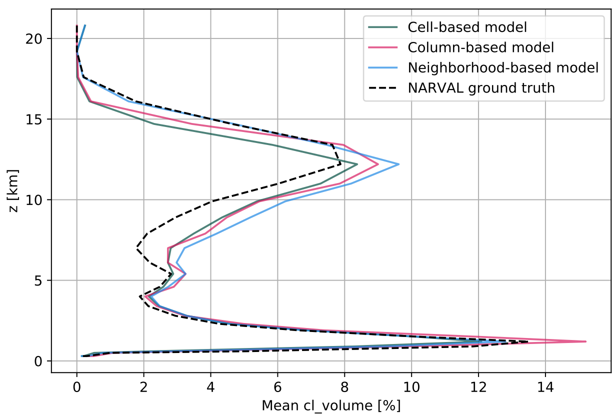
From global to regional

We first study the capability of QUBICC-trained models to generalize to the NARVAL data (see Figure 7). 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 ≈ 0.2) when compared to the models' performance on its training data (Figure 6). 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 succesfully to the NARVAL data. A sign of overfitting the training data is discernible: While the column-based model had emulated the training data better than the other two model types, it generalizes slightly worse to the NARVAL data (see e.g. Figure 7c).

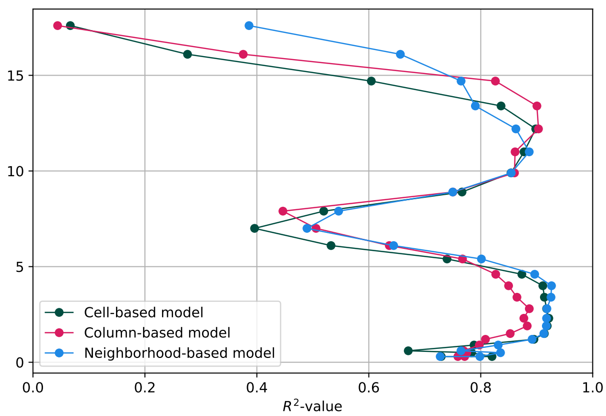
A considerable bias that pertains all three NN types is a consistent overprediction of both cloud volume and cloud area fraction between 6 and 9 km. In this altitude range, this is visible in all four plots, either through the mismatch in mean cloud cover or the dip in R^2-value. This striking behavior will be further investigated in section 4.5.

From regional to global

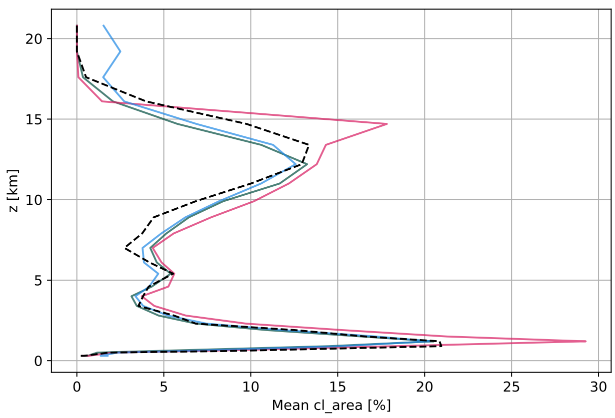
We have seen that the NNs are able to reproduce the cloud cover distribution of the storm-resolving NARVAL simulation, limited to its tropical region. We coarse-grain the QUBICC data to the same R2B4 grid resolution that the NARVAL NNs were trained with. This helps us to investigate to what extent the NNs can actually generalize to out-of-training regimes. We focus on the tropics first, extending the evaluation from the NARVAL region (68W-15E, 10S-20N) to the entire tropical band (23.4S-23.4N). 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 repre-



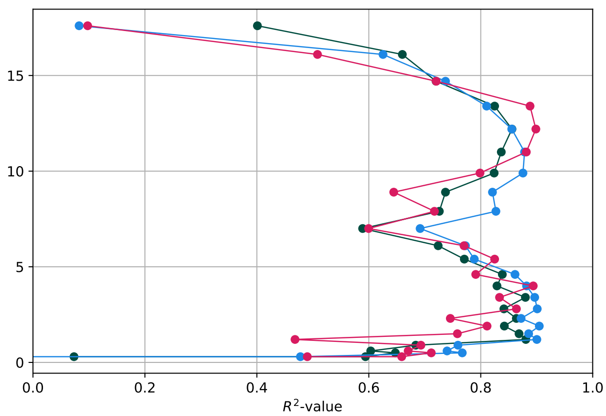
(a) Cloud volume fraction profiles



(b) Cloud volume fraction R^2-values



(c) Cloud area fraction profiles

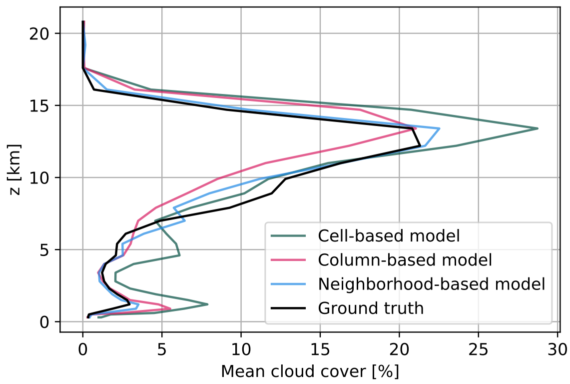


(d) Cloud area fraction R^2-values

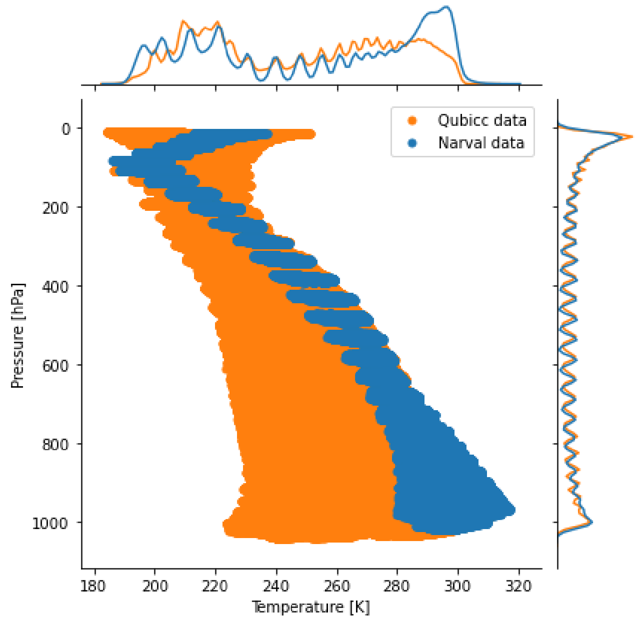
Figure 7. Evaluation of QUBICC R2B5 cloud volume and cloud area models on NARVAL R2B5 data. The average R^2-values of the cell-, column-, and neighborhood-based models shown in b) are (0.74, 0.74, 0.79) and in d) are (0.72, 0.71, 0.72).

sented 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, 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 8b). 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 1000 hPa in the tropical training regime of the NARVAL data. This circumstance is particularly challenging for the neighborhood- and column-based



(a) Cloud cover profiles: Tropics



(b) Joint and marginal distributions of temperature and pressure

Figure 8. Panel a): Evaluation of NARVAL R2B4 models (NARVAL region: 68W-15E, 10S-20N) on QUBICC R2B4 data over the tropical zone (23.4S - 23.4N). We plot the means over 10 days (Nov. 20 - Nov. 29, 2004). Different NNs of the same type produce consistent mean vertical cloudiness profiles ($\pm 1\%$). 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.

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 1000 hPa.

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, i.e. 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', i.e. 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').

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 (Tables 3, 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 (containing 10000 samples) of the respective training data sets. 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 7). 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) despite the doubling of the horizontal resolution.

The subset of NARVAL R2B5 data 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 9) 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 originally believed the cause for this to be that the QUBICC training data has a very distinct average cloudiness profile than the NARVAL data that we apply the model to. After all, the Shapley values have to bridge the gap between the base values and the new model predictions. However, constructing the base values to be much closer to the average NARVAL cloudiness profile does not decrease the magnitude of the Shapley values of the QUBICC model (see Figure S6). We also find that such a drastic change of the base value barely impacts the qualitative information that we can extract from the plots. An alternative explanation goes as follows: During training, the QUBICC model was confronted with a large variety of climatic conditions across the entire globe implying a larger variance of cloud cover. The NN is thus used to deviate from

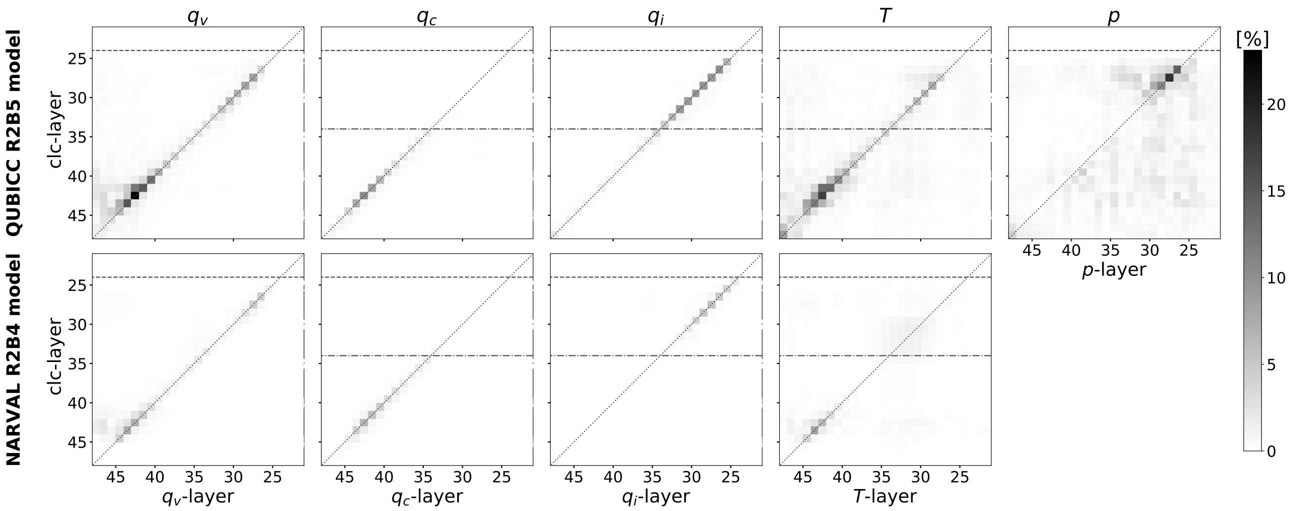


Figure 9. Average absolute SHAP values of the QUBICC R2B5 and the NARVAL R2B4 column-based models when applied to a sufficiently large subset of the NARVAL R2B5 data. We use the conventional ICON-A numbering of vertical layers from layer 21 (at a height of $\approx 20.8\,\text{km}$) decreasing in height to layer 47, which coincides with Earth's surface. The dashed line shows the tropopause, here at $\approx 15\,\text{km}$, the dash dotted line shows the freezing level (i.e. where temperatures are on average below 0 degrees celsius), here at $\approx 5\,\text{km}$. 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 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 set 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. Due to the Clausius-Clapeyron relation, relative humidity depends most strongly on temperature. Taking into account that throughout the troposphere relative humidity 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 7).

4.5 Understanding Model Errors

In this section, our goal is to understand the source of flawed NN predictions. Which input features are most responsible for erroneous predictions in one NN, while in another NN they cause no bias?

In the evaluation of the QUBICC (R2B5) cloud volume fraction models on NARVAL R2B5 data (Figure 7) 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, which corresponds to this altitude, is the 32^{nd} 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), 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 A.

Figure 10a 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 10b and c we find that the local values have the largest effect on cloud cover. This local importance is also corroborated by Figure 9. 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 10d and e). 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 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 storm-resolving model 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 4, 6). 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 7). Using SHAP we compare regionally and globally trained

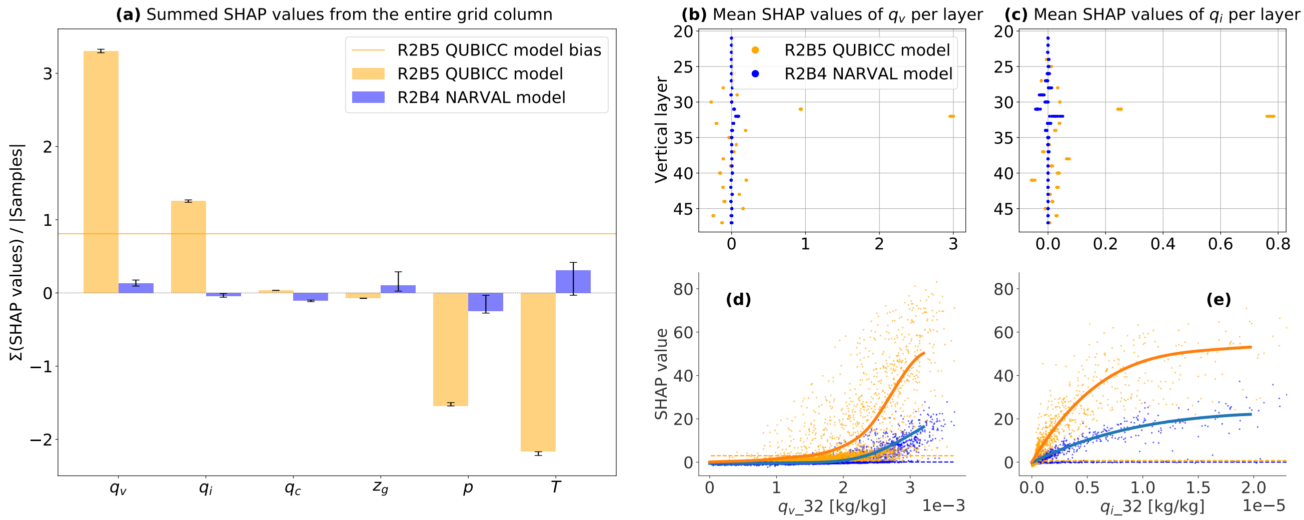


Figure 10. SHAP/Shapley value statistics per input feature for cloud cover predictions on vertical layer 32 (at ≈ 7 km) of the column-based models with a focus on q_v and q_i in **(b)**-**(e)**. Input features the models have not in common are neglected. As in Figure 9, the Shapley values are computed on a set of 10^4 random NARVAL R2B5 samples (using ten 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 QUBICC R2B5 model bias (of 0.95%) on layer 32 (cf. Figure 7a) 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 ten 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)**.

NNs to understand the relationship between predicted cloud cover and its thermodynamic environment and vertical structure (Figure 9). We are able to uncover that specific humidity and cloud ice are the drivers of one NN's largest tropospheric generalization error (Figure 10).

We implement three different types of NNs in order to assess the degree of (vertical) locality 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 4, 6). However, the more local cell- and neighborhood-based NNs show slightly fewer signs of overfitting the training data (Figure 7). Generally we find that none of 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 9). 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 2).

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. As we are not planning to further train the NNs in this coupled mode, the implementation itself should be relatively straightforward (using e.g. the Fortran-Keras bridge from Ott et al. (2020)).

The presence of condensate-free clouds in the training data (they make up $\approx 7\%$ of all cloudy grid cells) shows inaccuracies that are present both in the NARVAL and the QUBICC training data. The most likely reason for their occurrence is a temporal mismatch between different model output variables from one common time step. Some parameterization schemes in the ICON model are processed sequentially, potentially causing such a temporal mismatch. However, this delay should not exceed the fast physics time step in the model, which was set to 40 seconds in the QUBICC and to five minutes in the NARVAL simulations.

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 climate so different that it circumvents our regularization measures (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, Pritchard, Peng, et al., 2020).

While we can achieve good results with our "vanilla" NNs, Bayesian NNs or adding dropout to our conventional NNs are promising ways of also estimating the uncertainty associated with NN predictions (Gal & Ghahramani, 2016). Furthermore, we have developed different types of NNs to test which information those NNs need to learn cloud cover. However, causal discovery methods would likely provide a more rigorous and physically consistent approach for defining the input features (Nowack et al., 2020; Runge et al., 2019).

From a climate science perspective, instead of diagnosing cloud cover from large-scale variables directly, one could also train a 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 Comparing Two Neural Networks With Shap

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)

$$\sum_{i \in I} \phi_{h,X,i} = h(X) - \mathbb{E}[h(X)], \tag{A1}$$

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 (A1) we get

$$\sum_{i \in I} \phi_{f,X,i} - \sum_{j \in J} \phi_{g,X,j} = f(X) - g(X) + B_f - B_g = f(X) - g(X), \tag{A2}$$

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 **i)** $\overline{f}$ and $\overline{g}$ are good approximations of the predicted averages of f and g on the entire NARVAL R2B5 data set (as shown in Figures 7a and S5a) and **ii)** 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 (A2) 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

$$\begin{aligned} \overline{f} - \overline{g} &= \sum_{i \in I \cap J} (\overline{\phi_{f,X,i}} - \overline{\phi_{g,X,i}}) + \sum_{i \in I \setminus J} \overline{\phi_{f,X,i}} - \sum_{i \in J \setminus I} \overline{\phi_{g,X,i}} \\ &\approx \sum_{i \in I \cap J} (\overline{\phi_{f,X,i}} - \overline{\phi_{g,X,i}}). \end{aligned} \tag{A3}$$

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 **i)** and **ii)**, we can generalize (A3) to the entire NARVAL R2B5 data set.

Acknowledgments

The neural network and analysis code can be found at `https://github.com/agrundner24/iconml_clc` and is preserved at `DOI: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`.

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). 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). We kindly acknowledge the computational resources of the Deutsches Klimarechenzentrum (DKRZ, Hamburg, Germany) that allowed the training and evaluation of the machine learning algorithms used in this study.

References

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

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

Barker, H. W., Stephens, G., Partain, P., Bergman, J., Bonnel, B., Campana, K., ... Yang, F. (2003). Assessing 1d atmospheric solar radiative transfer models: Interpretation and handling of unresolved clouds. *Journal of Climate*, *16*(16), 2676–2699.

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.* doi: 10.1109/IGARSS39084.2020.9324569

Beucler, T., Pritchard, M. S., Peng, L., Rasp, S., Gentine, P., & Gupta, A. (2020). Climate-invariant nets: Using physical rescalings to help neural networks generalize to out-of-sample climates. In *Agu fall meeting 2020.*

Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parameterizations of convection. *Journal of the Atmospheric Sciences*. doi: 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. doi: 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. doi: 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*. doi: 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. doi: 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*. doi: https://doi.org/10.1002/qj.49712656318

Doms, G., Förstner, J., Heise, E., Herzog, H., Mironov, D., Raschendorfer, M., ... others (2011). A description of the nonhydrostatic regional cosmo model, part ii: Physical parameterization. *Deutscher Wetterdienst, Offenbach, Germany*.

Eyring, V., Mishra, V., Griffith, G. P., Chen, L., Keenan, T., Turetsky, M. R., ... van der Linden, S. (2021). Reflections and projections on a decade of climate science. *Nature Climate Change*, *11*(4), 279-285. doi: 10.1038/s41558-021-01020-x

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. doi: 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). doi: 10.1029/2020ms002268

Giorgetta, M. A., Crueger, T., Brokopf, R., Esch, M., Fiedler, S., Hohenegger, C., ... Stevens, B. (2018). Icon-a, the atmosphere component of the icon earth system model: I. model description. *Journal of Advances in Modeling Earth Systems*, *10*(7), 1638-1662. doi: 10.1029/2017ms001233

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). doi: 10.1029/2020ms002076

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

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, *4*(2), 251–257.

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. doi: 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. doi: 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*. doi: 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*. doi: 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., ... Roeckner, E. (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. doi: 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 Atmospheric Sciences*, *64*(11), 4113–4126.

Meehl, G., Senior, C., Eyring, V., Flato, G., Lamarque, J., Stouffer, R., ... Schlund, M. (2020). Context for interpreting equilibrium climate sensitivity and transient climate response from the cmip6 earth system models. *Science Advances*. doi: 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: Atmospheres*, *102*(D14), 16663–16682.

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

Mooers, G., Tuyls, J., Mandt, S., Pritchard, M., & Beucler, T. (2020). Generative modeling of atmospheric convection. In *Proceedings of the 10th interna-*

tional conference on climate informatics. doi: 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). doi: 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. doi: 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. *Scientific Programming*, *2020*, 1-13. doi: 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. doi: 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. doi: 10.1002/jame.20027

Prill, F., Reinert, D., Rieger, D., Zängl, G., Schröter, J., Förstner, J., ... Vogel, B. (2019, April). Icon tutorial: Nwp mode and icon-art [Computer software manual].

Raddatz, T., Reick, C., Knorr, W., Kattge, J., Roeckner, E., Schnur, R., ... Jungclaus, J. (2007). Will the tropical land biosphere dominate the climate–carbon cycle feedback during the twenty-first century? *Climate dynamics*, *29*(6), 565–574.

Randall, D., Khairoutdinov, M., Arakawa, A., & Grabowski, W. (2003). Breaking the cloud parameterization deadlock. *Bulletin of the American Meteorological Society*, *84*(11), 1547-1564. doi: 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. *PNAS*, *115*(39), 9684-9689. doi: 10.1073/pnas.1810286115

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

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. doi: 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, October). *Cdo user guide.* doi: 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). doi: 10.1029/2020ms002301

Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. In *International conference on machine learning* (p. 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., ... Zängl, G. (2020). The added value of large-eddy and storm-resolving models for simulating clouds and precipitation. *Journal of the Meteorological Society of Japan*, *98*(2), 395-435. doi: 10.2151/jmsj.2020-021

Stevens, B., Ament, F., Bony, S., Crewell, S., Ewald, F., Gross, S., ... Zinner, T. (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. doi: 10.1175/bams-d-18-0198.1

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., ... Zhou, L. (2019). Dyamond: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Progress in Earth and Planetary Science*, *6*(1). doi: 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*.

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. doi: 10.1006/jcph.2001.6897

Tompkins, A. (2005). The parametrization of cloud cover. *ECMWF Moist Processes Lecture Note Series Tech. Memo*, *25*.

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*. doi: 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. doi: 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*. doi: https://doi.org/10.1175/1520-0493(1994)122⟨1021:CCAIRT⟩2.0.CO;2

Xu, K.-M., & Krueger, S. K. (1991). Evaluation of cloudiness parameterizations using a cumulus ensemble method. *Monthly Weather Review*. doi: 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. doi: 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). doi: 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. doi: 10.1002/qj.2378

Supporting Information for "Deep learning based cloud cover parameterization for ICON"

Arthur Grundner[1,2], Tom Beucler[3], Fernando Iglesias-Suarez[1], Pierre Gentine[2], 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]Columbia University, Center for Learning the Earth with Artificial intelligence And Physics (LEAP), New York, NY 10027, USA

[3]University of Lausanne, Institute of Earth Surface Dynamics, Lausanne, Switzerland

[4]Max Planck Institute for Meteorology, Hamburg, Germany

[5]University of Bremen, Institute of Environmental Physics (IUP), Bremen, Germany

Contents

Introduction This supplementary information provides more detailed information concerning the data and the neural networks (NNs). It describes the variables that were used as input features for the NNs, illustrates the architecture of the three NN types, and the preprocessing and amount of (training) data for each network. Table S1 specifies the parameterization schemes used in the NARVAL and QUBICC simulations. The cross-validation split for the QUBICC (R2B5) models is depicted in Figure S3. Figure

S1 illustrates the coefficients of a multiple linear model trained on the NARVAL (R2B4) data. Figures S4 and S5 cover aspects of the generalization capability of the NARVAL networks across regions and resolutions. Lastly, Figure S6 shows that SHAP values do not strongly depend on the base value.

---

1. Definition and Choice of Input Parameters for the NNs

1. **land**: The land fraction (in $[0, 1]$) is used in the ICON-A cloud cover scheme to discern whether one might have to artificially increase relative humidity in order to take thin maritime stratocumuli into account.

2. **lake**: The lake fraction (in $[0, 1]$) is a parameter closely related to the land fraction. A supply of moisture from the ground very likely influences the distribution of moisture in the atmospheric column above, especially in the presence of convection.

3. **Cor.**: The Coriolis parameter (in $1/s$) allows the cloud cover parameterization to vary between different latitudes, which can be especially useful with global training data.

4. $\mathbf{q_v}$, $\mathbf{T}$, $\mathbf{p}$, $\mathbf{z_g}$: Specific humidity (in kg/kg), air temperature (in K), pressure (in Pa) and geometric height at full levels (in m). These are the most important input variables for the original ICON-A cloud cover scheme (to compute relative humidity).

5. $\mathbf{q_c}$, $\mathbf{q_i}$: The specific cloud water content and the specific cloud ice content (in kg/kg). They have a direct influence on cloudiness as their presence is a necessary requirement for the presence of clouds. In this spirit, they are for instance used in an alternative 0-1 cloud cover scheme in ICON-A, which sets cloud cover to 1 when a certain threshold of cloud condensate is crossed.

6. $\boldsymbol{\rho}$: Air density (in kg/m^3). We left it out for the R2B5 NNs, since air density can mostly be derived from p, T and q_v by using the ideal gas law and is therefore redundant.

7. **u**, **v**: Zonal/eastward wind and meridional/northward wind (in m/s). Vertical wind shear can induce a large difference between cloud area fraction and cloud cover.

8. $\mathbf{clc_{t-1}}$: The cloud cover estimate (in $[0, 100]\%$) from the previous timestep (1 hour before). Undeniably, clouds have a memory effect on this time scale. However, a model that relies on previous cloudiness cannot be used in the first time step.

The features ρ, u, v are also used in the Tompkins scheme of cloud cover (Tompkins, 2002).

2. Preprocessing

The preprocessing, which we define as distinct from coarse-graining, consists of up to four steps:

1. **For all cell-based and QUBICC neighborhood-based models (N1, Q1 and Q3)**: Ensure that the amount of data samples with $clc \neq 0$ is as large (for the Q1 model twice as large to reduce the data size) as the one with $clc = 0$, by downsampling the latter class of cloud-free data samples.

2. **For the neighborhood-based NARVAL models (N3)**: Remove the cloud cover from the first time step of each day of the NARVAL data from the output. We cannot predict it, because there is no previous cloud cover value which the neighborhood-based NARVAL model would require as input.

3. **QUBICC data**: Remove the first time steps of the simulations because that output incorrectly consists of an entirely cloud-free atmosphere. Scale the cloud cover to be in $[0, 100]\%$. Convert the data from float64 to float32 to reduce the data size.

4. **For the QUBICC cell- and neighborhood-based models (Q1 and Q3)**: Subsample only every third hour from the QUBICC data set to reduce the data size. Assuming a high temporal correlation, we should not lose a lot of information. Remove condensate-free clouds ($\sim 7\%$ of all clouds).

5. **For all models (N1-N3, Q1-Q3)**: Normalize the actual training data so that each input feature to the NN is distributed according to a Gaussian with zero mean and unit variance. In the column-based models this means that the normalization is done on a level-by-level basis and for the cell-based and neighborhood-based models we have one level-independent mean and standard deviation per input feature. According to Brenowitz and Bretherton (2019), we expect the impact on our results due to these different choices of normalization to be very small. This step of normalization can only be done after splitting the set of all training data samples into subsets of training, validation and test data.

References

Barker, H. W., Stephens, G., Partain, P., Bergman, J., Bonnel, B., Campana, K., ... Yang, F. (2003). Assessing 1d atmospheric solar radiative transfer models: Interpretation and handling of unresolved clouds. *Journal of Climate*, *16*(16), 2676–2699.

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. doi: 10.1029/2019ms001711

Doms, G., Förstner, J., Heise, E., Herzog, H., Mironov, D., Raschendorfer, M., ... others (2011). A description of the nonhydrostatic regional cosmo model, part ii: Physical parameterization. *Deutscher Wetterdienst, Offenbach, Germany.*

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 Atmospheric Sciences*, *64*(11), 4113–4126.

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: Atmospheres*, *102*(D14), 16663–16682.

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. doi: 10.1029/2019MS001621

Raddatz, T., Reick, C., Knorr, W., Kattge, J., Roeckner, E., Schnur, R., ... Jungclaus, J. (2007). Will the tropical land biosphere dominate the climate–carbon cycle feedback during the twenty-first century? *Climate dynamics*, *29*(6), 565–574.

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

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).

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

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*. doi: https://doi.org/10.1175/1520-0469(2002)059⟨1917:APPFTS⟩2.0.CO;2

Table S1. Parameterizations used in the NARVAL and QUBICC simulations

	NARVAL	**QUBICC**
Cloud Cover	Diagnostic PDF	All-or-nothing scheme based on cloud condensate
Microphysics	Single-moment scheme (Doms et al., 2011; Seifert, 2008)	Single-moment scheme (Doms et al., 2011; Seifert, 2008)
Radiation	RRTM scheme (Barker et al., 2003; Mlawer et al., 1997)	RTE+RRTMGP scheme (Pincus et al., 2019)
Turbulence	Prognostic TKE (Raschendorfer, 2001)	Total turbulent energy scheme (Mauritsen et al., 2007)
Land	Tiled TERRA (Schrodin & Heise, 2001; Schulz et al., 2015)	JSBach4-lite (Raddatz et al., 2007)

Table S2. Amount of training data samples for the NNs. The tuples denote either (time steps, vertical layers, horizontal fields) or (time steps, horizontal fields). Note that for the R2B4 neighborhood-based model we trained one NN per vertical layer, so the number of training samples is equal to the number of training samples for the R2B4 column-based model. Grid columns containing grid cells that were omitted during coarse-graining are excluded in the 'After coarse-graining'-column and are also not used for training.

	Original data ($\leq 21\,\text{km}$)	After coarse-graining	After preprocessing
Cell-based			
R2B4 NARVAL	$5.6 \cdot 10^{11}$ $(1721, 66, 4887488)$	$4.5 \cdot 10^{7}$ $(1635, 27, 1024)$	$3.7 \cdot 10^{7}$
R2B5 QUBICC	$3.9 \cdot 10^{12}$ $(2162, 87, 20971520)$	$4.6 \cdot 10^{9}$ $(2162, 27, 78069)$	$8.8 \cdot 10^{8}$
Neighborhood-based			
R2B4 NARVAL	$8.4 \cdot 10^{9}$ $(1721, 4887488)$	$1.7 \cdot 10^{6}$ $(1632, 1024)$	$1.7 \cdot 10^{6}$
R2B5 QUBICC	$3.9 \cdot 10^{12}$ $(2162, 87, 20971520)$	$4.6 \cdot 10^{9}$ $(2162, 27, 78069)$	$1.2 \cdot 10^{9}$
Column-based			
R2B4 NARVAL	$8.4 \cdot 10^{9}$ $(1721, 4887488)$	$1.7 \cdot 10^{6}$ $(1635, 1024)$	$1.7 \cdot 10^{6}$
R2B5 QUBICC	$4.5 \cdot 10^{10}$ $(2162, 20971520)$	$1.7 \cdot 10^{8}$ $(2162, 78069)$	$1.7 \cdot 10^{8}$

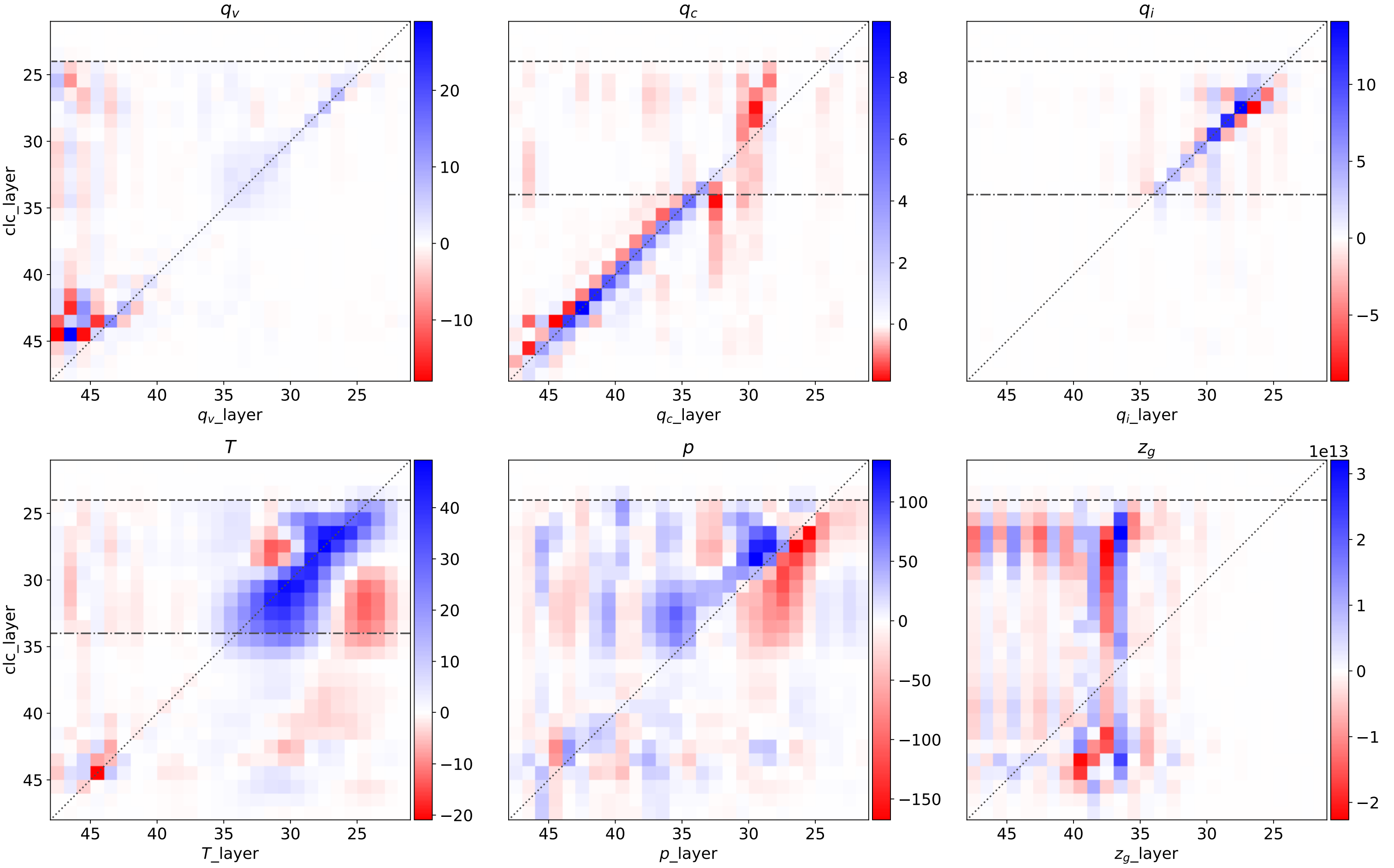


Figure S1. Coefficients of the best multiple linear model on standardized NARVAL R2B4 data. The dashed line shows the tropopause (≈ 15 km), the dash dotted line shows the freezing level (i.e. where temperatures are on average below 0 degrees) (≈ 5 km) and the dotted line visualizes the diagonal. The coefficients suggest that the problem of diagnosing cloud cover is non-local. The zg coefficients seem to dominate. An elevated grid cell on level 15 increases cloud cover significantly. However, due to the nature of the vertical grid, the layers below will also be elevated, driving a decrease of cloud cover. An increase in specific humidity, cloud water (at altitudes below the freezing level) and cloud ice (at altitudes above the freezing level) increase cloudiness in the same grid cell. In the upper troposphere, when we increase the pressure, we force the condensation of water vapor at the given level and above.

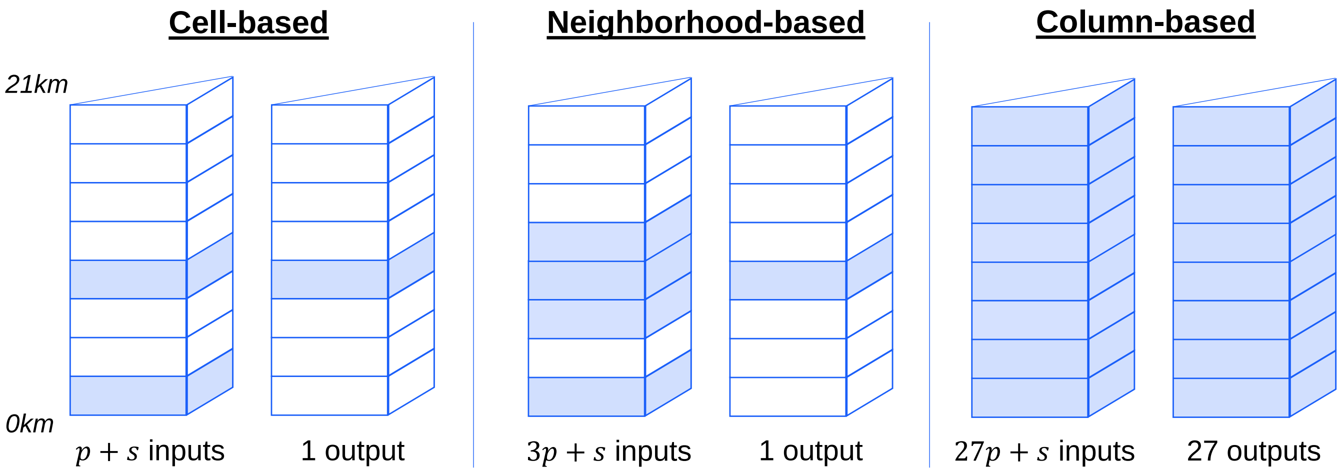


Figure S2. A sketch of the three NN types based on one grid column. The variable p denotes the number of input features from the grid cells and s is the number of extra variables from the surface. In this sketch, the neighborhood-based model uses two neighboring cells, which is only true for our QUBICC-trained NN.

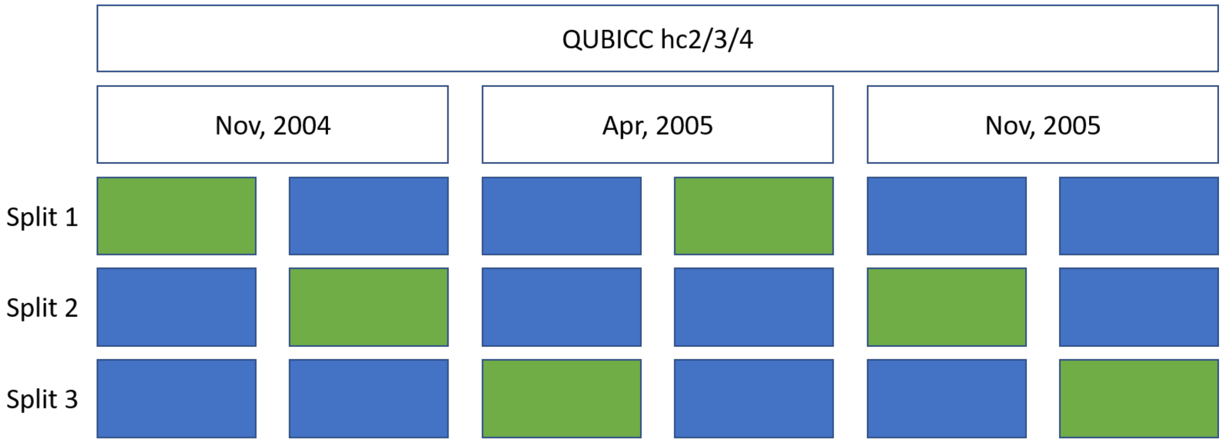


Figure S3. We split the R2B5 data using a three-fold temporally coherent cross-validation split. In each split, we train a network on the blue folds and validate it on the green folds. One fold covers approximately 15 days.

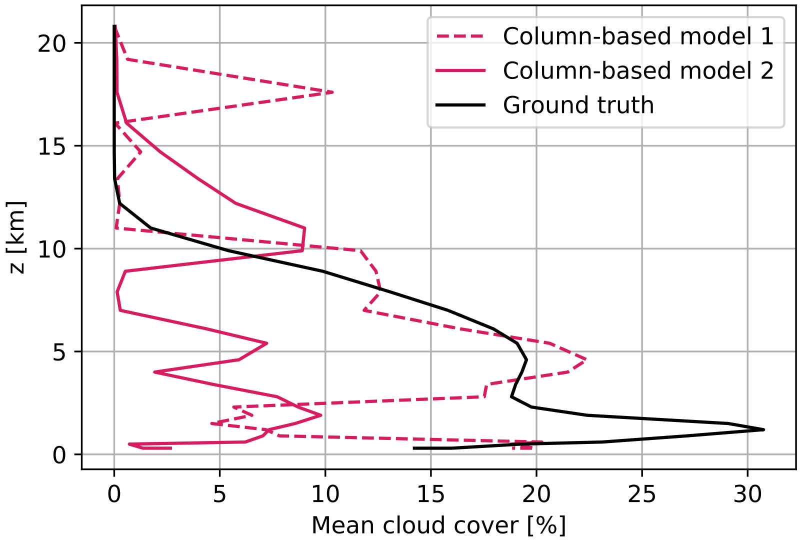
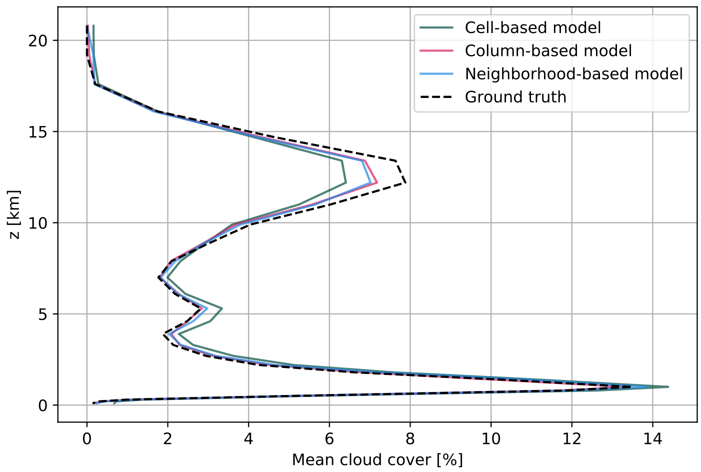
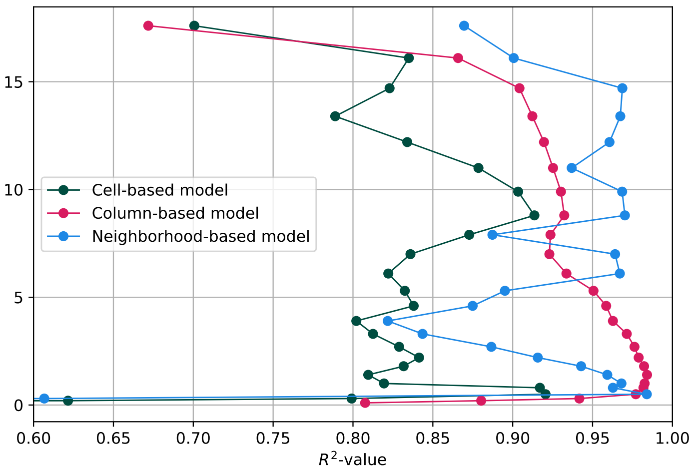


Figure S4. Two different column-based models trained on NARVAL R2B4 data evaluated on QUBICC R2B4 data over the Southern Ocean and Antarctica (< 60S). Models from the same type stop being consistent and deviate significantly from the ground truth.



(a) Cloud cover profiles



(b) Coefficients of determination (best value: 1)

Figure S5. The NNs trained on NARVAL R2B4 data evaluated on the coarse-grained and preprocessed NARVAL R2B5 data.

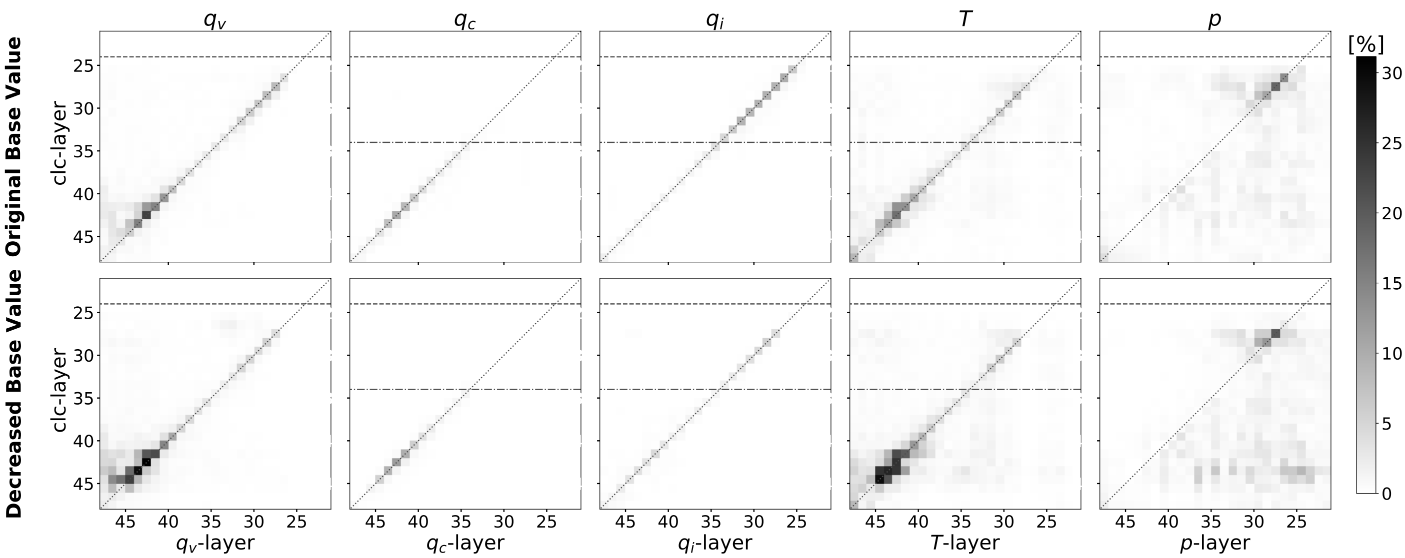


Figure S6. Average absolute SHAP values of the QUBICC R2B5 column-based model when applied to a sufficiently large subset of the NARVAL R2B5 data. By repeatedly drawing an appropriate training sample from the QUBICC training data we decrease its base values, aligning them closely with the cloud cover profile of the NARVAL R2B5 data. Tests with ten different seeds have shown the values from the lower row to be robust, with pixel values not differing absolutely by more than 1 or relatively by more than 20%. The input features that are not shown exhibit smaller absolute SHAP values ($z_g < 0.8\%$, $land/lake < 0.22\%$) everywhere and are thus omitted.