Check for updates

JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS002984

Special Section:
Machine learning application to Earth system modeling

Non-Local Parameterization of Atmospheric Subgrid Processes With Neural Networks

Peidong Wang[1], **Janni Yuval**[1], **and Paul A. O'Gorman**[1]

[1]Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

Abstract Subgrid processes in global climate models are represented by parameterizations which are a major source of uncertainties in simulations of climate. In recent years, it has been suggested that machine-learning (ML) parameterizations based on high-resolution model output data could be superior to traditional parameterizations. Currently, both traditional and ML parameterizations of subgrid processes in the atmosphere are based on a single-column approach, which only use information from single atmospheric columns. However, single-column parameterizations might not be ideal since certain atmospheric phenomena, such as organized convective systems, can cross multiple grid boxes and involve slantwise circulations that are not purely vertical. Here we train neural networks (NNs) using non-local inputs spanning over 3×3 columns of inputs. We find that including the non-local inputs improves the offline prediction of a range of subgrid processes. The improvement is especially notable for subgrid momentum transport and for atmospheric conditions associated with mid-latitude fronts and convective instability. Using an interpretability method, we find that the NN improvements partly rely on using the horizontal wind divergence, and we further show that including the divergence or vertical velocity as a separate input substantially improves offline performance. However, non-local winds continue to be useful inputs for parameterizating subgrid momentum transport even when the vertical velocity is included as an input. Overall, our results imply that the use of non-local variables and the vertical velocity as inputs could improve the performance of ML parameterizations, and the use of these inputs should be tested in online simulations in future work.

Plain Language Summary Current global climate models cannot resolve small-scale processes, such as clouds and convection, which are crucial for accurate simulations of climate, and the effect of these processes is approximated using parameterizations. Traditionally, these parameterizations rely on simple conceptual models, but in recent years machine learning (ML) has also been used to develop new parameterizations. Both traditional and ML parameterizations rely on a simple approach in which the vertical structure of a single atmospheric column is used to predict the effect of unresolved small-scale processes on the column itself. Here we use ML to show that this single-column approach might hamper the accuracy of parameterizations. We demonstrate that a machine-learning parameterization that uses information from multiple atmospheric columns simultaneously (rather than information from a single atmospheric column) better predicts the effects of small-scale processes compared to the same approach but only using information from a single column. We show that non-local inputs are especially important for parameterizing subgrid momentum transport and for mid-latitude situations with atmospheric fronts. Including only neighboring columns is sufficient to improve the parameterization in climate model simulations, and therefore the increase in computational expense should not be a barrier in the implementation of non-local parameterizations.

Key Points:

- Using non-local inputs from neighboring atmospheric columns improves offline performance of a neural network parameterization
- Improvements in mid-latitudes are associated with cases with mid-latitude fronts, where subgrid squall-line features are present
- An explainable artificial intelligence technique shows non-local winds are especially useful for parameterizing subgrid momentum transport

Supporting Information:

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

Correspondence to:

P. Wang and J. Yuval,
pdwang@mit.edu;
janniy@mit.edu

Citation:

Wang, P., Yuval, J., & O'Gorman, P. A. (2022). Non-local parameterization of atmospheric subgrid processes with neural networks. *Journal of Advances in Modeling Earth Systems*, *14*, e2022MS002984. https://doi.org/10.1029/2022MS002984

Received 3 JAN 2022
Accepted 27 AUG 2022

© 2022 The Authors. Journal of Advances in Modeling Earth Systems published by Wiley Periodicals LLC on behalf of American Geophysical Union. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

1. Introduction

Accurate climate projections are of great societal relevance (e.g., in assessing the risk from heavy rainfall events) and scientific interest (e.g., in understanding the dynamics of the climate system). These projections rely on global climate models that typically have grid spacing of a few tens to a hundred kilometers and thus, cannot resolve processes that occur on smaller scales (i.e., subgrid processes). Because subgrid processes, such as convection and clouds, have important consequences for Earth's climate, there is a need to represent them using parameterizations. These parameterizations approximate the effects of unresolved processes on the resolved fields. Traditional parameterizations rely partly on physics but also on simple conceptual models and heuristic approximations, and they are a major source of uncertainties in climate models and climate projections (Bony et al., 2015; Schneider et al., 2017; Sherwood et al., 2014; Wilcox & Donner, 2007).

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In recent years, it has been suggested that data-driven parameterizations based on machine learning (ML) could be computationally more efficient and/or more accurate than traditional parameterizations (Gentine et al., 2018; Krasnopolsky et al., 2013; O'Gorman & Dwyer, 2018). For example, several studies have used output from the super-parameterized Community Atmosphere Model to emulate its super-parameterization for aquaplanets and more realistic configurations (Han et al., 2020; Mooers et al., 2020; Rasp et al., 2018). Other studies have learned from output from three-dimensional high-resolution simulations which resolve processes that are usually subgrid in global climate models (Brenowitz & Bretherton, 2018, 2019; Yuval & O'Gorman, 2020; Yuval et al., 2021). In this approach, the output from the high-resolution simulation is first coarse-grained (i.e., averaged onto a coarser grid), and then an ML algorithm is used to predict the effect of the small-scale processes on the (coarse-grained) prognostic variables.

Both traditional and ML-based parameterizations of subgrid atmospheric processes rely on a single column framework. In this framework, the vertical profiles of the moisture, temperature and winds in an atmospheric column are typically used as the inputs to the parameterization. In turn, the parameterization predicts the effect of small-scale processes on the resolved prognostic fields in the atmospheric column. The motivation to use single-column parameterization relies on the idea that the subgrid processes primarily rearrange mass, momentum and energy in the vertical, and that a single-column framework is adequate to model these vertical processes (Stensrud et al., 2015). In addition, a single column formulation does not require horizontal communication between columns which may be stored on different processors, thus increasing computational efficiency and simplifying code development. Thus, traditional and ML parameterizations for the atmosphere have been based on a single-column approach, with the exception of some stochastic parameterizations in which the stochasticity is implemented non-locally (Palmer, 2001). We note also that an ML parameterization has been developed that includes non-locality in time (Han et al., 2020) which may be related to non-locality in space for propagating weather systems.

However, a single-column parameterization structure may not be ideal when predicting the effect of certain subgrid atmospheric processes or for certain atmospheric conditions. For example, slantwise convection, which is related to conditional symmetric instability that is prevalent in the subtropics and mid-latitudes (Chen et al., 2018), is not a purely vertical process and may not be well predicted using only inputs from a single vertical column. Moreover, mesoscale convective systems are driven by small-scale convective processes which are not resolved by current global climate models, but are organized in coherent tilted structures that are larger than a single grid box of a global climate model (Houze, 2004). In such multiscale convective systems, unresolved convective processes in an atmospheric column might have some statistical relation with the atmospheric state of neighboring columns, and therefore including information from neighboring columns in parameterizations could potentially improve their performance. In addition, knowledge about the three dimensional structure of the winds may help in estimating the magnitude and direction of subgrid momentum transport.

The examples mentioned above provide the motivation for this study in which we test whether using non-local information in the horizontal could improve ML parameterizations and potentially also traditional parameterizations. We learn from coarse-grained output of a high-resolution three-dimensional simulation, noting that the alternative of emulating a super-parameterization is less attractive in this case since the super-parameterization already imposes a single-column structure. In related recent work, a non-local convolutional neural network (NN) was used to predict the horizontal subgrid eddy momentum forcing for an ocean gyre circulation using inputs over 40×40 horizontal blocks of grid boxes (Bolton & Zanna, 2019). Here, given our focus on convection, we use inputs from the 3×3 grid columns in the immediate neighborhood of the target column. In a distributed computing environment, such neighboring inputs would typically be available on a given processor through a halo around the subdomain for that processor, and thus using only 3×3 columns helps to limit inter-processor communication and computational expense.

We also investigate which non-local features are important to improve the predictions of the parameterization. We find that the non-local variables that ML parameterizations rely on are mostly the wind fields. For some outputs, the patterns of inputs used suggest that the parameterization is learning the horizontal wind divergence, which is related to the vertical velocity through the mass continuity equation. This motivates us to also study the use of the (local) vertical velocity as an input to a single-column parameterization. Some conventional convection schemes rely on closures in terms of the vertical velocity (Ooyama, 1969) or the moisture convergence which is closely related to the vertical velocity (Kuo, 1974), although this is less common in climate models compared

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

to closures based on measures of instability (see Table 2 of Pathak et al. (2019)). Whether variables such as the vertical velocity or moisture convergence should be included as inputs to convection and cloud parameterizations is the subject of debate (Emanuel et al., 1994; George et al., 2021). In particular, using convergence as an input may lead to reverse causation since convergence is both a cause and a consequence of convection (Back & Bretherton, 2009), and this is potentially an issue in the context of ML parameterizations which can be unstable as a result of learning non-causal relations between inputs and outputs (Brenowitz et al., 2020). Given the uncertainty as to whether the divergence or vertical velocity should be included as inputs in ML parameterizations, we choose to present results for parameterizations with and without these inputs.

We organize the paper as follows. In Section 2, we describe the high-resolution simulation used to build our training and testing data sets (Section 2.1), the NN parameterizations (Section 2.2), and an explainable ML technique, called layer-wise relevance propagation, that we use to interpret NN parameterizations (Section 2.3). Then in Section 3, we begin to analyze the results by quantifying the performance of the parameterizations and comparing parameterizations that use non-local inputs to parameterizations that use a single column structure, with and without vertical wind as an input (Section 3.1). Next, we focus on mid-latitudes where non-local inputs are especially useful and identify atmospheric conditions in which the non-local parameterization substantially improves or does not improve the predictions (Section 3.2). We use layer-wise relevance propagation to understand on which non-local inputs the NN parameterizations rely (Section 3.3). We also briefly discuss the results for the deep tropics (Section 3.4). Finally, in Section 4 we give a summary and the conclusions of the study.

2. Data and Methods

2.1. Simulations

The high resolution data was obtained from a quasi-global aquaplanet simulation (referred to as hi-res) on an equatorial beta plane using the System for Atmospheric Modeling (SAM) version 6.3 (Khairoutdinov & Randall, 2003). The domain of the simulation has a meridional extent of 17, 280 km and a zonal width of 6, 912 km, equivalent to a latitude range from −78.5° to 78.5° and a longitudinal extent of 62.2° at the equator. To reduce the computational resources necessary to run a quasi-global simulation that resolves deep convection, we use a horizontal grid spacing of 12 km combined with a hypohydrostatic-rescaling factor of 4. Hypohydrostatic rescaling increases the horizontal length scale of convection without affecting the larger-scale flow (Boos et al., 2016; Fedorov et al., 2019; Kuang et al., 2005). The sea surface temperature is specified to be the "qobs" distribution of Neale and Hoskins (2000), which is zonally and hemispherically symmetric and peaks at the equator. The default time step is 24 s, but this time step is reduced if the Courant-Friedrichs-Lewy condition would otherwise be violated. There are 48 vertical layers that extend up to 28.7 km. The first 100 days are considered as spin-up, and we use three-dimensional instantaneous snapshots from a subsequent 337.5 days that were saved every 3 model hours. Detailed description of this simulation can be found in Yuval and O'Gorman (2020).

To obtain training data, we follow the coarse-graining protocol described in Yuval et al. (2021). In short, for each three-dimensional snapshot from hi-res, we coarse grain the prognostic variables as well as the temperature, vertical advective fluxes of energy, non-precipitating water and momentum, tendency of precipitating water due to cloud microphysics, turbulent diffusivity and radiative heating. Coarse-graining is performed by a spatial averaging to a horizontal grid spacing of 192 km (16×16 grid boxes). Subgrid tendencies and fluxes are then calculated using the equations of the model. SAM uses an Arakawa C-grid (Arakawa & Lamb, 1977) which introduces some challenges regarding how to coarse grain variables that are not found on the same horizontal grid (see discussion in the Supporting Information of Yuval and O'Gorman (2021)). Here we choose to coarse grain the data such that the coarse-grained data is found on a collocated grid (see Figure S1 in Yuval and O'Gorman (2021)). In other words, the coarse-grained quantities, including prognostic variables, tendencies, and fluxes, are all on the same horizontal grid.

2.2. Neural-Network Parameterizations

We train both a single-column parameterization (NN1D) and a non-local parameterization (NN3D), where the non-local parameterization uses inputs from the 3×3 atmospheric columns on the coarse grid (Figure 1). The atmospheric variables that are used as the default inputs for NN1D and NN3D are the coarse-grained absolute temperature (T), non-precipitating water mixing ratio (sum of water vapor, cloud liquid water and cloud

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

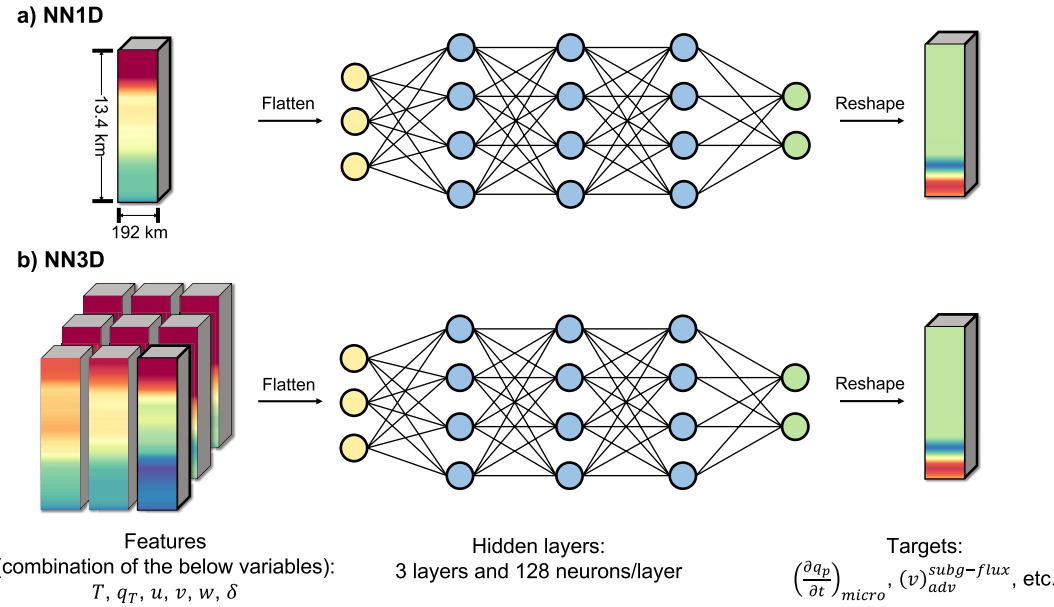

Figure 1. Schematic of the neural-network (NN) architectures we use in this study. The single-column parameterization (NN1D) uses inputs taken from a single atmospheric column that is the same as the target column containing the outputs, whereas the non-local parameterization (NN3D) uses inputs taken from the 3 × 3 columns of data centered on the target column. A combination of the following atmospheric variables are used as inputs: temperature (T), non-precipitating water mixing ratio (q_T), zonal, meridional and vertical velocities (u, v, w), and the horizontal wind divergence (δ), but the schematic only shows one input variable for illustration purposes. The targets are single-columns of eight output variables, including the tendency of total precipitating water mixing ratio due to microphysics $\left((\partial q_p/\partial t)_{\text{micro}}\right)$ and the subgrid meridional momentum flux due to vertical advection $\left((v)_{\text{adv}}^{\text{subg-flux}}\right)$.

ice mixing ratios; q_{T}), and zonal (u) and meridional (v) velocities. As discussed in the introduction, we also train single-column NN parameterizations that use the above four local input features as well as an additional input of the local vertical wind (this NN is referred to as NN1D + w) or the horizontal wind divergence (this NN is referred to as NN1D + δ where $\delta = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$). From symmetry considerations, in the Southern Hemisphere the meridional coordinate is flipped and the meridional velocity is changed in sign before training so that the NNs we train can generalize across hemispheres. The number of inputs is 120 for NN1D (four feature variables, each variable with 30 vertical levels), 1,080 for NN3D (four feature variables, each variable with 30 vertical levels and 3 × 3 columns), and 150 for NN1D + w and NN1D + δ (five feature variables, each variable with 30 vertical levels).

We train NNs to predict the following eight target variables simultaneously: the tendency of total precipitating water mixing ratio due to microphysics $\left((\partial q_p/\partial t)_{\text{micro}}\right)$; the subgrid fluxes of total non-precipitating mixing ratio due to vertical advection $\left((q_T)_{\text{adv}}^{\text{subg-flux}}\right)$ and sedimentation $\left((q_T)_{\text{sed}}^{\text{subg-flux}}\right)$; the tendency due to radiation $\left((\partial H_L/\partial t)_{\text{rad}}\right)$; the subgrid energy flux due to vertical advection $\left((H_L)_{\text{adv}}^{\text{subg-flux}}\right)$; the coarse-grained turbulent diffusivity $\left(\overline{D}\right)$; and the subgrid zonal and meridional momentum flux due to vertical advection $((u)_{\text{adv}}^{\text{subg-flux}}$ and $(v)_{\text{adv}}^{\text{subg-flux}})$. A detailed description of how the outputs are calculated is found in Yuval et al. (2021), except for the subgrid momentum fluxes which are described in Yuval and O'Gorman (2021).

For NN3D the outputs are only predicted at the center column, such that both local and non-local NNs have the exact same outputs. We only use inputs and outputs from the lowest 30 vertical levels of the model (extending to 13.4 km) to prevent the NNs from predicting near the sponge layer which is active above 20 km, and also because previous studies suggest that using stratospheric information can lead to numerical instabilities when parameterizations are implemented online (Brenowitz & Bretherton, 2019; Yuval et al., 2021).

We use 3-hourly snapshots taken from 337.5 days, resulting in 2,700 time snapshots. The coarse-grained data contains 90 × 36 atmospheric columns (samples) for each snapshot, but the two outermost columns of data in

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

each snapshot are not used to simplify both the training of the non-local parameterization and the calculation of the horizontal divergence which is used later. Therefore, each time snapshot contains $(90-4) \times (36-4) = 2{,}752$ samples, resulting in total of 7,430,400 samples. We use the first 50% of the simulated data for training (3,715,200 samples), the middle 10% of the simulated data for validation (743,040 samples) and the remaining 40% of the simulated data for testing (2,972,160 samples). All the results shown in this paper are based only on the testing set. The reason for using only 50% of data for training is because we want to have a large testing data set to ensure robust results for case composites. We verified that training on 80% of the data leads to similar coefficient of determination (R^2) values compared to when training on 50% of the data (Figure S1 in Supporting Information S1).

Before training, the input variables are standardized such that each input at each vertical level has a mean of zero and a standard deviation of one. The outputs are also standardized by removing the mean and rescaling by the standard deviation, but the mean and standard deviation are calculated over all vertical levels, such that the output standardization consists of a single mean value and a single standard deviation value for all levels of each output variable. We use the mean squared error as a loss function, and we use the Adam optimizer (Kingma & Ba, 2014) to update the weights and biases. The training process is the same for all networks we train. The NNs are first trained for seven epochs, with a cyclic learning rate (Smith, 2017) bounded by 2×10^{-4} to 2×10^{-3}. They are then trained for another five epochs with reduced cyclic learning rate bounded by 2×10^{-5} to 2×10^{-4}. We apply "early stopping" by using the validation data to evaluate the network after each epoch, and we choose the NN weights and biases that performed best on the validation data. The default NN architecture we use in this study has three hidden layers where each layer has 128 neurons, and we use rectified linear unit activations (ReLu) except in the output layer. Figure S2 in Supporting Information S1 shows the training and validation loss versus epoch for NN1D, NN3D, NN3D + δ and NN1D + w, as well as the learning rate during each epoch.

2.3. Layer-Wise Relevance Propagation

We use layer-wise relevance propagation (LRP; Bach et al. (2015)) to better understand the inputs that are most important for NN3D. LRP propagates the relevance score from the output layer back to the input layer. Therefore, each input has an associated relevance score, where a higher relevance score indicates that the NN relies more on this input for the specific sample that is tested. The propagation rules used are given in Text S1 in Supporting Information S1.

Different output variables with distinct physical processes can rely on different inputs. The relevance score from the multi-target NN might be hard to interpret because it entangles the physical processes from all the eight target variables. Therefore, when applying LRP for a given output variable, we retrain the NN parameterizations to have only the single output variable we are interested in. These single-output NNs have similar offline performance as the multi-target NN. LRP provides the relevance score for each input variable and grid box at 30 vertical levels. Because we want to focus on the reliance of NNs on horizontally non-local inputs which is the novelty of this paper, we sum the relevance across heights, treating height levels as color channels in image recognition problems. However, the relevance score can be positive or negative, and both signs in relevance are physically meaningful. Therefore, we sum over the absolute value of the relevance score across height levels. We then normalize the vertically summed relevance sample by sample, such that the relevance values for all variables and columns add up to one in each sample. This better illustrates the relative importance of each variable for each column, and the magnitude of vertically summed relevance can be easily interpreted.

3. Results

3.1. Performance of Parameterizations

To measure the performance of NN1D, NN3D, NN1D + w, and NN1D + δ, we calculate the global R^2 values by concatenating vertical columns for each target variable (Figure 2). The non-local parameterization NN3D improves on NN1D for almost all variables. The only exceptions are the tendency due to radiation and the coarse grained diffusivity for which the non-local parameterization is either a disimprovement or unaltered, respectively. For most output variables, NN1D + w has better performance than NN3D, with the important exception that NN3D outperforms NN1D + w for the zonal and meridional subgrid momentum fluxes. As discussed in the introduction, it is not clear whether the vertical velocity should be included as an input from the point of view of

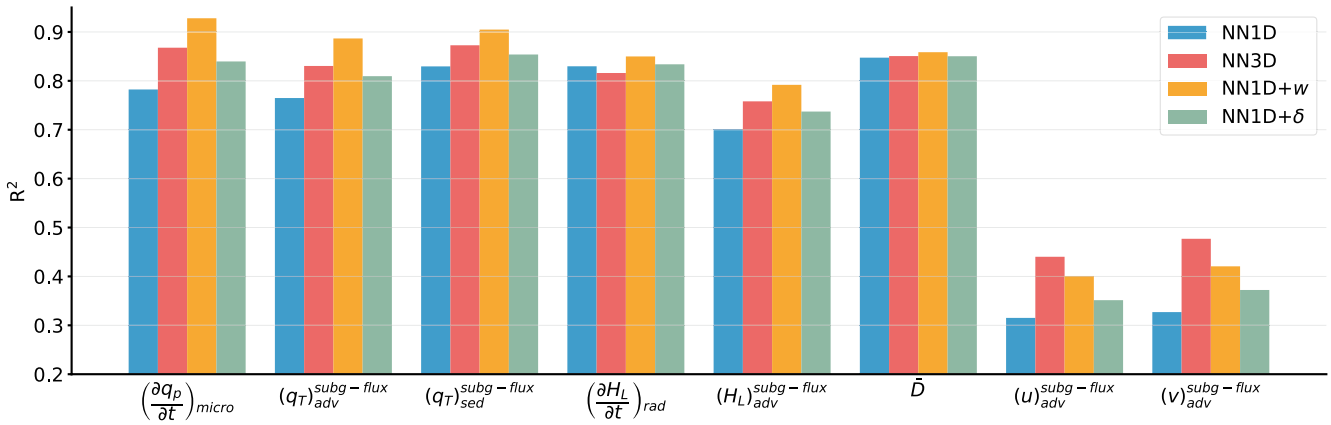

Figure 2. Global coefficient of determination (R^2) for a variety of output variables using a single-column parameterization (NN1D; blue), a non-local parameterization (NN3D; red), a single-column parameterization with the vertical wind (w) being an additional input feature (NN1D + w; yellow), and a single-column parameterization with the horizontal wind divergence (δ) being an additional input feature (NN1D + δ; green). All four neural networks predict all the presented variables simultaneously (multi-output prediction). The predicted outputs are the tendency of total precipitating water mixing ratio due to microphysics $\left((\partial q_p/\partial t)_{\text{micro}}\right)$; the subgrid flux of total non-precipitating mixing ratio due to vertical advection $\left((q_T)_{\text{adv}}^{\text{subg-flux}}\right)$ and sedimentation $\left((q_T)_{\text{sed}}^{\text{subg-flux}}\right)$; the subgrid tendency due to radiation $\left((\partial H_L/\partial t)_{\text{rad}}\right)$; the subgrid energy flux due to vertical advection $\left((H_L)_{\text{adv}}^{\text{subg-flux}}\right)$; the coarse-grained diffusivity $\left(\overline{D}\right)$; and the subgrid zonal and meridional momentum flux due to vertical advection ($(u)_{\text{adv}}^{\text{subg-flux}}$ and $(v)_{\text{adv}}^{\text{subg-flux}}$).

causality and robustness in online simulations, and thus we present results with and without the vertical velocity as an input.

The performance of NN1D + δ is better than NN1D but worse than the performance of NN1D + w and NN3D, and the relation of the horizontal wind divergence (δ) to the non-local wind inputs in NN3D will be discussed further in Section 3.3. It may seem surprising that using the horizontal wind divergence as an input is not equivalent to using the coarse-grained vertical velocity as an input given that these are related by the anelastic mass continuity equation in SAM. However, the coarse-grained (i.e., horizontal averaged) vertical velocity is directly related by mass continuity to line averages of the horizontal winds on the boundaries of the grid cells. These line averages are effectively subgrid compared to the coarse-grained horizontal winds. Thus, including the vertical velocity as an input is not equivalent to including the horizontal divergence as an input, and we present separate results for parameterizations using these inputs.

One caveat of the global R^2 values is that they could be overstated because they partly reflect the NN parameterization correctly predicting the mean at each vertical level. To verify that the NN parameterizations predict accurately beyond the means at each vertical level, we also tested the performance after the means at each vertical level are first removed (Figure S3 in Supporting Information S1). When the means at each vertical level are first removed, R^2 values for all the targets are slightly lower (and substantially so for the turbulent diffusivity), but the differences between different NNs remain similar.

From now on, we primarily focus on two output variables: the tendency of total precipitating water mixing ratio due to microphysics $\left((\partial q_p/\partial t)_{\text{micro}}\right)$ and the subgrid meridional momentum flux due to vertical advection $\left((v)_{\text{adv}}^{\text{subg-flux}}\right)$. We choose to focus on these two output variables because (a) their prediction is most accurate for NNs that use different inputs, (b) $(\partial q_p/\partial t)_{\text{micro}}$ is directly related to surface precipitation (Text S2 in Supporting Information S1) which is a variable of great interest, and (c) subgrid momentum transport has previously been found to be challenging to predict (Yuval & O'Gorman, 2021), and therefore it is especially interesting to investigate how its prediction can be improved using non-local inputs. The results for $(u)_{\text{adv}}^{\text{subg-flux}}$ are similar in most regards to the results for $(v)_{\text{adv}}^{\text{subg-flux}}$, but the improvement from using non-local inputs is slightly greater for $(v)_{\text{adv}}^{\text{subg-flux}}$, and thus we focus on it throughout the paper.

We next consider R^2 values calculated at each latitude for the two output variables of interest (Figures 3a and 3b). Both NN3D and NN1D + w perform substantially than NN1D at all latitudes, where NN1D + w predicts more accurately $\left(\partial q_{\text{p}}/\partial t\right)_{\text{micro}}$, and NN3D predicts more accurately $(v)_{\text{adv}}^{\text{subg-flux}}$. Unsurprisingly given the close relation of

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

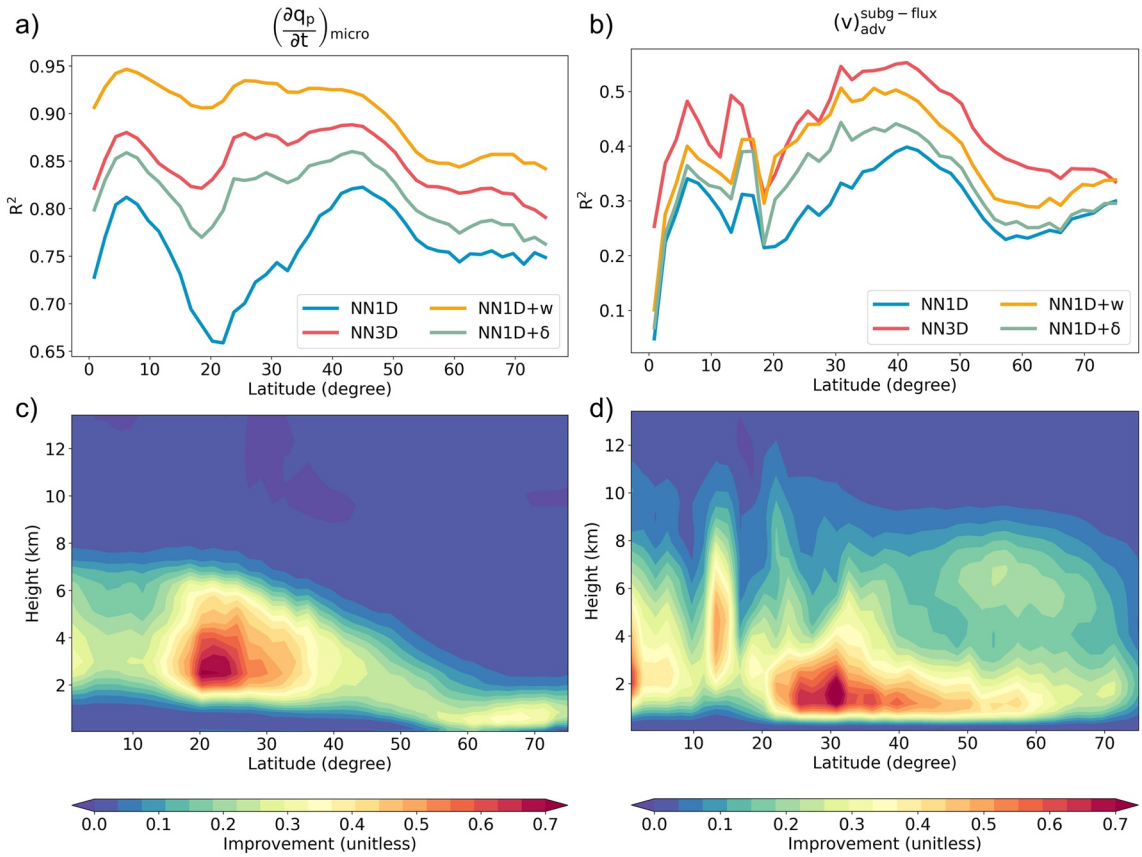

Figure 3. (a, b) Coefficient of determination (R^2) as a function of latitude for (a) the precipitating water tendency due to microphysical processes $\left(\left(\partial q_p/\partial t\right)_{\text{micro}}\right)$ and (b) the subgrid meridional momentum flux due to vertical advection $\left((v)_{\text{adv}}^{\text{subg-flux}}\right)$ for different neural networks: a single-column parameterization (NN1D; blue), a single-column parameterization with an additional 1D vertical velocity input (NN1D + w; yellow), a single-column parameterization with an additional 1D wind divergence input (NN1D + δ; green), and a non-local parameterization (NN3D; red). (c, d) The zonal-mean improvement index (Equation 1) for NN3D compared to NN1D averaged over all test samples as a function of latitude and height for (c) $\left(\partial q_p/\partial t\right)_{\text{micro}}$ and (d) $(v)_{\text{adv}}^{\text{subg-flux}}$.

$\left(\partial q_{\text{p}}/\partial t\right)_{\text{micro}}$ to precipitation rates, both NN3D and NN1D + w outperform NN1D in predicting extreme precipitation (Figure S4 in Supporting Information S1).

To verify that the improvement is robust to changes in the network architecture, and is not due to the larger number of tunable parameters used in NN3D and NN1D + w compared to NN1D, we also train networks with different complexities by adjusting the number of hidden layers and neurons. Specifically, we compare the default architecture of three hidden layers and 128 neurons per layer to a shallow network with a single hidden layer and 64 neurons, and a more complex architecture with eight hidden layers and 256 neurons per layer. In all cases, performance improves when changing from a single hidden layer to three hidden layers, but performance stays similar or disimproves when changing from three hidden layers to a deeper network with eight hidden layers (Figure S5 in Supporting Information S1). More importantly we find that NN3D or NN1D + w even with a single hidden layer outperforms NN1D for all architectures. This implies that the additional information provided as inputs to NN3D and NN1D + w improves their performance. For the rest of the paper, we will only focus on the results using three hidden layers and 128 neurons per layer.

3.2. In Which Atmospheric States Does Non-Locality Help the Parameterization in Midlatitudes?

Next, we want to understand for which atmospheric states NN3D predicts better than NN1D. We first define an improvement index for each individual test sample for each of the output variables and each atmospheric level as:

$$\text{Improvement} = \frac{(\text{NN1D} - \text{true})^2 - (\text{NN3D} - \text{true})^2}{\sigma_{true}^2}, \tag{1}$$

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

where $(\mathrm{NN1D}-\mathrm{true})^2$ is the squared error of NN1D output, $(\mathrm{NN3D}-\mathrm{true})^2$ is the squared error of NN3D output, and σ^2_{true} is the variance of the ground truth over the column and over all testing samples which is a latitude-dependent variable. A large improvement index indicates that the difference between the squared errors of NN1D and NN3D is large compared to the climatological variance at a given latitude. By taking the zonal mean of the improvement indices at every altitude and latitude, we find that most of the improvements occur in subtropical and mid-latitude tropospheric regions for both $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$ and $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$, although the improvements in $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$ are more spread in the vertical and maximize at lower altitudes (Figures 3c and 3d). We note that for $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$, the absolute improvement in the tropics is even larger than the absolute improvement in the mid-latitudes, but the variance in the tropics is very large, and therefore the relative improvement is smaller in that region. For $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$, the biggest improvement occurs in the mid-latitudes in the both absolute and relative sense.

We will focus on the midlatitude band 20°–40° in both hemispheres which shows large improvements for NN3D. The region of greatest improvement actually extends somewhat further equatorward than 20° latitude, but we focus on 20°–40° to avoid mixing the tropical and mid-latitude dynamical regimes. We next compare groups of cases in this latitude band with large and small improvement indices. Each case is a testing sample with 3×3 horizontal grid boxes and 30 vertical levels, and thus has a horizontal length scale of 574 km and a vertical height scale of 13.4 km. We find that cases with large improvement tend to have heavier precipitation on average than cases with no improvement. Therefore, in order to make a fair comparison between cases with and without improvement, we only select cases that have true instantaneous precipitation between 50–70 mm day^{-1} (corresponding to 99.5th–99.9th percentile of precipitation over the mid-latitude band). From cases with these precipitation rates, we use the improvement criterion (Equation 1) to select the 300 cases for each output variable that have the largest column-mean improvement index, which we refer to as the "cases with largest improvement." We also select the 300 cases with the smallest column-mean improvement index that is still positive, which we refer to as the "cases with little improvement." Among "cases with largest improvement," the improvement indices range between 11.5–92.7 for $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$, and 16.6–511.9 for $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$. And among "cases with little improvement," the improvement indices range between $2.2 \times 10^{-3} - 0.6$ for $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$, and $1.5 \times 10^{-3} - 0.6$ for $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$. Forty-one percent of the testing samples have negative improvement indices for $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$ and 46% for $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$, but most of the negative improvements are close to zero.

We find that cases where NN3D improves the prediction tend to have different cloud shapes compared to the cases where NN3D does not improve the prediction (Figure 4 for $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$ and Figure S6 in Supporting Information S1 for $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$). Specifically, for cases with improvement the clouds tend to be organized in coherent linear and narrow features which are squall lines associated with atmospheric fronts. Coarse-graining smears out sharp boundaries such as fronts and narrow convective features such as squall lines making the prediction of subgrid tendencies and fluxes more difficult. Therefore single-column coarse-grid inputs might not be informative enough to make an accurate prediction when such strong subgrid variability is present, but a non-local parameterization may be able to use non-local information in order to better understand the atmospheric conditions and improve the prediction of the effect of subgrid processes on the resolved scales. For cases without improvement, clouds are more uniformly distributed, and the inputs from non-local columns does not convey information which assists in the prediction of the effect of subgrid processes. We find that doing a similar analysis for NN1D + w gives similar results (e.g., 260 cases out of the 300 improved cases are shared between NN3D and NN1D + w for $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$).

To better understand in which atmospheric conditions a non-local parameterization is superior to a single-column parameterization, we calculate the convective available potential energy (CAPE) from the high resolution data for cases with largest improvement and cases with little improvement. CAPE is calculated assuming reversible ascent with convective inhibition being removed as described in Text S3 in Supporting Information S1 following Muller et al. (2011). We find that the CAPE is larger in cases with the largest improvement compared to cases with little improvement (Figure 5 for $\left(\partial q_p/\partial t\right)_{\mathrm{micro}}$, and similar results in Figure S7 in Supporting Information S1 for $(v)^{\mathrm{subg-flux}}_{\mathrm{adv}}$), suggesting the atmosphere is more unstable to convection in the improved cases. The CAPE composites show higher CAPE at the equatorward and eastward ends of the 3×3 subdomain, and this is because of the general increase in CAPE equatorward and the structure of the flow for these cases which are all precipitating. We also find that the improved cases are more unstable for two other measures of instability as shown in Figures S8 and S9 in Supporting Information S1: the saturation potential vorticity (negative values are indicative of condi-

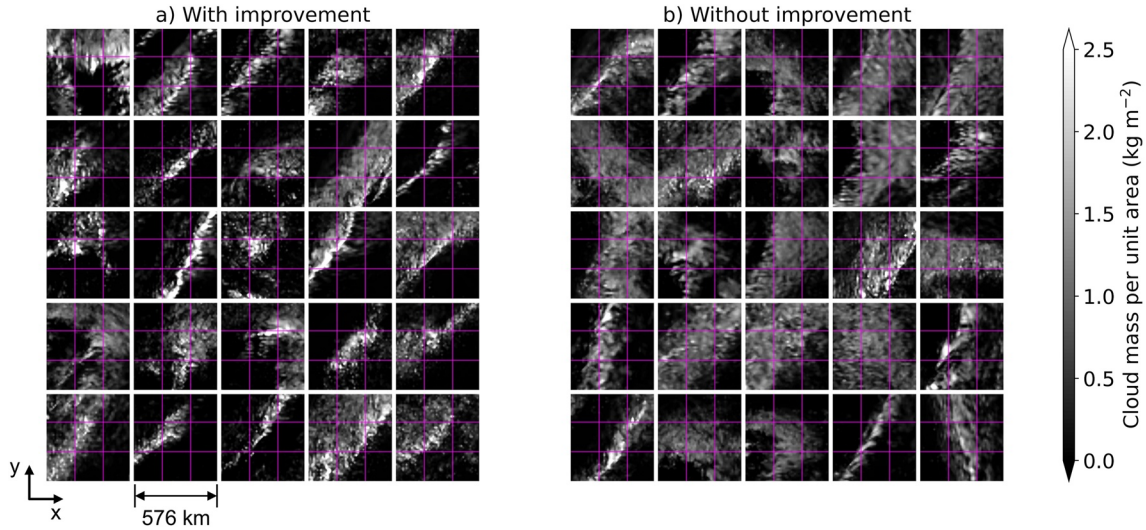

Figure 4. Snapshots of the column-integrated cloud mass per unit area plotted from the hi-res output for mid-latitude cases with and without improvement when using the non-local parameterization (NN3D) to predict $\left(\partial q_p/\partial t\right)_{micro}$ (similar results are found for $(\upsilon)_{adv}^{subg-flux}$ as shown in Figure S6 in Supporting Information S1). Shown are 25 randomly chosen cases from (a) the 300-member group with large-improvement when the non-local parameterization is used, and (b) the 300-member group with little improvement when the non-local parameterization is used. See Section 3.2 for details on how these groups were chosen which involves only selecting cases with heavy precipitation in the latitude band 20°–40°. The domains are equivalent to 3 × 3 coarse-grained grid boxes (576 km in each horizontal direction) with the edges of the coarse-grained grid boxes plotted in magenta.

tional symmetric instability which could give rise to slantwise convection) and the vertical derivative of the saturation moist static energy (negative values are indicative of upright conditional instability). Calculation of these instability metrics is described in Text S4 in Supporting Information S1. One possibility for the improvement in non-local parameterization is that it could be especially relevant for correctly predicting slantwise convection since slantwise convection is not a purely vertical process and is sensitive to horizontal gradients that can be estimated by the non-local parameterization, and since slantwise convection may also involves more than one coarse atmospheric column. Interestingly, the latitude band we find that has the greatest relative improvement (roughly

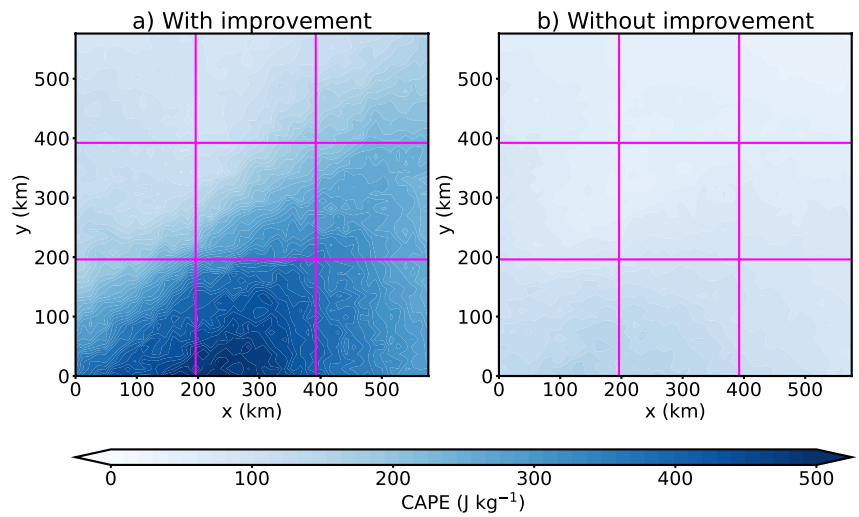

Figure 5. Composite of convective available potential energy (CAPE) for cases in the mid-latitudes (20°–40°) for (a) the group of 300 cases with largest improvement when the non-local parameterization (NN3D) is used to predict $\left(\partial q_p/\partial t\right)_{micro}$, and (b) the group of 300 cases with little improvement in predicting $\left(\partial q_p/\partial t\right)_{micro}$. Each case is equivalent to a single testing sample, and see Section 3.2 for details of how the groups of cases are chosen. CAPE is calculated from the hi-res output. The domain shown is equivalent to 3 × 3 coarse-grained grid boxes (576 km in each horizontal direction) with the edges of the coarse-grained grid boxes plotted in magenta. Similar results for $(\upsilon)_{adv}^{subg-flux}$ are shown in Figure S7 in Supporting Information S1.

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

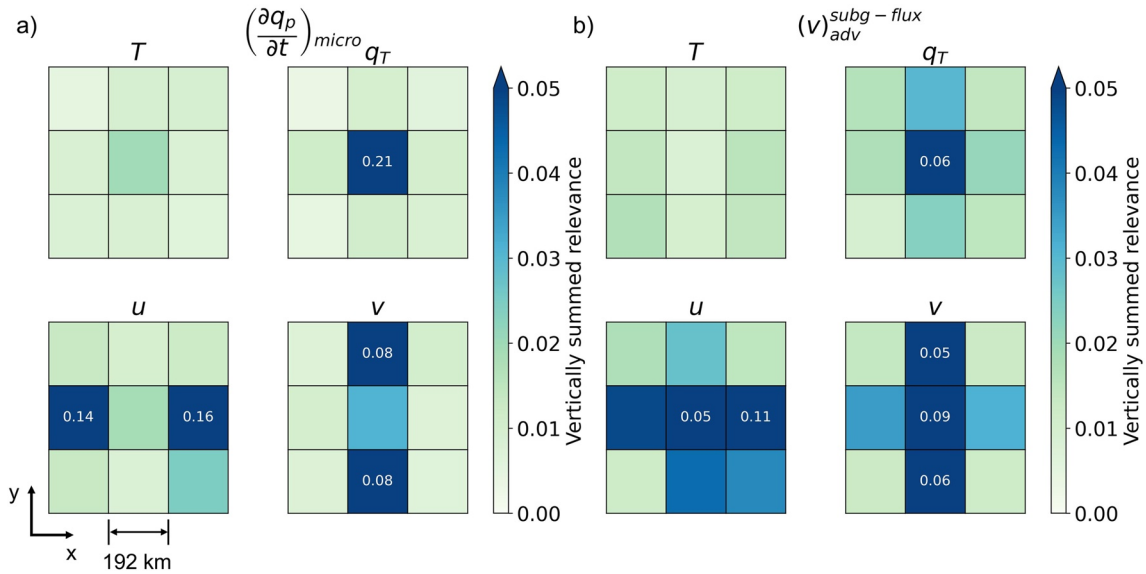

Figure 6. Vertically summed relevance (as calculated from layer-wise relevance propagation) averaged over the group of 300 mid-latitude (20°–40°) cases with largest improvement when the non-local parameterization (NN3D) is used to predict (a) $\left(\partial q_p/\partial t\right)_{micro}$ and (b) $(v)_{adv}^{subg-flux}$. Each sub-panel shows the vertically summed relevance for the 3 × 3 atmospheric input columns of a different input variable: temperature (T), non-precipitating water mixing ratio (q_T), zonal wind (u), and meridional wind (v). To better illustrate the variations in relevance, we set the upper limit of the color bar to 0.05 and explicitly write the relevance for the inputs that saturate the color bar. See Section 3.2 for details of how the cases were chosen.

10°–40°) is similar to the latitude band in which conditional symmetric instability is most favored over upright convective instability in reanalysis data (see Figure 2b of Chen et al. (2018)). However, we note that NN3D being better at recognizing conditional symmetric instability is only one possible reason for the improvement, and further work is needed to support this hypothesis.

3.3. Which Non-Local Inputs Are Useful in Midlatitudes and How do They Relate to Horizontal Wind Divergence?

We first test whether extending the non-locality beyond using information from the closest neighbors (3 × 3 atmospheric columns) gives further improvements. To do this, we trained NNs with non-local information from 5 × 5 and 7 × 7 atmospheric columns (Figure S10 in Supporting Information S1). For almost all output variables, the best performing networks rely on 3 × 3 atmospheric columns. We conclude that extending the non-locality beyond the closest neighboring atmospheric columns is not helpful to further improve the parameterization.

We next use LRP to calculate the relevance score for the cases showing the largest improvements from NN3D (Figure 6). A separate NN3D is trained to predict each of the two output variables on their own in order to perform the LRP analysis with only one output variable. This is to prevent mixing the relevance for different physical processes that matter for different target variables. The groups of cases with largest improvement are the same groups of cases selected in previous sections for consistency. Each grid box in Figure 6 indicates a single column of data for a different input variable, with the absolute value of the relevance score being vertically summed and then averaged across cases. Figure S11 in Supporting Information S1 shows that the LRP results are robust to changes in the LRP parameters (described in Text S1 in Supporting Information S1).

Starting with the NN3D predicting $\left(\partial q_p/\partial t\right)_{micro}$, we find that for the temperature and non-precipitating water input variables, the most relevant grid box is the center column, which means that thermodynamic and moisture variables follow the reasoning of using a single column for parameterizations (Figure 6a). However, the most relevant grid boxes for the wind input variables are found at the non-local atmospheric columns. Interestingly, for the zonal wind the most relevant columns are east and west of the center column, whereas for the meridional wind the most relevant columns are north and south of the center column, and the relevance is symmetric about the center column, which suggests that the NN is using the non-local winds to reconstruct the horizontal wind divergence $\delta = \frac{\partial u}{\partial x} + \frac{\partial v}{\partial y}$. For the NN3D predicting $(v)_{adv}^{subg-flux}$, the local moisture input is still important, but non-local moisture

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

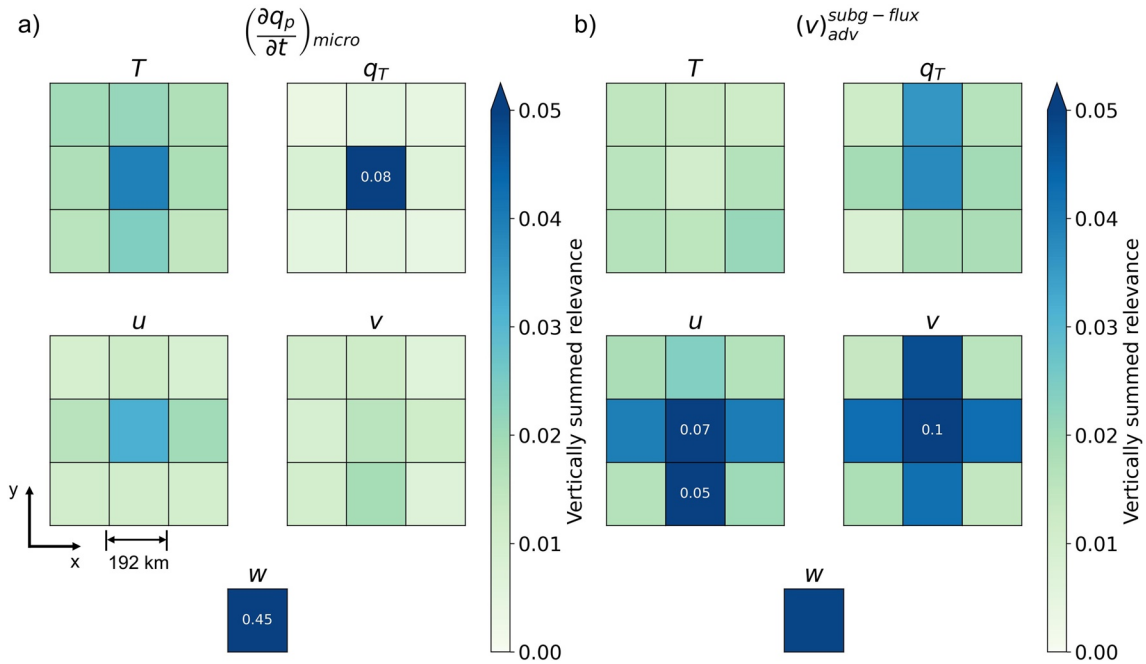

Figure 7. As in Figure 6, but the relevance is plotted for non-local neural-net parameterizations that include the vertical velocity at the center column as an additional input (NN3D + w) for cases where there are large improvements for NN3D + w compared to NN1D.

inputs also contribute (Figure 6b). Furthermore, NN3D relies heavily on both the local and non-local columns of horizontal wind fields. Unlike for the NN3D that predicts $\left(\partial q_p/\partial t\right)_{micro}$, the relevance for NN3D that predicts $(v)^{subg-flux}_{adv}$ is not symmetric around the center column, and there are high-relevance regions that are not necessary to calculate the horizontal wind divergence, which implies that this NN uses features beyond the horizontal wind divergence. Overall, these patterns imply that NN3D is likely reconstructing the horizontal wind divergence from non-local wind fields to help predict $\left(\partial q_p/\partial t\right)_{micro}$. This conclusion from LRP is consistent with the increase in performance when the divergence is added as an input to NN1D (Figure 3a). The horizontal wind divergence may also be reconstructed to help predict $(v)^{subg-flux}_{adv}$ but other aspects of the non-local winds are also being used for that output.

We next investigate the relevance scores for cases with largest improvement for a non-local NN that additionally gets as an input the vertical wind at the center column (this network is referred to as NN3D + w). LRP results for NN3D + w could potentially show the competing effects between local vertical velocity and non-local horizontal wind fields if the non-local winds are used to construct the horizontal divergence and thus approximate the vertical wind. For NN3D + w that predicts $\left(\partial q_p/\partial t\right)_{micro}$, we find that the relevance for non-local wind inputs reduces dramatically compared to NN3D, and 45% of the total relevance score comes from the vertical wind at the center column (Figure 7a). This result provides evidence that NN3D tries to estimate the vertical wind through the horizontal wind divergence in order to better predict $\left(\partial q_p/\partial t\right)_{micro}$, and hardly relies on non-local information beyond the horizontal wind divergence. Furthermore, there is a large reduction in the relevance of q_T in NN3D + w compared to NN3D. This reduction is likely due to a correlation between large values of w and large values of q_T during precipitation and convection events, and therefore some of the informtion contained in q_T is not independent of the information contained in w. For NN3D + w that predicts $(v)^{subg-flux}_{adv}$, we find that the relevance for the non-local wind inputs reduces only by a moderate amount (Figure 7b). Furthermore, only 4.7% of the total relevance score comes from the vertical wind at the center column. These LRP results provide evidence that when NN3D is predicting $(v)^{subg-flux}_{adv}$, it indeed relies on non-local information beyond that necessary to construct the horizontal wind divergence, and this is also consistent with the ability of NN3D to outperform NN1D + w when predicting subgrid momentum transport (Figure 3b).

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

3.4. Cases With Improvement and Role of Non-Local Inputs in the Tropics

Following the same general approach as for the mid-latitudes, we select cases with largest improvement and cases with little improvement when using the non-local parameterization but for the deep tropics defined as latitudes between 10°S and 10°N. Cases with improvement are selected as the 300 cases with the largest column-mean improvement indices and with instantaneous precipitation between 100–150 mm day^{-1} (corresponding to the 99.5th–99.9th percentile in the deep tropics), resulting in column-mean improvement indices between 3.9–35.6 for $\left(\partial q_p/\partial t\right)_{\text{micro}}$, and 14.3–249.5 for $(v)_{\text{adv}}^{\text{subg-flux}}$. Cases with little improvement are selected as the 300 cases with the smallest column-mean improvement indices that are still positive and with instantaneous precipitation between 100–150 mm day^{-1}, resulting in column-mean improvement indices between $1.5 \times 10^{-3} - 0.3$, and $1.3 \times 10^{-3} - 1.1$ for $\left(\partial q_p/\partial t\right)_{\text{micro}}$ and $(v)_{\text{adv}}^{\text{subg-flux}}$ respectively.

Although NN3D has better performance compared to NN1D at all latitudes, the cases showing improvement appear to be different in the deep tropics and mid-latitudes. In the deep tropics, the cloud shapes for cases that have large improvements when using NN3D are not clearly distinguishable from cases that do not have an improvement as shown for example, for the $\left(\partial q_p/\partial t\right)_{\text{micro}}$ output in Figure S12 in Supporting Information S1. Furthermore, NN3D improves the prediction for more unstable conditions in the mid-latitudes (Figure 5) but for indistinguishable or slightly more stable conditions in the tropics, as shown for example, for cases with largest improvements for the $\left(\partial q_p/\partial t\right)_{\text{micro}}$ output in Figure S13 in Supporting Information S1. However, LRP results suggest similar columns of non-local data are most relevant for both the tropics and mid-latitudes (compare Figure 6 and Figure S14 in Supporting Information S1). Thus, the atmospheric conditions under which improvement is found differs between the tropics and mid-latitudes, but similar non-local information is used in both regions.

4. Conclusions

In this study we show that a neural-net (NN) parameterization using inputs of temperatures, moisture and horizontal winds from 3 × 3 atmospheric columns (NN3D; non-local in the horizontal) is generally more accurate compared to a single column NN parameterization (NN1D) in predicting the tendencies and fluxes due to different subgrid processes. We find that the relative improvement is especially large for the subtropics and mid-latitudes. Cases in mid-latitudes with the largest improvements have heavy precipitation and squall-line feature associated with fronts, and they tend to be convectively and symmetrically more unstable than cases with little improvement. We hypothesize that such unstable cases with large subgrid variations are more amenable to improvement through non-local information.

We use layer-wise relevance propagation, which is an interpretable ML technique, to determine which non-local features the parameterization relies on for its predictions. We focus on the prediction of two different target variables: the subgrid tendency of total precipitating water mixing ratio due to microphysics $\left((\partial q_p/\partial t)_{\text{micro}}\right)$, and the subgrid meridional momentum fluxes due to vertical advection $\left((v)_{\text{adv}}^{\text{subg-flux}}\right)$. The non-local parameterizations rely on different features for these two output variables. The prediction of $(\partial q_p/\partial t)_{\text{micro}}$ relies locally on temperature and moisture, but non-locally on horizontal wind variables. Interestingly, for this output variable NN3D uses the non-local wind features to construct the horizontal wind divergence in order to approximate the vertical wind, and it barely relies on other non-local variables. This result motivated us to train a single-column NN parameterization that includes the vertical wind as an input, and we find that such a single-column parameterization outperforms the non-local parameterization for $(\partial q_p/\partial t)_{\text{micro}}$. By contrast, the prediction of $(v)_{\text{adv}}^{\text{subg-flux}}$ relies both locally and non locally on the moisture and wind variables, and uses non-local wind information beyond that needed for the horizontal wind divergence. The non-local parameterization outperforms a single-column NN that includes the vertical velocity as an input for the output $(v)_{\text{adv}}^{\text{subg-flux}}$. Overall, we find that both non-local features and the local vertical velocity (or horizontal wind divergence) can substantially improve the offline performance of parameterizations, and that non-local features are especially important for the parameterization of subgrid momentum fluxes.

The large-scale horizontal wind and moisture divergence and the vertical velocity are both a cause and a consequence of convection, and there is debate as to whether these features should be included as an input to convection and cloud parameterizations (Emanuel et al., 1994; George et al., 2021). Detailed testing would be needed to determine how including these inputs in an ML parameterization affects simulation of the general circulation and

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

transient disturbances, and whether their use in an ML parameterization may affect the robustness or numerical stability of the simulations.

In the context of developing parameterizations by coarse-graining output from high-resolution data, including the vertical velocity as an input is not equivalent to including the horizontal divergence as an input. The underlying reason is that the coarse-grained output does not necessarily exactly obey the same equations as the high-resolution data. As an example, one can consider the anelastic mass continuity equation, through which the coarse-grained vertical velocity is directly related to line averages of the horizontal winds on the boundaries of the grid cells, rather than to the coarse-grained horizontal winds. However, these line averages are effectively subgrid compared to the coarse-grained horizontal winds, and indeed, we find that a single-column NN parameterization that includes the vertical velocity as an input perform substantially better than a single-column NN parameterization that includes the horizontal wind divergence. This inconsistency between the coarse-grained output and the anelastic continuity equation, raises an interesting question as to which quantities can we expect to have similar statistics between a coarse-resolution simulation run with an ML parameterization and the coarse-grained output of the high-resolution simulation on which the parameterization was trained, even in the limit of a perfect parameterization.

Our offline results suggest that for some outputs the non-local parameterizations can be more accurate than a single-column parameterization at all latitudes, but in this work we mostly focused on characterizing cases with improvement at mid-latitudes, and further work is needed to explain the reasons for improvements in the tropics. Furthermore, we focused only on spatial non-locality and future studies should investigate the opportunity of using non-local information in time (e.g., Han et al. (2020)) which may be related to non-locality in the horizontal for propagating weather systems. Finally, our focus here is on investigating the potential of using non-local inputs (and the related issue of using the local divergence or vertical velocity as an input) and understanding the situations in which improved prediction occurs, and we leave to future work the important next step of implementing and testing such parameterization approaches in climate-model simulations. We note that because the non-local inputs we used are in a close neighborhood, it is expected that they could be incorporated in parameterizations to improve simulations without greatly increasing computational expense. Another potential route forward is to adapt traditional (physics-based) parameterizations to take advantage of non-local inputs or horizontal gradients.

Data Availability Statement

Code and data used in this study are available at https://doi.org/10.5281/zenodo.6672908.

Acknowledgments
We thank Bill Boos for the output from the high-resolution simulation and Caroline Muller for the code to calculate CAPE consistent with SAM. This research received support by the generosity of Eric and Wendy Schmidt by recommendation of the Schmidt Futures as part of its Virtual Earth System Research Institute (VESRI). PW acknowledges support from NSF 1906719. PAO'G acknowledges support from NSF AGS 1749986. The authors appreciate comments from the three anonymous reviewers on the role of the vertical velocity and the degree of non-locality in subgrid parameterizations.

References

Arakawa, A., & Lamb, V. R. (1977). Computational design of the basic dynamical processes of the UCLA general circulation model. *General Circulation Models of the Atmosphere*, *17*(Supplement C), 173–265.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One*, *10*(7), 1–46. https://doi.org/10.1371/journal.pone.0130140

Back, L. E., & Bretherton, C. S. (2009). On the relationship between SST gradients, boundary layer winds, and convergence over the tropical oceans. *Journal of Climate*, *22*(15), 4182–4196. https://doi.org/10.1175/2009jcli2392.1

Bolton, T., & Zanna, L. (2019). Applications of deep learning to ocean data inference and subgrid parameterization. *Journal of Advances in Modeling Earth Systems*, *11*(1), 376–399. https://doi.org/10.1029/2018ms001472

Bony, S., Stevens, B., Frierson, D. M. W., Jakob, C., Kageyama, M., Pincus, R., et al. (2015). Clouds, circulation and climate sensitivity. *Nature Geoscience*, *8*(4), 261–268. https://doi.org/10.1038/ngeo2398

Boos, W. R., Fedorov, A., & Muir, L. (2016). Convective self-aggregation and tropical cyclogenesis under the hypohydrostatic rescaling. *Journal of Advances in Modeling Earth Systems*, *73*(2), 525–544. https://doi.org/10.1175/jas-d-15-0049.1

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

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

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019ms001711

Chen, T.-C., Yau, M. K., & Kirshbaum, D. J. (2018). Assessment of conditional symmetric instability from global reanalysis data. *Journal of the Atmospheric Sciences*, *75*(7), 2425–2443. https://doi.org/10.1175/jas-d-17-0221.1

Emanuel, K. A., David Neelin, J., & Bretherton, C. S. (1994). On large-scale circulations in convecting atmospheres. *Quarterly Journal of the Royal Meteorological Society*, *120*(519), 1111–1143. https://doi.org/10.1002/qj.49712051902

Fedorov, A. V., Muir, L., Boos, W. R., & Studholme, J. (2019). Tropical cyclogenesis in warm climates simulated by a cloud-system resolving model. *Climate Dynamics*, *52*(1–2), 107–127. https://doi.org/10.1007/s00382-018-4134-2

Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G., & Yacalis, G. (2018). Could machine learning break the convection parameterization deadlock? *Geophysical Research Letters*, *45*(11), 5742–5751. https://doi.org/10.1029/2018gl078202

George, G., Stevens, B., Bony, S., Klingebiel, M., & Vogel, R. (2021). Observed impact of mesoscale vertical motion on cloudiness. *Journal of the Atmospheric Sciences*, *78*, 2413–2427. https://doi.org/10.1175/jas-d-20-0335.1

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A moist physics parameterization based on deep learning. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002076. https://doi.org/10.1029/2020ms002076

Houze, R. A., Jr. (2004). Mesoscale convective systems. *Reviews of Geophysics*, *42*(4), RG4003. https://doi.org/10.1029/2004rg000150

Khairoutdinov, M. F., & Randall, D. A. (2003). Cloud resolving modeling of the ARM summer 1997 IOP: Model formulation, results, uncertainties, and sensitivities. *Journal of the Atmospheric Sciences*, *60*(4), 607–625. https://doi.org/10.1175/1520-0469(2003)060<0607:crmota>2.0.co;2

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

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*, 485913. https://doi.org/10.1155/2013/485913

Kuang, Z., Blossey, P. N., & Bretherton, C. S. (2005). A new approach for 3D cloud-resolving simulations of large-scale atmospheric circulation. *Geophysical Research Letters*, *32*(2), L02809. https://doi.org/10.1029/2004GL021024

Kuo, H. L. (1974). Further studies of the parameterization of the influence of cumulus convection on large-scale flow. *Journal of the Atmospheric Sciences*, *31*(5), 1232–1240. https://doi.org/10.1175/1520-0469(1974)031<1232:fsotpo>2.0.co;2

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentine, P. (2020). Assessing the potential of deep learning for emulating cloud super-parameterization in climate models with real-geography boundary conditions. arXiv preprint arXiv:2010.12996.

Muller, C. J., O'Gorman, P. A., & Back, L. E. (2011). Intensification of precipitation extremes with warming in a cloud-resolving model. *Journal of Climate*, *24*(11), 2784–2800. https://doi.org/10.1175/2011jcli3876.1

Neale, R. B., & Hoskins, B. J. (2000). A standard test for AGCMs including their physical parametrizations. II: Results for the Met Office Model. *Atmospheric Science Letters*, *1*(2), 108–114. https://doi.org/10.1006/asle.2000.0020

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548–2563. https://doi.org/10.1029/2018ms001351

Ooyama, K. (1969). Numerical simulation of the life cycle of tropical cyclones. *Journal of the Atmospheric Sciences*, *26*(1), 3–40. https://doi.org/10.1175/1520-0469(1969)026<0003:nsotlc>2.0.co;2

Palmer, T. N. (2001). A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Quarterly Journal of the Royal Meteorological Society*, *127*(572), 279–304. https://doi.org/10.1002/qj.49712757202

Pathak, R., Sahany, S., Mishra, S. K., & Dash, S. K. (2019). Precipitation biases in CMIP5 models over the South Asian region. *Scientific Reports*, *9*(1), 9589. https://doi.org/10.1038/s41598-019-45907-4

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

Schneider, T., Teixeira, J., Bretherton, C. S., Brient, F., Pressel, K. G., Schär, C., & Siebesma, A. P. (2017). Climate goals and computing the future of clouds. *Nature Climate Change*, *7*(1), 3–5. https://doi.org/10.1038/nclimate3190

Sherwood, S. C., Bony, S., & Dufresne, J.-L. (2014). Spread in model climate sensitivity traced to atmospheric convective mixing. *Nature*, *505*(7481), 37–42. https://doi.org/10.1038/nature12829

Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on applications of computer vision (WACV)* (pp. 464–472). IEEE.

Stensrud, D., Coniglio, M., Knopfmeier, K., & Clark, A. (2015). Numerical models—Model physics parameterization. In G. R. North, J. Pyle, & F. Zhang (Eds.), *Encyclopedia of atmospheric sciences* (2nd ed., pp. 167–180). Academic Press.

Wilcox, E. M., & Donner, L. J. (2007). The frequency of extreme rain events in satellite rain-rate estimates and an atmospheric general circulation model. *Journal of Climate*, *20*(1), 53–69. https://doi.org/10.1175/jcli3987.1

Yuval, J., & O'Gorman, P. A. (2020). Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, *11*(1), 3295. https://doi.org/10.1038/s41467-020-17142-3

Yuval, J., & O'Gorman, P. A. (2021). Neural-network parameterization of subgrid momentum transport in the atmosphere [preprint]. *Earth and Space Science Open Archive*. https://doi.org/10.1002/essoar.10507557.1

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. *Geophysical Research Letters*, *48*(6), e2020GL091363. https://doi.org/10.1029/2020gl091363

19422466, 2022, 10, Downloaded from https://agupubs.onlinelibrary.wiley.com/doi/10.1029/2022MS002984, Wiley Online Library on [19/10/2022]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License