JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2023MS003668

Special Section:
Machine learning application to Earth system modeling

Neural Network Parameterization of Subgrid-Scale Physics From a Realistic Geography Global Storm-Resolving Simulation

Oliver Watt-Meyer[1], Noah D. Brenowitz[2], Spencer K. Clark[1,3], Brian Henn[1], Anna Kwa[1], Jeremy McGibbon[1], W. Andre Perkins[1], Lucas Harris[3], and Christopher S. Bretherton[1]

[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA, [2]NVIDIA Corporation, Santa Clara, CA, USA, [3]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ, USA

Key Points:

- Effective sources of heat and moisture are computed from a global storm-resolving simulation accounting for semi-resolved dynamics
- A neural network is trained to predict columns of the effective sources using profiles of temperature and specific humidity
- When used online, stable month-long simulations are possible although skill is not yet comparable to a previous corrective approach

Correspondence to:
O. Watt-Meyer,
oliverwm@allenai.org

Citation:
Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., et al. (2024). Neural network parameterization of subgrid-scale physics from a realistic geography global storm-resolving simulation. *Journal of Advances in Modeling Earth Systems*, *16*, e2023MS003668. https://doi.org/10.1029/2023MS003668

Received 9 FEB 2023
Accepted 8 DEC 2023

Abstract Parameterization of subgrid-scale processes is a major source of uncertainty in global atmospheric model simulations. Global storm-resolving simulations use a finer grid (less than 5 km) to reduce this uncertainty by explicitly resolving deep convection and details of orography. This study uses machine learning to replace the physical parameterizations of heating and moistening rates, but not wind tendencies, in a coarse-grid (200 km) global atmosphere model, using training data obtained by spatially coarse-graining a 40-day realistic geography global storm-resolving simulation. The training targets are the three-dimensional fields of effective heating and moistening rates, including the effect of grid-scale motions that are resolved but imperfectly simulated by the coarse model. A neural network is trained to predict the time-dependent heating and moistening rates in each grid column using the coarse-grained temperature, specific humidity, surface turbulent heat fluxes, cosine of solar zenith angle, land-sea mask and surface geopotential of that grid column as inputs. The coefficient of determination R^2 for offline prediction ranges from 0.4 to 0.8 at most vertical levels and latitudes. Online, we achieve stable 35-day simulations, with metrics of skill such as the time-mean pattern of near-surface temperature and precipitation comparable or slightly better than a baseline simulation with conventional physical parameterizations. However, the structure of tropical circulation and relative humidity in the upper troposphere are unrealistic. Overall, this study shows potential for the replacement of human-designed parameterizations with data-driven ones in a realistic setting.

Plain Language Summary Numerical models used for projecting climate change impacts must use ad-hoc assumptions about the effects of unresolved small-scale processes. These assumptions contribute to uncertainty in predicting how rainfall and temperature will change in the future. Expensive fine-grid simulations which eliminate the need for some of these assumptions are possible to run for shorter (month-to year-long) duration. We use such a simulation to train a data-driven representation of the effects of processes, like clouds, which are poorly simulated by a cheaper coarse-grid model. The data-driven representation (a neural network) predicts rates of temperature and moisture change in each column using inputs from that grid column. This approach has been previously shown to work for models with idealized boundary conditions, but not for the realistic setting we use. When this neural network is used in a coarse-resolution model, the realism of many global skill metrics is as good or better than a baseline model with traditional representation of small-scale processes. However, some features are degraded, such as the time-evolving pattern of rainfall in the tropics and humidity in the upper atmosphere. This work is a first step toward the use of data-driven representations of unresolved processes in realistic global atmospheric models.

1. Introduction

The parameterization of subgrid-scale phenomena, such as clouds and turbulence, is a fundamental challenge in climate modeling. These processes occur at spatial and temporal scales that are too small to be adequately resolved by the atmospheric models used for decadal and longer climate forecasts. As a result, human-designed parameterization schemes are used to represent their effects on the grid-resolved atmospheric state. Traditionally, these schemes have been based on empirical relationships and physical closure assumptions derived from theoretical considerations, observations and process modeling studies. While these methods have been fairly successful, they struggle to accurately represent the complex and nonlinear interactions within the atmosphere important in processes such as cumulus convection (Derbyshire et al., 2004; Guichard et al., 2004) or flow through

mountainous terrain (Pithan et al., 2016; van Niekerk et al., 2016). Biases in the simulated climate can often be attributed to parameterization choices (e.g., Chen et al., 2007; Song & Zhang, 2018; Woelfle et al., 2018; Zhao, 2014) and they are a leading cause of uncertainty in climate change projections (Caldwell et al., 2016). As a result, there is a growing interest in the use of machine learning techniques to improve the parameterization of these processes.

Machine learning algorithms can learn complex patterns and relationships from data, making them well-suited for this task. Framing parameterizations in terms of machine learning algorithms allows the use of the rapidly growing suite of software tools for training and inference, potentially leading to improvements in computational efficiency at runtime on modern computing architectures like GPUs. This would be particularly advantageous if machine learning parameterizations could run in concert with a dynamical core that runs efficiently on GPUs (e.g., Dahm et al., 2022). It can also enable novel strategies for the prediction of subgrid-scale phenomena, such as the use of data from neighboring columns (P. Wang et al., 2022).

Our approach is to train machine learning models on high-resolution simulation data to learn the relationships between these processes and the large-scale state, and to use these models to parameterize the processes in a lower-resolution numerical model. This approach has been demonstrated in aquaplanet simulations (Brenowitz & Bretherton, 2018; Brenowitz & Bretherton, 2019; Yuval et al., 2021; Yuval & O'Gorman, 2020) but not in atmospheric models with realistic boundary conditions and a land-surface model. In this paper, we show that a neural network can be used to skillfully predict the heating and moistening rates due to all subgrid-scale processes (i.e., the apparent sources of heat and moisture, typically predicted by human-designed parameterization schemes) as diagnosed by coarse-graining a global storm-resolving simulation. This is the first study, to our knowledge, to show stable online simulations with reduced pattern error of surface precipitation rate and near-surface temperature using a coarse-graining based machine learning parameterization in a model with realistic boundary conditions. A previous approach of replacing all parameterizations except clear-sky radiation and turbulence with machine learning did enable 10-day simulations (Brenowitz et al., 2020). This study differs from Brenowitz et al. (2020) in the definition of the machine learning target to account for semi-resolved dynamics and it uses a larger set of input variables for prediction. Finally, learning the tendencies of embedded cloud-resolving models (Gentine et al., 2018; Rasp et al., 2018) has also shown promise, including for realistic boundary conditions (Y. Han et al., 2020; Mooers et al., 2021; X. Wang et al., 2022; Y. Han et al., 2023).

2. Methods

The goal of this study is to machine learn the apparent sources of heat and moisture from a global storm-resolving simulation. These apparent sources (Yanai et al., 1973) represent the total unresolved sources at a given location. Once such a model is trained, it can be used in place of the physical parameterizations of a coarse-resolution atmospheric general circulation model. We do not attempt to machine-learn momentum tendencies (e.g., due to gravity wave drag or boundary layer turbulence) or any cloud condensates (i.e., cloud water/ice and rain/snow/graupel). Formulating the apparent momentum source consistent with our pressure-level coarse-graining approach is challenging in mountainous regions, because of the effect of fine-grid pressure drag on coarse-grained horizontal velocity tendencies at coarse pressure levels that intersect the fine-grid terrain (this diagnostic was not computed or saved). Instead, we use the coarse grid model's parameterizations to account for the effects of subgrid-scale processes on the winds. Although we do not directly predict cloud condensates their effects on heating and moistening rates are accounted for in our target data set.

2.1. Atmospheric Models

To produce training data, we use the GFDL X-SHiELD global storm-resolving model to make a 40-day simulation with specified sea surface temperature (SST) initialized on 1 August 2016 (Cheng et al., 2022; Harris et al., 2023; Stevens et al., 2019). This is the same fine-grid simulation used in Bretherton et al. (2022) (hereafter B22) and more details of the configuration can be found therein. X-SHiELD uses the FV3 dynamical core (Harris et al., 2021) run at C3072 resolution (approximately 3-km) and 79 vertical levels with physics modified from the Global Forecast System (GFS) (Zhou et al., 2019). The time step for the physics parameterizations is 225 s. This simulation was run on 13,824 cores on NOAA's Gaea computing system, taking approximately 140 min of wall clock time per simulated day. For the coarse resolution simulation, we use the related model FV3GFS, which is

the atmospheric component of the Unified Forecast System (UFS) (UFS Community, 2020). The top-level routines of the Fortran model are wrapped by Python (McGibbon et al., 2021) to enable doing inference with ML models online, as well nudging and prescribing of physics tendencies, as discussed further in Section 2.3. The coarse model is run at C48 resolution (approximately 200-km), so there is a 64x coarse-graining between the fine and coarse resolutions. The time step for the coarse model is 900 s, four times greater than the fine-resolution model's time step. The simulations were run on Google Cloud Platform, typically using 24 CPUs with a duration of about 17 min of wall clock time per simulated day. Doing neural network inference online (as described in later sections) increased runtime by approximately 20%.

Unlike B22, in this study we specify the sea surface temperature to equal the coarsened value from the fine-resolution simulation. In B22, the sea surface temperature was initialized with the coarsened fine-resolution values but slowly relaxed toward a climatology. This results in a slight 0.02 K decrease in ocean- and time-mean surface temperature for simulations in this study compared to B22. Locally, the time-mean differences are mostly less than 1 K except at high latitudes where they reach up to 3 K. This configuration change is one reason for the slight difference in the baseline model performance between this study and B22.

2.2. Computing Apparent Sources of Heat and Moisture

We use a global storm-resolving simulation to generate training data for learning the apparent sources of heat and moisture (Yanai et al., 1973). Following B22, we coarse-grain along surfaces of constant pressure. Since the X-SHiELD and FV3GFS models use a terrain-following vertical coordinate, this requires a vertical interpolation of fine-resolution fields to the coarse-grained pressure surfaces before doing horizontal coarse-graining, as detailed in Section 2.4 of B22.

We define the apparent sources as the coarse-grained tendency due to the physics parameterizations of the fine-resolution model plus the vertical eddy-flux convergence:

$$Q_T = \overline{\left.\frac{\partial T}{\partial t}\right|_{phys}} - \frac{\partial \overline{\omega' T'}}{\partial p} \tag{1}$$

$$Q_q = \overline{\left.\frac{\partial q}{\partial t}\right|_{phys}} - \frac{\partial \overline{\omega' q'}}{\partial p} \tag{2}$$

where $\overline{\cdot}$ represents horizontal coarse-graining and $\omega' = \omega - \overline{\omega}$ (similarly for temperature T and specific humidity q). The first term on the right-hand sides of Equations 1 and 2 is the coarsened tendency due to all physics parameterizations. For specific humidity, this is the sum of tendencies due to radiation, shallow convection, turbulence, and microphysics, including the fast saturation adjustment within the dynamical core (Zhou et al., 2019). For temperature, we additionally include the tendency due to the "meteorological" nudging of temperature in the fine-resolution simulation (see Section 2.1 of B22). Given its gentle 24-hr nudging timescale, this latter term is typically much smaller than the others. No nudging of specific humidity was done in the fine-resolution simulation, and so there is no corresponding term in $\left.\frac{\partial q}{\partial t}\right|_{phys}$.

Unlike our previous corrective ML approach which primarily required coarse-graining temperature, humidity, pressure thickness and horizontal winds (Bretherton et al., 2022), the apparent sources depend strongly on the coarsened vertical pressure velocity ω. This exposes some issues with our coarsening algorithm in the lowest model layer over ocean (see Appendix A for details). Nevertheless, the prescribed physics simulation described in the next section is able to compensate for the resulting biases in coarse-grained pressure velocity.

Saving all of the desired data at the native C3072 X-SHiELD resolution before coarse-graining offline would consume an unwieldy 4PB of storage, so we instead do an eight-fold coarse-graining to C384 resolution and 4-step temporal averaging during simulation. The final stored output is about 70 TB. We do a subsequent coarse-graining to C48 resolution offline. Some of these saved outputs were not directly used in the calculation of $Q_{T,q}$, but were used for ad-hoc analysis to better understand the contribution of different processes to these terms. The eddy flux term due to eddies between C3072 and C384 is also computed online and then added to the offline calculation of eddy flux due to eddies between C384 and C48 resolution:

$$\overline{\omega' T'} = \overline{\widehat{\omega^* T^*}} + \overline{\hat{\omega}\hat{T}} - \overline{\omega}\,\overline{T} \tag{3}$$

where $\hat{\cdot}$ represents coarse-graining from C3072 to C384 and $\cdot^*$ is the corresponding eddy (and similarly for the eddy flux of moisture). In addition to a horizontal coarse-graining, we do an online time-averaging over the four fine-resolution timesteps that make up each coarse model timestep (e.g., for the tendencies due to the fine-grid parameterizations). For the eddy fluxes, the $\overline{\widehat{\omega^* T^*}}$ term is computed at native time resolution then averaged over time while the other two terms on the right hand side of Equation 3 are computed using data at 900 s resolution. Ideally these two terms would be computed on the fine-grid model's time resolution, but the necessary data was not saved from the fine-grid simulation. The resultant data set for $Q_{T,q}$ has 900 s temporal resolution, equivalent to the physics time step used in the coarse-grid model simulations described in the next section.

2.3. Prescribed Physics Simulation

In principle, the apparent sources Q_T and Q_q include the effects of all unresolved motions on scales finer than C48 (approximately 200 km). However, the dynamical core does not accurately advect disturbances with wavelengths comparable to its grid resolution—the errors of which strongly depend on the numerical method used, as discussed in Chapter 3 of Durran (2010). Therefore, as noted by B22, there are "semi-resolved" dynamics of the fine-grid simulation which can be represented on the coarse grid but are nevertheless not accurately simulated by the coarse model advection scheme. We call the resulting errors in representing the coarsened fine-grid advective tendencies of heat and moisture the "apparent dynamics tendencies," labeled ΔQ_T^{dyn} and ΔQ_q^{dyn} (and whose calculation will be described shortly). These can also account for any algorithmic differences between the dynamical cores of the fine and coarse simulations or errors in the estimation of Q_T and Q_q, for example, due to issues with coarse-graining. Unlike the apparent sources, the apparent dynamics tendencies depend on the coarse model and how it is configured to track the evolving coarsened fine model state. The apparent dynamics tendencies must be added to the apparent sources to get "effective" sources that, in concert with the coarse-model dynamics, can reproduce the evolution of the fine-grid reference simulation:

$$Q_T^+ = Q_T + \Delta Q_T^{dyn}, \tag{4}$$

$$Q_q^+ = Q_q + \Delta Q_q^{dyn}. \tag{5}$$

These effective sources $Q_{T,q}^+$ will be our machine-learning target.

To compute the apparent dynamics tendencies $\Delta Q_{T,q}^{dyn}$, we do a prescribed-physics coarse-model simulation (to be specified mathematically in Equation 6 and schematically described in Figure 1). The tendencies of temperature and specific humidity due to physical parameterizations in the coarse model are overwritten online by the apparent sources computed from the fine-grid reference simulation. All other prognostic variables are updated by the normal physical parameterizations. This simulation is initialized from a coarsened snapshot of the fine-grid run. An obvious approach would be to calculate the apparent dynamics tendency at each time as a single time step tendency difference between the coarsened-fine reference output and the coarse model dynamics initialized from that reference output and forced by the apparent sources (e.g., Brenowitz & Bretherton, 2019). Similar to B22, the approach fails here due to initialization shock of the coarse model, in the form of strong spurious vertical motions that contaminate the apparent dynamics tendencies and take hours to settle down.

Thus, like B22, we instead nudge the coarse model state to the time-dependent reference state with a timescale long enough to minimize spurious vertical motions but short enough to keep the evolving atmospheric state close to the desired reference, so that we can associate the apparent sources with atmospheric state consistently with the fine-grid reference simulation. Specifically, we nudge the temperature, specific humidity, horizontal winds and pressure thickness, all using a 3-hr timescale.

The tendency of a scalar quantity a in the prescribed physics simulation is described by

$$\frac{\partial a}{\partial t} = -\nabla \cdot (\mathbf{v} a) + \underbrace{\tilde{Q}_a^p + \Delta Q_a^{dyn}}_{Q_a^+ \text{if } a \in [T,q]} \tag{6}$$

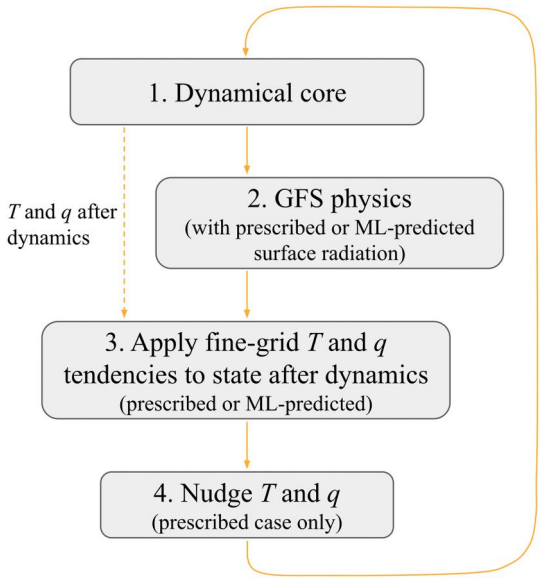

Figure 1. The control flow for both coarse grid simulations: prescribed physics run (Section 2.3) and online run (Section 2.5). The dynamical core (step 1) and standard GFS physics module (step 2) are run in sequence, with the surface radiation within the physics either prescribed from the fine-grid data or predicted by a standalone ML model in the online run. Then the column tendencies of T and q due to subgrid processes are applied (step 3). These tendencies are either prescribed from the fine-grid run ($Q_{T,q}$) or are ML-predicted $\left(\hat{Q}^{+}_{T,q}\right)$. The tendencies are applied to the T, q state after dynamics, thereby overwriting the GFS physics updates for these variables. Finally, for the prescribed case only, the T and q variables are nudged toward the fine-grid simulation data (step 4).

where $\tilde{Q}^{p}_{a}$ is the tendency due to the coarse-resolution physics parameterizations, Q^{p}_{a}, for all variables except temperature and specific humidity, for which it is replaced by the apparent source from the fine-grid reference simulation:

$$\tilde{Q}^{p}_{a} = \begin{cases} Q_a & \text{if } a \in [T,q] \\ Q^{p}_{a} & \text{otherwise.} \end{cases} \tag{7}$$

In this nudged methodology, the apparent dynamics tendency is the nudging tendency:

$$\Delta Q^{dyn}_{a} = -\frac{a-\overline{a}}{\tau}, \tag{8}$$

and the nudging keeps the coarse state a close to $\overline{a}$, defined as the coarsened a from the reference fine-grid simulation. See Figure 1 for a schematic of the control flow in the prescribed physics simulation.

A very similar, but not identical, "dynamics nudging tendency" was discussed in Section 7 of B22. There it was computed as the difference between the apparent sources $Q_{T,q}$ and the sum of the coarse model's physics parameterization tendency plus the nudging tendency in a nudged coarse model run without any prescribed physics. This was a natural choice for the corrective ML approach used there.

2.4. Machine Learning Data set and Architecture

The machine learning targets (outputs) are the effective sources of temperature or humidity in all grid columns and time steps in the training data set. We make the approximation of column-local physics. That is, the ML inputs used to predict Q^{+}_{T} and Q^{+}_{q} in a grid column are the vertical profiles of temperature and specific humidity and five additional scalar inputs, the cosine of the solar zenith angle, surface geopotential, land-sea mask and the surface sensible and latent heat fluxes, in that grid column. During training, the temperature and specific humidity are taken from the coarsened fine-grid data set, while the scalar inputs are from the coarse-grid prescribed physics simulation. For use as machine learning inputs, the surface turbulent heat fluxes are computed by the coarse model's piggy-backed physics parameterizations (Figure 1). Using nearby columns as additional inputs could lead to more accurate predictions (P. Wang et al., 2022), in particular for the apparent dynamics tendency. However, it is not clear whether the offline improvements demonstrated in predicted moisture and temperature tendencies would translate to online skill improvements, so here we retain the single-column approximation for simplicity.

We use a neural network (multi-layer perceptron) for predictions. It has three hidden layers and width of 128. For regularization, we add Gaussian noise with a standard deviation of 0.1 to each layer's input during training. These hyperparameters were chosen by hand, balancing offline skill and short-term (2-day) online stability. Inputs and outputs are normalized by subtracting their means and dividing by the standard deviation. A separate mean and standard deviation is computed for each vertical level. The loss function uses mean-squared error and is computed on normalized outputs. For the inputs only, a small value $\epsilon = 1 \times 10^{-7}$ is added to the standard deviation before normalization. Models are trained for 10 different random seeds, and we primarily present results from the random seed which gave the best online performance (see Section 5.1 for more detail).

The fine-grid reference simulation used for this study (and B22) spans 40 days, initialized at 0 UTC 1 August 2016, following the DYAMOND specification (Stevens et al., 2019). The prescribed physics run (see Section 2.3) is initialized at 0 UTC 5 August 2016 (to allow prior spin-up of the fine-grid simulation) and run for 35 days. We randomly sample 150 timesteps from the 35-day prescribed physics run to compute $Q^{+}_{T,q}$ for training. Given $6 \cdot 48 \cdot 48$ samples per timestep, this is about 2.07M samples. The model is trained with a learning rate of 5×10^{-5} for

400 epochs using the Adam optimizer (Kingma & Ba, 2014). The training for this simple neural network was done on a CPU and took approximately 1 hour. For validation and testing, a further 50 timesteps are randomly selected from the remaining timesteps not used for training.

Following previous work, to ensure stability of online simulations we did not use the uppermost 30 model levels (with typical air pressures less than 270 hPa) of the $\overline{T}$ and $\overline{q}$ profiles as inputs (Brenowitz & Bretherton, 2019; Clark et al., 2022).

We additionally train a model to predict surface radiative fluxes and use these predictions to force the land-surface model (following B22, Clark et al. (2022) and Kwa et al. (2022)). Specifically, using the method described in Section 2.5 of Clark et al. (2022), we predict the shortwave transmissivity of the atmospheric column—defined as the ratio of the downward surface shortwave radiative flux to the downward top-of-atmosphere shortwave flux—and the downward surface longwave radiative flux. Then the coarse model's surface albedo and downward top-of-atmosphere shortwave flux can be used to derive the downward and net shortwave surface radiative flux (necessary inputs to the land-surface model) along with the downward longwave surface radiative flux which is directly predicted by the ML model. The inputs are the same as those used for the neural network that predicts the effective sources except we do not use sensible or latent heat flux. The architecture (depth, width) and optimization (loss function, learning rate) are also the same as for the effective source neural network. The predicted shortwave transmissivity is limited to between 0 and 1 by applying $\min(\max(0, y), 1)$ to the output layer and the predicted downward longwave radiative flux is forced to be positive in a similar fashion.

2.5. Online Simulation

To use the trained model online, we follow the same procedure as in the prescribed-physics simulation (Section 2.3), but with the apparent sources $Q_{T,q}$ at each physics timestep replaced by their ML-inferred estimates $\hat{Q}^{+}_{T,q}$ when updating T and q (Figure 1). No nudging is done in the online simulation.

We still use the coarse-model physical parameterizations to predict momentum tendencies and to diagnose latent and sensible heat fluxes, which are used as ML inputs. Although the coarse-model physics also prognoses the condensate fields, these are not used as inputs for the ML and hence do not directly affect the simulation, at least to the extent that the tendencies of momentum from the coarse-resolution physics do not depend on condensates. That is, clouds are treated implicitly, affecting the simulation only indirectly by modifying the effective sources $Q^{+}_{T,q}$ and the predicted downwelling surface radiation.

The machine learning parameterization interacts with the land and ocean surface in three ways. First, the sensible and latent heat fluxes from the coarse model's physics are used as inputs for the ML predictions. Second, a surface precipitation rate is diagnosed from the evaporation and predicted column-integrated atmospheric moistening following conservation of water:

$$\hat{P} = E - \langle \hat{Q}^{+}_{q} \rangle = E - \frac{1}{g} \int_{p_{top}}^{p_s} \hat{Q}^{+}_{q} \, dp. \tag{9}$$

This surface precipitation rate is passed to the land-surface model, and can feed back on the ML parameterization via changes in soil moisture and hence the surface turbulent surface fluxes. Third, the predicted downward surface radiative fluxes are used to force the land model.

For comparison, in Section 5 we will also compare our results with a baseline (no-ML) coarse-resolution simulation and with a nudge-to-fine simulation in which the ML only predicts corrections to the coarse-resolution parameterizations instead of replacing them. The latter approach has shown skill in various contexts (Bretherton et al., 2022; Clark et al., 2022; Kwa et al., 2022).

For evaluation of time-mean diagnostics, we run a 35-day simulation initialized at 0 UTC 5 August 2016 (the same period used for the prescribed physics run). For evaluation of weather timescale metrics, we run four 10-day forecasts, initialized at 0 UTC of August 5, 13, 21, and 29, 2016. Differently than B22, here we use the full 35 days of simulation when computing the time-mean RMSEs and biases. In the previous study, the first 6 days of simulation were discarded as a spin-up period.

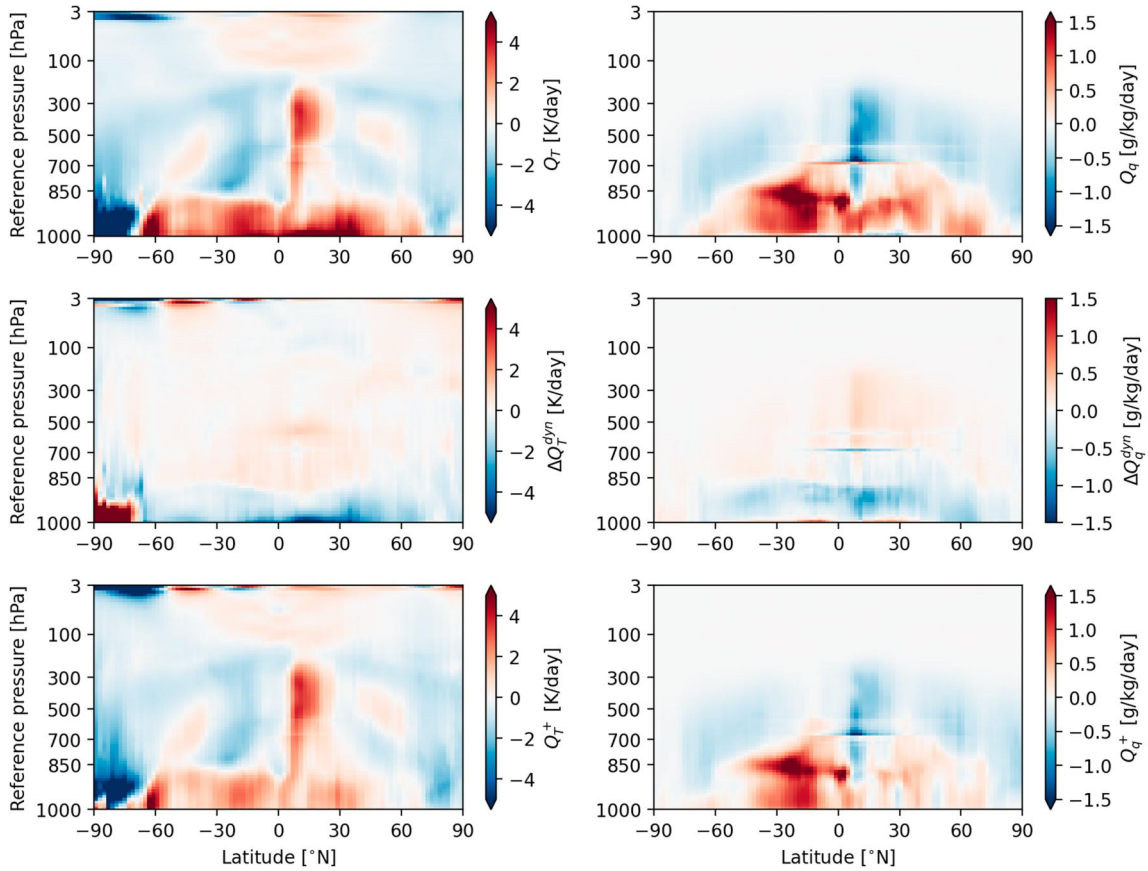

Figure 2. Zonal- and time-mean tendencies from prescribed physics run: (left) temperature tendency and (right) specific humidity tendency. Rows show the (top) apparent sources $Q_{T,q}$, (middle) apparent dynamics tendency $\Delta Q_{T,q}^{dyn}$ and (bottom) their sum, the effective sources $Q_{T,q}^{+}$. Data is zonally averaged on model levels, but is plotted in terms of reference pressure assuming a surface pressure of 1,000 hPa.

3. Prescribed Physics Simulation Results

The 35-day time-mean zonal averages of the apparent sources Q_T and Q_q (prescribed from the coarsened fine-grid reference outputs) are shown in the top row of Figure 2. As expected, we see strong heating and drying in the free troposphere in regions of convection (approximately 5°N to 20°N given the average is over late boreal summer); cooling in most other regions of the free troposphere and heating and moistening in the boundary layer at most latitudes. Some features are more unexpected: for example, the sharp change in Q_q at single vertical levels (at approximately 550 hPa and 700 hPa). The lower and more prominent of these artifacts can be traced to a threshold placed on the shallow convection scheme in the fine-grid reference simulation (J. Han & Pan, 2011; J. Han et al., 2017) which limits the vertical extent at which the scheme can be active and is a particular issue when running without a deep convection scheme. Another unexpected feature is large positive values of Q_T in the lowermost grid levels in the Northern Hemisphere subtropics. This persists even if averaged only across columns over ocean (not shown). There is also a sharp gradient in Q_q over the lowest few model levels (moistening above but drying at the very lowest level).

The prescribed physics simulation provides the apparent dynamics tendency $\Delta Q_{T,q}^{dyn}$. The middle row of Figure 2 shows that at all latitudes, the mean apparent dynamics tendency tends to be much smaller in magnitude than the apparent sources except near the surface. In general the apparent dynamics tendencies slightly warm and moisten the free troposphere and cool and dry the boundary layer. In some isolated regions, the nudging tendency can be of similar magnitude to the apparent source: near the surface in the tropics, subtropics and Antarctica, and at specific model levels where there is a sharp jump in Q_q. In these cases, the apparent dynamics tendency tends to compensate the apparent source. Their sum, our ML target the effective source, is therefore vertically smoother and in some regions has significantly smaller magnitude than the apparent source (bottom row of Figure 2).

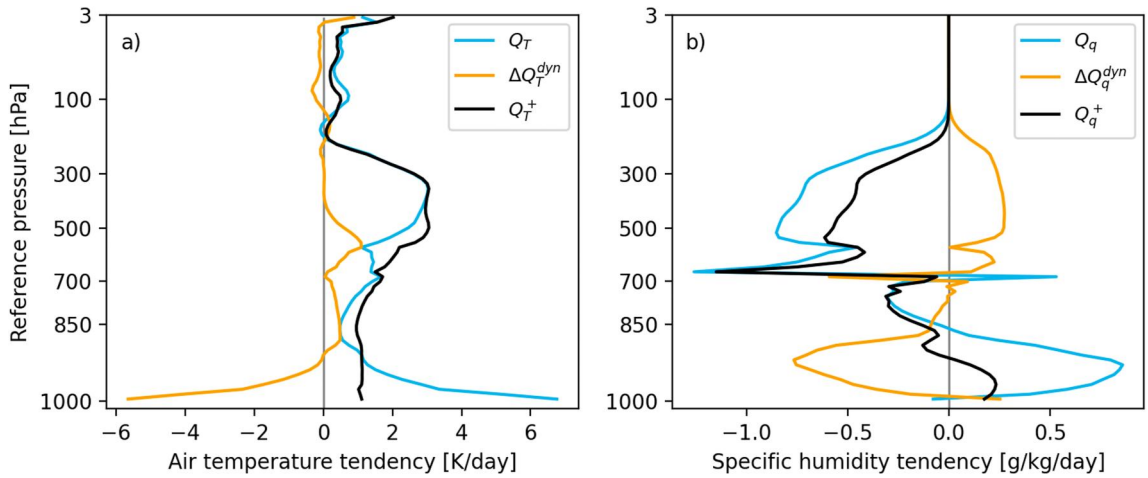

Figure 3. Tropical ascent region and time-mean tendencies from prescribed physics run: (left) temperature tendency and (right) specific humidity tendency. Spatial average is from 5°N to 20°N. Data is averaged on model levels, but is plotted in terms of reference pressure assuming a surface pressure of 1,000 hPa.

Figure 3 shows the vertical profiles of heating and moistening rates averaged over the tropical convective belt from 5 to 20°N. In the boundary layer, we again see the aforementioned compensation between the apparent dynamics tendency and the apparent sources, leading to an effective source that is relatively uniform in the vertical. Further examination of the calculation of the apparent sources in the lowest model layers has revealed artifacts from the coarse-graining approach implemented in the fine-grid reference simulation (see Appendix A); these likely help lead to the very strong vertical gradients in Q_T and Q_q near the surface. The effective sources (black curves in Figure 3) that are our ML target do not have this large gradient near the surface because they mainly depend on the evolution of $\overline{T}$ and $\overline{q}$ in the fine-grid reference simulation.

In the upper troposphere (150–500 hPa) it is evident that the coarse resolution model's dynamics does not vertically advect enough moisture to compensate for the drying due to the apparent source of moisture (Figure 3b). Therefore, in the time-mean, the apparent dynamics tendency ΔQ_q^{dyn} must do moistening that is about a quarter to a third of the strength of the apparent drying. Overall, the effect of the apparent dynamics tendency is to transport heat and moisture upward, suggesting that this transport may be numerically damped in the coarse-grid simulation compared to its more accurate representation in the fine-resolution simulation.

To verify the calculation of the apparent sources, we compare the vertically integrated apparent moistening $\langle Q_q \rangle$ with the flux of moisture through the surface (evaporation minus precipitation; $E - P$). Ideally these terms would equal each other, but due to the complications of coarse-graining vertically resolved fields (see Appendix A and Section 2.4 of B22) this may not be exactly true. Nevertheless, we find close agreement between the column-integrated apparent moistening and surface moisture flux across individual timesteps and grid cells, with a squared Pearson correlation coefficient of 0.90 and a linear relationship that closely follows the 1:1 line (Figure 4a). However, the machine learning target necessary for representing the evolution of the fine-grid model is the effective source Q_q^+. As seen in Figures 2 and 3, there is a time-mean compensation of Q_q and ΔQ_q^{dyn} in most regions. This compensation is also true for individual timesteps resulting in a column-integrated effective moistening $\left\langle Q_q^+ \right\rangle$ that is typically of smaller magnitude than the corresponding $E - P$ (Figure 4b). We interpret this as strong localized atmospheric drying and latent heating events in the fine-grid simulation (e.g., tropical convection) being more difficult for the coarse resolution model to balance with the resolved coarse model dynamics, hence requiring stronger positive apparent dynamics moistening.

4. Offline Machine Learning Skill

For a machine-learning parameterization based on coarse-graining to be skillful online, it must first be skillful offline. Figure 5 shows a snapshot of offline column-integrated target and ML-predicted effective heating rates $\langle Q_T^+ \rangle$. Brackets denote a mass-weighted column integral, multiplied by c_v, the specific heat capacity of air at constant volume, to scale it to energy units appropriate for forcing the coarse model's nonhydrostatic dynamics

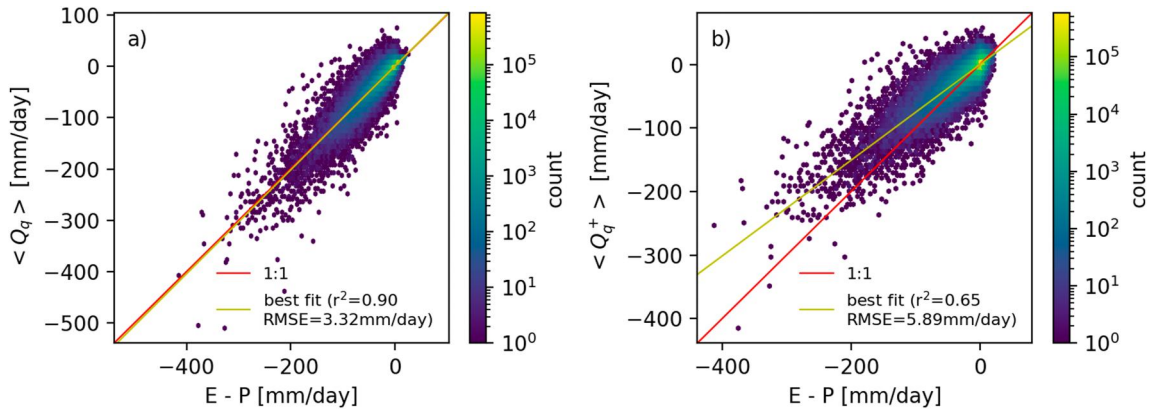

Figure 4. Hexbin plot of joint histogram between $E - P$ (evaporation minus precipitation) from fine-grid simulation and the column-integrated (a) apparent moistening $\langle Q_q \rangle$, (b) effective moistening $\left\langle Q_q^+ \right\rangle$. Data are computed on the coarse C48 resolution grid and the histogram consists of timesteps spanning the prescribed physics simulation at a 5-hourly frequency. The r^2 value shown is the Pearson correlation coefficient squared.

(see Section 9.3 of Harris et al., 2021). The strongest effective heating occurs in the Intertropical Convergence Zone (ITCZ), West Pacific warm pool and midlatitude fronts. A pattern of heating during daytime and cooling elsewhere is overlaid. In general the offline ML prediction (Figure 5b) skillfully captures both the global-scale radiative heating and cooling as well as the strong heating within convective regions. However, the predictions tend to be somewhat horizontally smoother than the test data, and in some cases the ML predicts strong heating where the true effective source of heating is small (e.g., off the south tip of Africa or south of the Aleutian islands). Nevertheless, this snapshot shows the possibility of predicting the effective sources using only the coarse-grained state as input.

To quantify offline skill, Figure 6 plots the coefficient of determination R^2 as a function of latitude and pressure (see definition in Appendix A2 of B22, but note here we use the zonal average instead of global average when computing the total sum of squares). In the troposphere, away from the highest latitudes, predictions of the apparent sources are generally skillful ($0.4 < R^2 < 0.8$), in particular for Q_T^+. This is comparable to the offline skill in Brenowitz and Bretherton (2019) (see their Figure 4) but lower than that in Yuval and O'Gorman (2020) (see their Figure 3e, noting their result is for 32x coarsening whereas we are doing 64x coarsening). Both of these previous studies used aquaplanet boundary conditions and we expect prediction to be more challenging for the full geography case given the greater non-stationarity of the data, including differences between land and ocean and changes in surface topography. In general, skill for apparent moistening Q_q^+ is poorer than for Q_T^+ and is particularly bad (below 0) south of 75°S, a cold, dry, sloped region in which the effective moisture source is small and involves a different balance of physical processes than in moister lower-latitude regions that contribute much more strongly to the ML loss function.

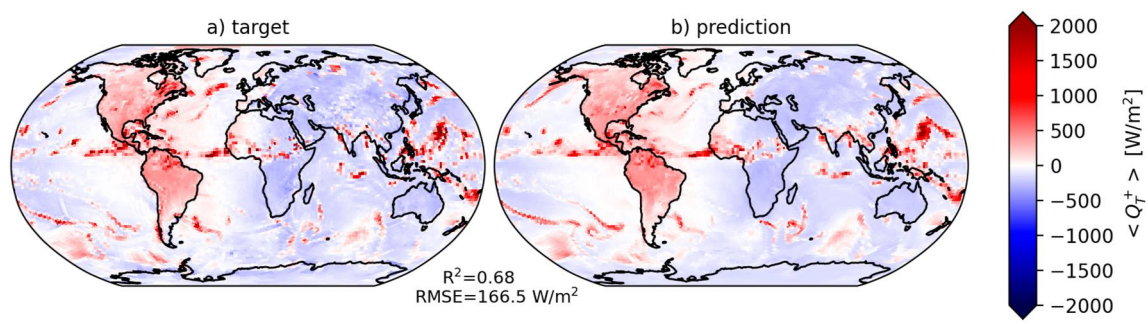

Figure 5. Column-integrated heating, $\langle Q_T^+ \rangle = \frac{c_v}{g} \int_{p_{top}}^{p_s} Q_T^+ \, dp$, at 17 UTC 2016-08-05 from (a) test data and (b) ML prediction. Text shows the area-weighted RMSE and R^2 of the column-integrated heating for this given snapshot.

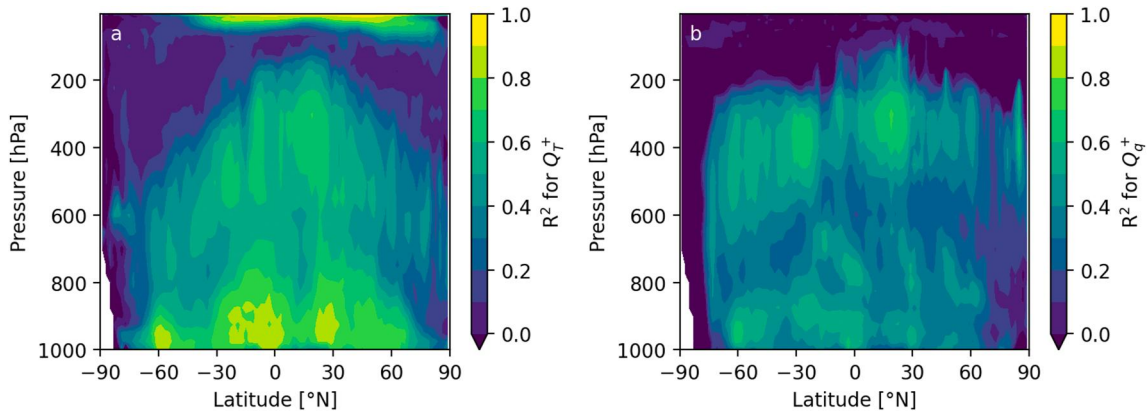

Figure 6. Offline skill R^2 for (left) Q_T^+ and (right) Q_q^+ as function of latitude and pressure.

The neural network trained to predict shortwave transmissivity and downward longwave radiative flux has high offline skill, with R^2 values of 0.97 and 0.99 for each of these outputs respectively.

5. Online ML Physics Results

The most important measure of the ML-based parameterization is its performance when it is coupled back to the dynamical core and the remaining coarse-res parameterizations (e.g., the land-surface model and parameterizations affecting momentum). In this evaluation, the coarse model is run with ML used online to predict the effective sources and update the temperature and specific humidity field; no nudging is applied. Good offline performance is important, but is no guarantee of stable or accurate behavior online (Brenowitz et al., 2020).

5.1. Random Seed Variability

One simple gauge of online performance that we have found useful is the evolution of global mean water vapor path. Given the approximate balance between evaporation and precipitation, we expect this quantity to remain roughly constant on weekly to monthly timescales. Because the dynamics can only move water vapor around and not create or destroy it, increases or decreases in the global mean water vapor path are indicative of the ML doing too little or too much precipitation (drying of the atmosphere).

Figure 7 shows this quantity for the verification data (fine-grid reference simulation), a no-ML baseline that uses conventional parameterizations, and 10 prognostic runs that use ML models trained with different random seeds. The fine-grid reference maintains a near-constant water vapor path. The baseline moistens by about 1.5 mm within 4 days and then stabilizes. All 10 prognostic runs simulate 10 days without crashing. Most seeds show a stable global mean water vapor path for about 4 days of simulation, followed by a rapid reduction in global water vapor path of about 1 mm in nearly all runs. As will be shown in Section 5.4 this is coincident with a significant reduction in the simulated strength of the tropical overturning circulation. After 10 days of simulation, the spread in water vapor path across prognostic simulations is about 1.5 mm and nine out of 10 runs have an overly dry atmosphere. For subsequent analysis, we will focus on the only seed that did not develop this dry bias (dashed green line in Figure 7; hereafter called seed-5), which produced the run with smallest global mean water vapor path bias over the 10-day forecast period. If we choose a different initialization date, the evolution is similar for each seed, and again seed-5 is the only run without a dry bias after 10 days of simulation, suggesting that this is a robust feature of this particular trained model. The approach of training an ensemble of ML models offline and choosing one based on online performance on some particular test case has also been used for ML emulation of embedded cloud resolving models (X. Wang et al., 2022).

5.2. Ten-Day Global Mean Drifts

A basic metric of whether the ML has improved a prognostic simulation is reduction of bias in important variables relative to some baseline model. We compare with the no-ML baseline (the dashed gray line in Figure 7) and a simulation using B22's "nudge-to-fine" method, in which nudging-trained ML is used to do online bias correction

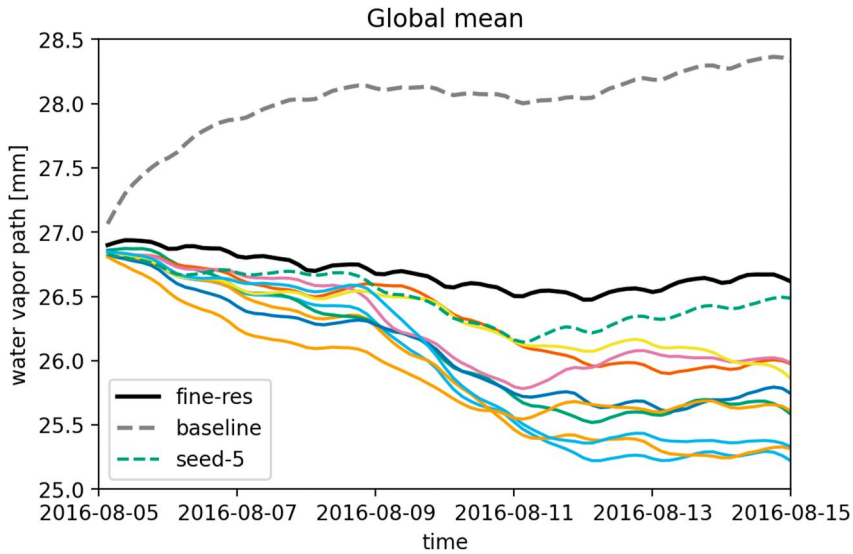

Figure 7. Global mean water vapor path for fine-grid reference (heavy black), no-ML baseline (heavy dashed gray) and 10 prognostic runs (colors) which differ only in the random seed used when training the neural network. The dashed green line is the simulation selected for further analysis in Sections 5.2–5.4.

of the coarse model rather than to replace the parameterized physics. Figures 8 and 9 show the evolution of the global mean bias in temperature and specific humidity as a function of time and vertical model level. These are for a single initialization starting at 0 UTC 5 August 2016. The ML-physics run here is seed-5, corresponding to the dashed green line in Figure 7. The baseline run has biases which develop very quickly (within 2 days) and then stabilize: too moist and cool in the free troposphere and too dry and warm in the boundary layer. The ML-physics run has significantly slower increases in bias for both temperature and humidity in the first few days. For temperature, the biases remain smaller in magnitude than the baseline run throughout the 10-day period. However, a large positive bias in specific humidity develops in the lowest model levels, suggestive of too little upward moisture transport via the combination of explicitly resolved dynamics and the ML parameterization. This feature is robust across all random seeds (not shown). The nudge-to-fine drifts (right panels of Figures 8 and 9) are mostly smaller than either the baseline or ML-physics runs. An exception is upper tropospheric temperature, which develops a strong warm bias after 10 days.

5.3. Comparison to Previous Corrective-ML Strategies

We extend the simulations described in Section 5.1 to 35 days to match the extent of the period with available verification data from the fine-res simulation. Of the 10 simulations, one crashes before reaching 35 days while

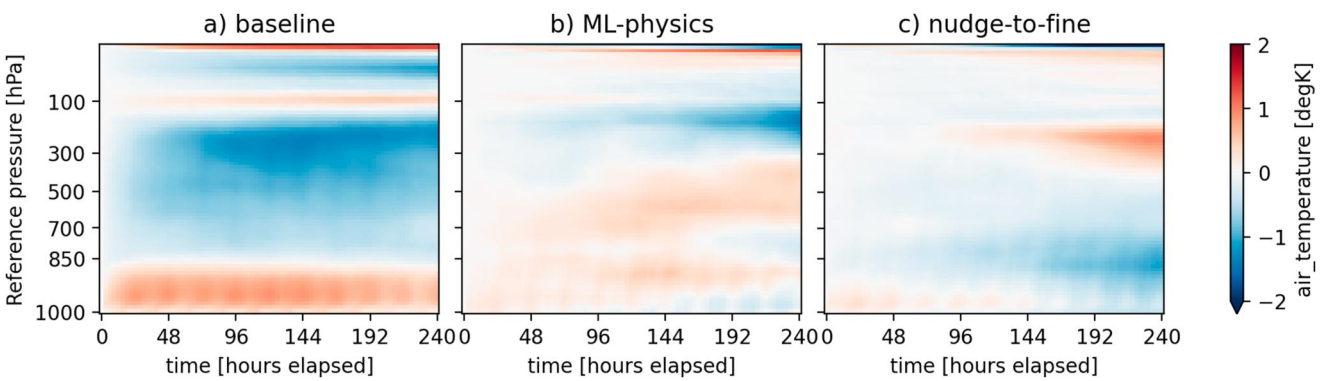

Figure 8. Global mean temperature bias compared to fine-resolution simulation for (left) no-ML baseline run, (middle) ML-physics prognostic run and (right) nudge-to-fine prognostic run.

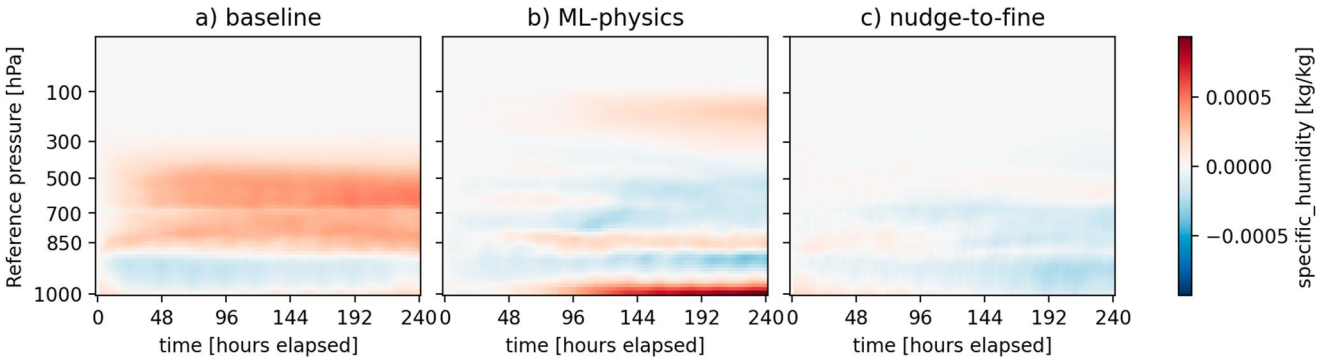

Figure 9. As in Figure 8 but for specific humidity.

the others are stable. Figure 10 shows the zonal mean temperature biases averaged over 35-days for the same runs as in Figures 8 and 9. The ML-physics prognostic run has significantly smaller temperature biases than the baseline over Antarctica, but otherwise has mostly slightly larger biases, comparable to the nudge-to-fine prognostic run. Over the Arctic, the ML-physics run reduces large temperature biases of the nudge-to-fine run to levels closer to the baseline simulation. Throughout much of the tropical and midlatitude troposphere, The ML-physics simulation has smaller relative humidity biases than the baseline simulation which is too moist almost everywhere (Figure 11). However, near the tropical tropopause and in Antarctica, the ML-physics simulation has a relative humidity that often far exceed 100%, a clearly unphysical result, leading to very large positive relative humidity biases. More work is needed to better understand the cause of these biases and how to ameliorate them. They may initiate in the ML-physics simulation due to ignoring the top 30 levels of temperature and specific humidity as inputs (required for stability) or the use of specific humidity instead of relative humidity as an input (Beucler et al., 2021). They may then grow large because the ML-physics parameterization does not build in any stabilizing feedbacks that create strong condensation and precipitation at high relative humidities exceeding 100%, since these are not encountered in the training data set. Such feedbacks are built into the physical parameterizations used in the baseline and nudge-to-fine simulations.

To more broadly measure the skill of the prognostic simulations, Table 1 includes five metrics which measure short-term weather forecast skill as well as the fidelity of the time-mean pattern of temperature and surface precipitation rate. These are the same metrics shown in Table 2 of B22 except additionally including the time-mean RMSE of 850 hPa atmospheric temperature. By all the measures shown except the 200 hPa temperature time-mean RMSE, the ML-physics prognostic run performs similarly or slightly better than the no-ML baseline. In particular, the RMSE of the time-mean precipitation rate and 850 hPa temperature are improved by 12% and 11% respectively, while the 200 hPa temperature time-mean RMSE is worsened by about 8%. It is reassuring that a column-local ML-physics scheme can modestly improve on the skill of a full suite of complex physical parameterizations for surface precipitation and lower tropospheric temperature. The ML-physics scheme is less

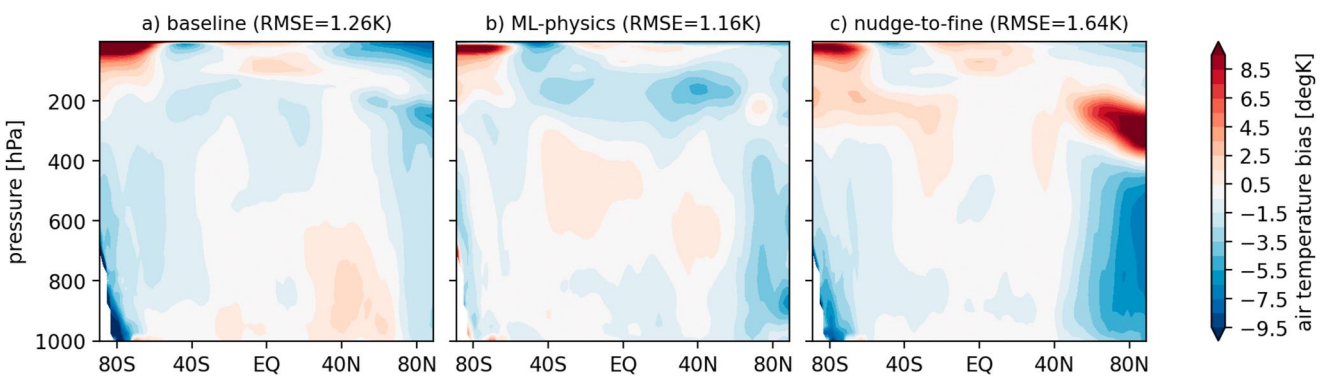

Figure 10. The zonal and time-mean air temperature bias with respect to the verification (fine-res) simulation, averaged over 35-day long simulations. For (left) the baseline, (middle) ML-physics and (right) nudge-to-fine simulations. The titles show area- and pressure-weighted RMSEs.

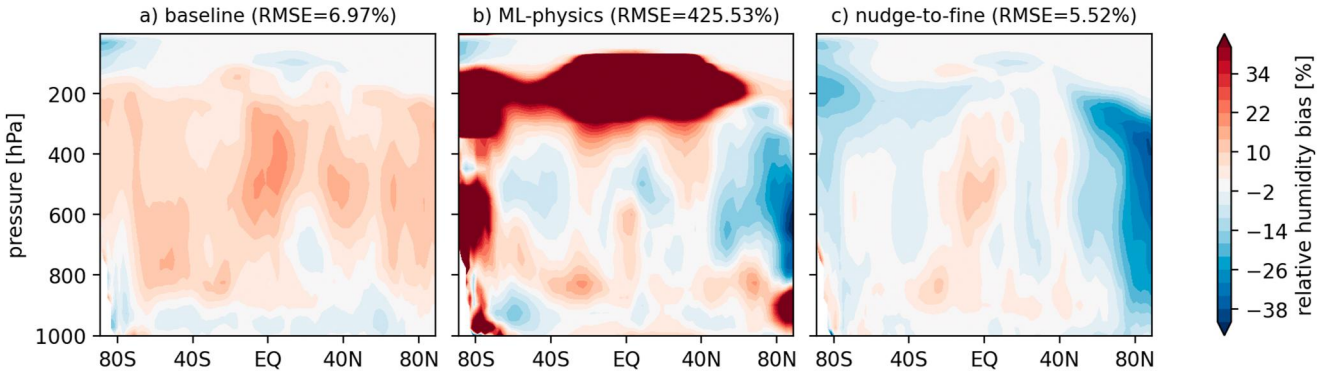

Figure 11. As in Figure 10 but for relative humidity. If the region for the RMSE calculation is restricted to below 350 hPa and equatorward of 60°, the values are 6.98%, 6.01%, 3.76% for the baseline, ML-physics and nudge-to-fine runs, respectively.

skillful than the nudge-to-fine method on the 3–7 day timescale weather forecast skill metrics and the surface precipitation rate time-mean RMSE, although it does improve the time-mean spatial pattern of 850 hPa temperature.

5.4. Tropical Precipitation Biases

The preceding results indicate good global performance of the ML-physics, but Figure 7 showed that most of our 10 ML random seeds led to prognostic simulations with overly dry atmospheres. This drying coincides with a weakening of the Hadley cell and upward mass transport in the ITCZ. Figure 12 compares the zonal mean surface precipitation rate for 35 days of ML-physics seed-5 and seed-7 simulations with the baseline, nudge-to-fine and reference fine-grid simulations. The latter three simulations all have an expected quasi-steady peak in precipitation slightly north of the equator. Both the baseline and nudge-to-fine simulations also develop a peak south of the equator (double ITCZ) that is not in the fine-grid run. On the other hand, the ML-physics prognostic runs (Figures 12b and 12c) show a less prominent peak in precipitation. A difference is also apparent between the two seeds shown for ML-physics. Seed-5 (Figure 12b; the seed for which results were shown in Figures 8–11), which does not dry globally, maintains a rainier deep tropics and drier subtropics. On the other hand, after about 10 days of simulation, seed-7 smears the precipitation uniformly over a wide band from 10°S to 20°N (see also Figure 12f), coincident with a weakening in upward mass transport in the ITCZ (not shown). However, even seed-5 shows signs of weakening ITCZ precipitation in the last 5 days of simulation shown in Figure 12.

Even the tropical circulation of the seed-5 ML physics simulation has some unrealistic characteristics. Figure 13 compares snapshots of the surface precipitation rate 4 days into the prognostic simulations. The baseline, nudge-to-fine and fine-res simulations all show a clear east-west band of precipitation slightly north of the equator in the

Table 1
Performance Metrics From 35-Day Prognostic Simulations

Metric	Units	Baseline	ML-physics	Nudge-to-fine
Z500 3–7 day RMSE	m	57.5	57.2	56.1
T850 3–7 day RMSE	K	2.87	2.77	2.57
Precip bias land-time-mean	mm/day	0.95	0.59	0.02
Precip time-mean RMSE	mm/day	3.24	2.74	2.39
T850 time-mean RMSE	K	2.06	1.84	2.27
T200 time-mean RMSE	K	2.08	2.25	2.02

Note. The 3–7 day RMSE metrics are averaged over four initialization dates. For the 3–7 day RMSE, first a global area-weighted RMSE is computed at each forecast lead time, and then these are averaged over the 3–7 day lead times. The time-mean RMSE is computed by first averaging over the 35-day simulations for the predicted and target fields, and then computing a global area-weighted RMSE (see Equation 5 of Sanford et al., 2023). In all cases, the coarsened fine-grid simulation is used as the target data.

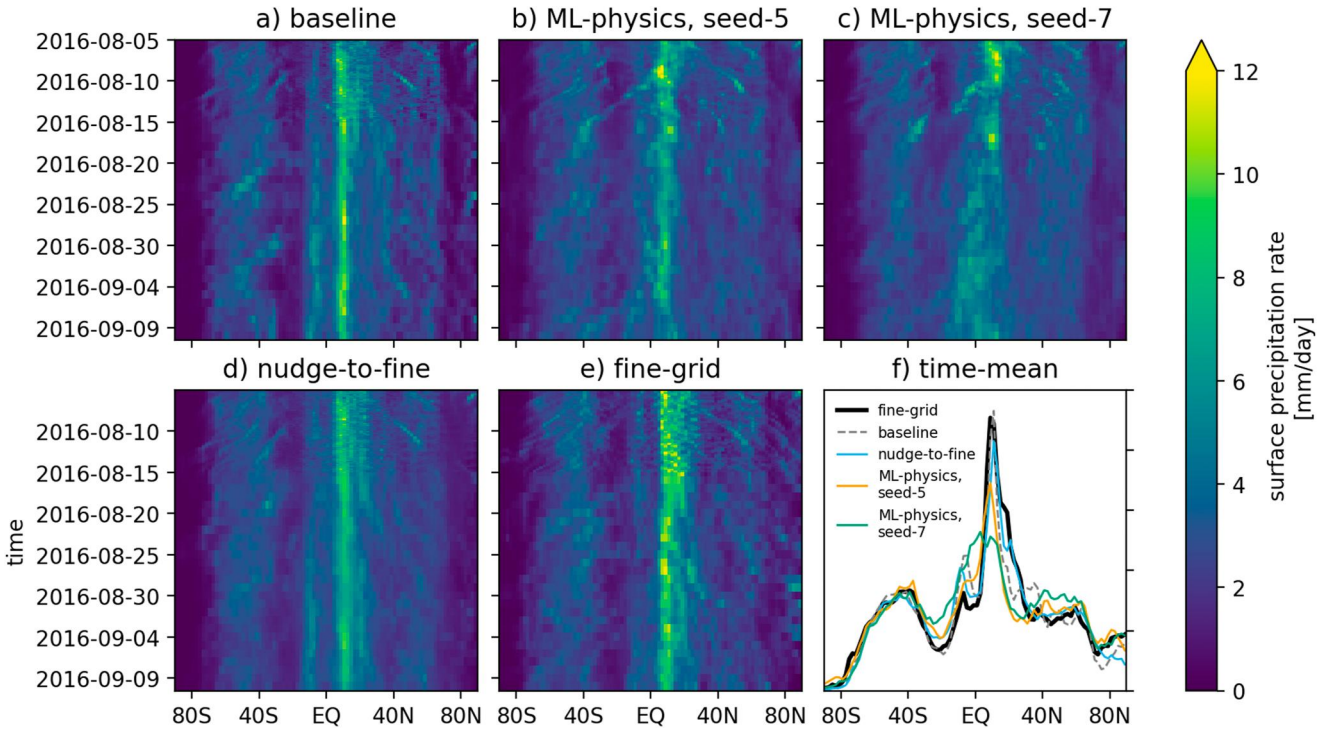

Figure 12. (a–e) The zonal mean precipitation rate as a function of time and latitude and (f) the zonal-mean precipitation averaged over 2016-08-15 to 2016-09-10. For panels (a–e), data is shown at 3-hourly frequency for the first 10 days and then as daily means thereafter.

East Pacific and Atlantic. On the other hand, the ML-physics simulation has developed individual storms spanning a few grid points and does not have an ITCZ-like structure in the East Pacific or Atlantic. Notably these are regions with strong meridional gradients in sea surface temperature which we expect to set the pattern of near-surface convergence (Back & Bretherton, 2009).

6. Ablations

Two algorithmic choices made in this study were predicting effective sources instead of apparent sources (i.e., including the apparent dynamics tendency in the ML target) and using the surface turbulent heat fluxes as ML

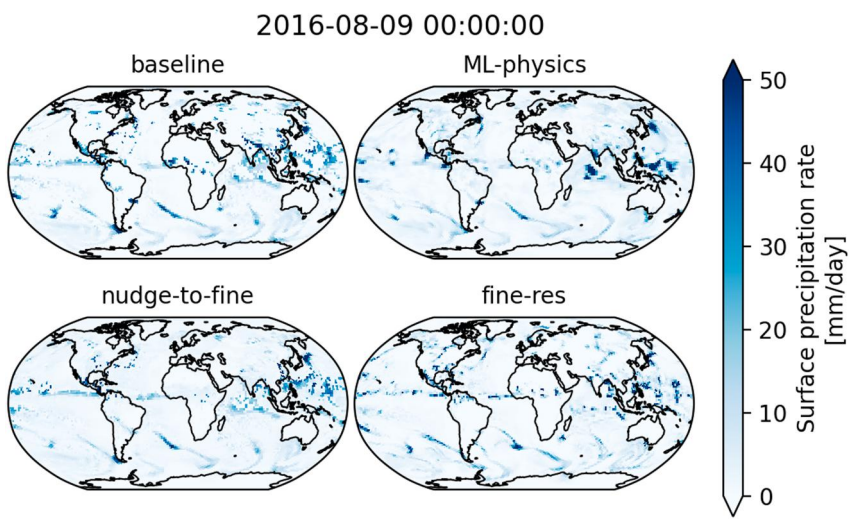

Figure 13. Snapshot of surface precipitation rate at 0 UTC 2016-08-09.

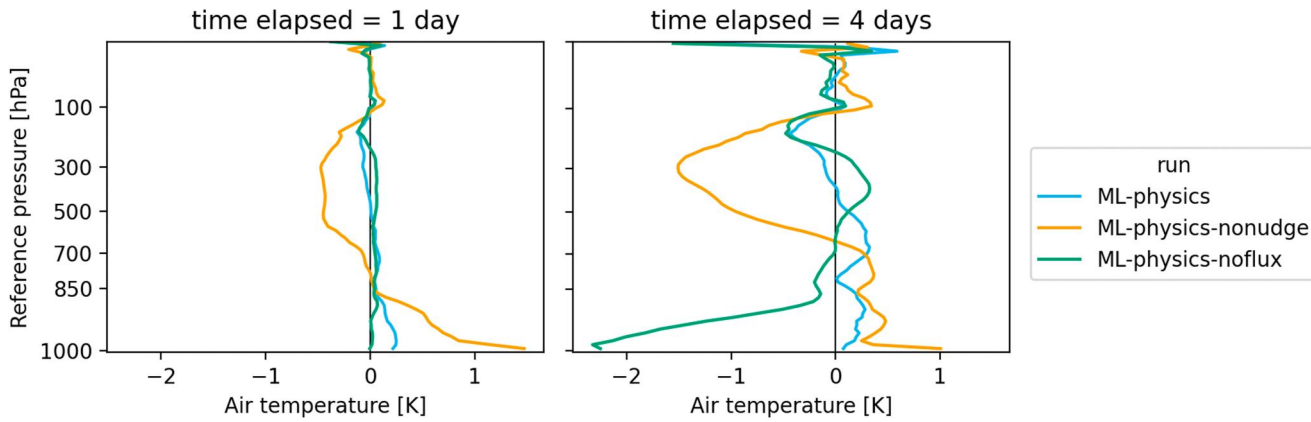

Figure 14. Global mean temperature bias for control simulation ("ML-physics"), one in which we do not include apparent dynamics tendency in ML target ("ML-physics-nonudge") and one in which we do not use the surface turbulent fluxes as predictors ("ML-physics-noflux"). For snapshots that are (left) one day and (right) 4 days into simulation.

predictors. This section demonstrates that these choices are necessary to avoid rapid drifts in the atmospheric state. Figure 14 shows the global mean bias in temperature relative to the fine-resolution run for three simulations. The first is the ML-physics run discussed in previous sections. In the second one, the surface turbulent heat fluxes are not used as ML inputs ("ML-physics-noflux"), and in the third we additionally do not include the apparent dynamics tendency in the ML target ("ML-physics-nonudge"). After one day of simulation, ML-physics-nonudge develops a roughly 1 K global mean positive bias in temperature near the surface and an approximately −0.5 K cool bias in the upper troposphere, both significantly larger than the corresponding biases in the ML-physics simulation. The changes in the T and q biases are approximately equal to the negative of the apparent dynamics tendency multiplied by one day (not shown), suggesting that no rapid feedback is preventing further divergence of the bias patterns of these two simulation; indeed their bias difference in the free troposphere has amplified 2–3 fold by day 4.

After 4 days (right panel of Figure 14), the importance of using surface turbulent heat fluxes as ML inputs is also apparent. The ML-physics-noflux simulation has developed a large −2 K global-mean near-surface temperature bias, unlike the ML-physics simulation. A large drift in lowest layer air temperature over ocean will, all else being equal, lead to a large change in sensible heat flux. Indeed the ML-physics-noflux simulation has an ocean-mean surface sensible heat flux (predicted by the piggy-backed physics parameterizations) of about 57 W/m^2 after 4 days of simulation, compared to the initial value of 14 W/m^2. However, since sensible heat flux is not used as an ML input in this simulation, the atmosphere does not respond to this strong positive heat flux and the negative near-surface air temperature bias continues to grow. To contrast, the ML-physics simulation, for which surface heat flux is an ML input, maintains an ocean-mean sensible heat flux around 14 W/m^2, with much reduced drift in air-sea temperature difference. Even ML-physics develops a negative bias in ocean-mean latent heat flux and an increase in near-surface humidity (Figure 9), possibly indicating too little vertical mixing of moisture by the ML physics. Previous machine learning based physics have used a more conventional turbulence scheme, either predicting a diffusivity (Yuval et al., 2021) or directly using an explicit planetary boundary layer parameterization (X. Wang et al., 2022), and this may be a strategy worth pursuing.

Beyond reducing global drifts, the algorithmic choices discussed in this section also lead to increases in stability of prognostic simulations. Compared to nine out of 10 simulations being stable to 35 days for ML-physics in a random seed ensemble, only two and six out of 10 are stable for ML-physics-noflux and ML-physics-nonudge, respectively.

7. Discussion and Conclusions

This study demonstrates promise for the use of machine learning for replacing the physical parameterization suite of a realistic geography coarse-grid global atmospheric model. Neural-network based ML is trained by coarse-graining outputs from a realistic geography global storm-resolving (3 km horizontal grid) simulation. Like

most physical parameterizations, the ML is restricted to be column-local, predicting temperature and moisture tendencies in each grid column and time based on temperature and humidity profiles from that grid column, plus five other physically relevant scalar inputs. Column-locality makes the resulting ML parameterization more amenable to physical interpretation, including column budgets of heat and moisture, and to comparison with traditional physical parameterizations.

Stable 35-day simulations are achieved, with most global skill metrics (Table 1) comparable to or slightly better than a baseline simulation with conventional physics parameterizations, although not as good as the ML-corrected nudge-to-fine methodology. The most important choice to enable this success was the inclusion of heating and moistening tendencies due to "semi-resolved" grid-scale motions that can be represented on the coarse grid, but are not accurately simulated by the coarse model dynamical core. These tendencies were calculated using a nudged prescribed-physics simulation. Use of the surface turbulent heat fluxes, as computed by the piggy-backed (Grabowski, 2019) coarse resolution model's physics, as ML inputs significantly reduced the drifts of near-surface temperature. In future, further improvements might be achieved by direct prediction or correction of the coarse model's estimates of the surface turbulent heat fluxes.

Despite these modest successes there are clear drawbacks to our ML parameterization. It is trained over a limited range of atmospheric states so there is no guarantee of good performance outside this range, for example, after mean state drifts. While it respects column-wise formulation of humidity conservation, it is not explicitly trained to obey physical principles such as the Clausius-Clapeyron equation, enabling the simulated relative humidity to reach unrealistically large values in the tropical upper troposphere and polar regions. In addition, for most choices of ML random seed, there is a rapid decrease in the strength of the zonal mean overturning circulation in prognostic simulations. This decrease was also noted, although to a lesser degree, in an aquaplanet simulation with a coarse-graining based ML parameterization (Brenowitz & Bretherton, 2019) but is not apparent in other aquaplanet-based studies (Yuval et al., 2021; Yuval & O'Gorman, 2020). The causes of the weakening overturning circulation and the differences between previous studies are unclear and merit further investigation. It is possible that it is necessary to apply sub-grid momentum tendencies derived from the fine-grid model in addition to the temperature and moisture terms discussed in this manuscript. Even for the best-case ML parameterization in this study, which does maintain a more realistic zonal-mean circulation, the tropical circulation quickly becomes dominated by relatively small (a few grid points) individual storms with inadequate large-scale organization. Since the reference simulation used for training is only 40 days long, this study did not aim to run prognostic simulations for a full year across the seasonal cycle. However, we did extend the seed-5 ML-physics prognostic simulation in order to probe its long-term behavior. After maintaining a reasonable global mean water vapor path for the first 35 days of the simulation, we find the simulation steadily dries out after this time. After about 102 days, the simulation crashed as a temperature field went outside the range of a saturation vapor pressure lookup table. It will be important to test whether training on a full year of data (recently available, e.g. Cheng et al., 2022; Kwa et al., 2022) would allow longer ML-physics simulations.

Multiple other lines of future research could be fruitful. The excessive accumulation of moisture near the surface could possibly be addressed by using ML to tune a conventional parameterization for the planetary boundary layer (as in Yuval et al., 2021; X. Wang et al., 2022). In addition, the extreme (much greater than 1) values of relative humidity are clearly unrealistic. Using relative humidity instead of specific humidity as an ML input could help the ML better learn how to mitigate this bias from the training data, could help generalization to other climates (Beucler et al., 2021), and could be used to enforce an artificial upper bound on relative humidity when it has gone outside the training range. To avoid rapid blow-ups of prognostic simulations, it was necessary for us to not use the uppermost 30 levels as inputs for ML prediction (Brenowitz & Bretherton, 2019), which can be interpreted as a crude way of enforcing causality when the training data set is coarse-grained in space and time. It would be satisfying to have another way to prevent instabilities, since our current strategy makes the ML completely insensitive to the upper tropospheric and stratospheric state, which will inevitably create large mean-state drifts in simulations longer than 35 days. Finally, we are still relying on the coarse-resolution parameterizations for horizontal wind tendencies and surface turbulent heat fluxes. ML prediction of the wind tendencies (Yuval & O'Gorman, 2021) could also further improve the prognostic simulation, especially given the large wind nudging tendencies over topography seen in the coarse-resolution model (Watt-Meyer et al., 2021). This would require additional coarse-grained outputs from the fine-grid reference model that were not easy to implement but are computable in principle.

To conclude, machine-learning parameterizations trained on coarse-grained realistic geographic global storm-resolving simulations are possible and on monthly timescales, they are approaching the skill of conventional parameterizations and the prognostic skill of corrective machine learning approaches.

Appendix A: Coarse-Graining Lowest Model Layer Over Ocean

In order to better maintain properties such as hydrostatic and thermal wind balance, in particular over regions of sloping topography, coarse-graining from C3072 to C48 was performed on surfaces of constant pressure (B22). This required vertical interpolation of the fine-grid data, since the model layers are terrain-following and not necessarily coincident with constant pressure. As described in Section 2.4 of B22, there is no downward extrapolation of C3072 grid columns: motivated by the variation in surface pressure in mountainous regions, the average is taken over only above-ground portions of C3072 columns, where the C3072 surface pressure is greater than the corresponding C48 coarse-grained pressure. However, this means that over ocean, any surface pressure gradient across a C48 grid cell leads to the masking of many C3072 grid cells in the lowermost model layer (Figure A1). If there is any correlation between surface pressure and the variable being coarse-grained this will lead to biases in the coarse-grained value at the lowest model layer.

Figure A2 shows the coarse-grained pressure velocity ω in the lowest model layer using the masked weights (left) and unmasked weights (right). For simplicity, here we show a one-step coarse-graining from C3072 to C48, although in practice we do this in two steps with an intermediate C384 resolution. In our standard procedure, we use the masked weights since this gives better performance over mountainous regions. However, due to correlations between surface pressure and ω in regions of convection, this leads to very large (up to 0.2 Pa/s) values of C48-averaged pressure velocity over ocean in the lowermost model layer (Figure A2a). If no masking is performed, the resulting C48 pressure velocity has much smaller magnitude in almost all grid cells (Figure A2b). The masking of certain C3072 grid cells for the lowest layer over ocean also affects the coarse-graining of other variable such as temperature and specific humidity, but not as much as the pressure velocity (not shown). Note this masking over ocean almost always occurs for the lowest model layer only.

Since the pressure-level coarse-graining from C3072 to C384 is performed online, changing the masking done over ocean would require rerunning the C3072 simulation which was not easily feasible given our computation resources. Nevertheless, the prescribed physics method is able to compensate for the impact of coarse-graining induced biases in ω and vertical eddy fluxes. Ongoing work is exploring a modified coarse-graining strategy that does not have this issue of masked grid cells over ocean.

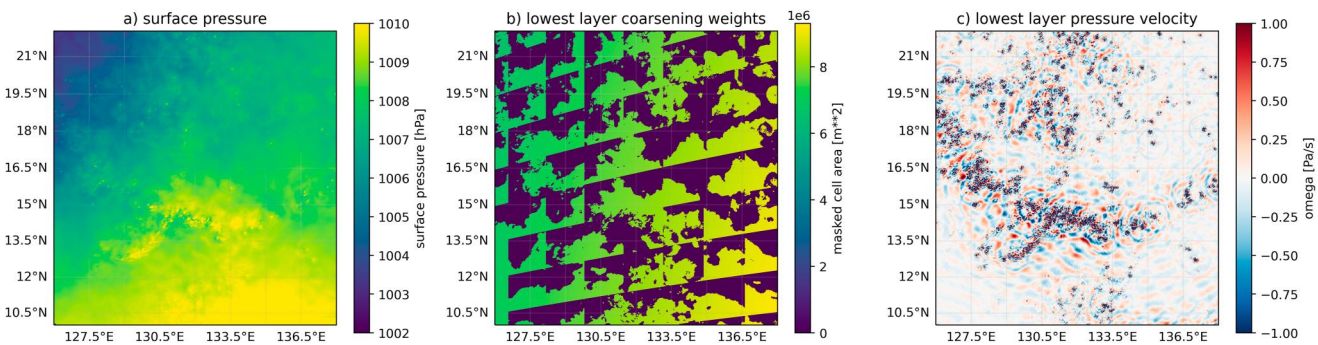

Figure A1. A snapshot of C3072 data over a region with active convection in the West Pacific warm pool. Showing (a) surface pressure, (b) the weights used for coarse-graining to C48 resolution at the lowest model layer, (c) pressure velocity ω in the lowest model layer. The weights for coarsening are set to zero wherever the C3072 surface pressure is less than the corresponding coarse-grained C48 surface pressure.

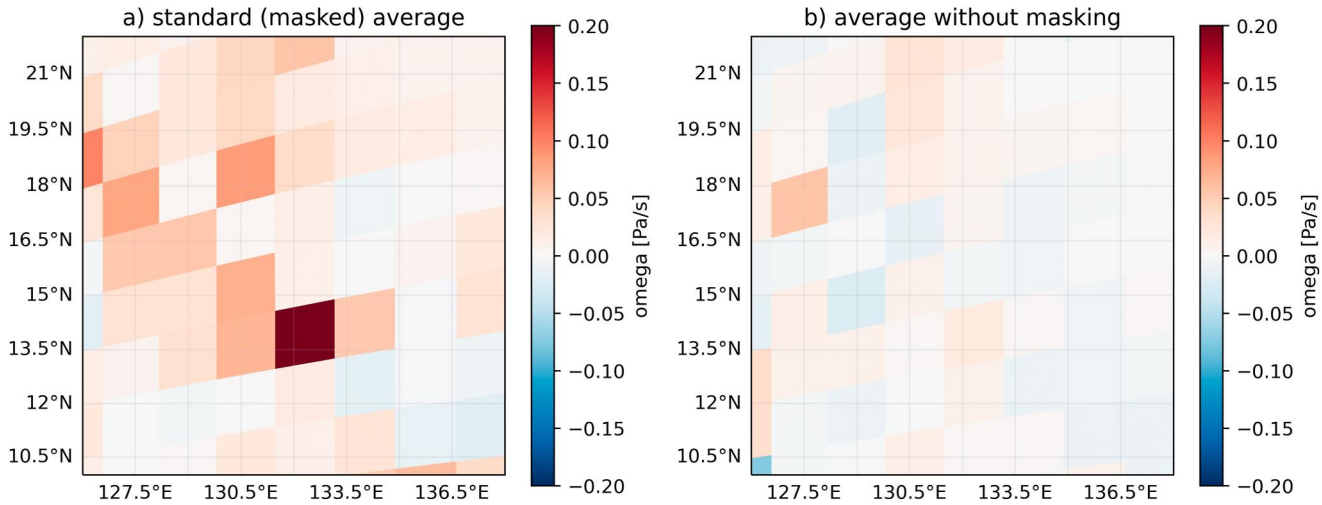

Figure A2. The C48 coarse-grained lowest layer pressure velocity ω from the same snapshot shown in Figure A1. Coarse-grained (a) using the masked weights shown in Figure A1b, (b) using unmasked cell area as weights.

Data Availability Statement

This work used the FV3GFS atmospheric model, also known as the UFS, available at https://github.com/ufs-community/ufs-weather-model (UFS Community, 2020). The model was forked to https://github.com/ai2cm/fv3gfs-fortran (AI2CM, 2021) for internal development to allow running simulations on Google Cloud Platform and interfacing with python. A repository with complete configuration of experiments and notebooks used for figure generation is available at https://github.com/ai2cm/fine-grid-ml-physics-manuscript (Watt-Meyer, 2023). The coarsened fine-grid data used for training is available upon request through a Google Cloud Storage requester pays bucket.

Acknowledgments
We thank the Allen Institute for Artificial Intelligence for supporting this work and NOAA-GFDL for running the 40-day reference X-SHiELD simulation using the Gaea computing system. We also acknowledge NOAA-GFDL, NOAA-EMC, and the UFS community for making code and software packages publicly available. We also thank two anonymous reviewers for insightful comments and questions that significantly improved the manuscript.

References

AI2CM. (2021). AI2 climate modeling fork of FV3GFS [Software]. Zenodo. https://doi.org/10.5281/zenodo.4470023

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

Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., et al. (2021). Climate-invariant machine learning. arXiv. https://doi.org/10.48550/ARXIV.2112.08440

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

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., et al. (2020). Machine learning climate model dynamics: Offline versus online performance. arXiv. https://doi.org/10.48550/ARXIV.2011.03081

Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGibbon, J., et al. (2022). Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *Journal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. https://doi.org/10.1029/2021MS002794

Caldwell, P. M., Zelinka, M. D., Taylor, K. E., & Marvel, K. (2016). Quantifying the sources of intermodel spread in equilibrium climate sensitivity. *Journal of Climate*, *29*(2), 513–524. https://doi.org/10.1175/JCLI-D-15-0352.1

Chen, G., Held, I. M., & Robinson, W. A. (2007). Sensitivity of the latitude of the surface westerlies to surface friction. *Journal of the Atmospheric Sciences*, *64*(8), 2899–2915. https://doi.org/10.1175/JAS3995.1

Cheng, K.-Y., Harris, L., Bretherton, C., Merlis, T. M., Bolot, M., Zhou, L., et al. (2022). Impact of warmer sea surface temperature on the global pattern of intense convection: Insights from a global storm resolving model. *Geophysical Research Letters*, *49*(16), e2022GL099796. https://doi.org/10.1029/2022GL099796

Clark, S. K., Brenowitz, N. D., Henn, B., Kwa, A., McGibbon, J., Perkins, W. A., et al. (2022). Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *Journal of Advances in Modeling Earth Systems*, *14*(9), e2022MS003219. https://doi.org/10.1029/2022MS003219

Dahm, J., Davis, E., Deconinck, F., Elbert, O., George, R., McGibbon, J., et al. (2022). Pace v0.1: A python-based performance-portable implementation of the FV3 dynamical core. *EGUsphere*, *2022*, 1–24. https://doi.org/10.5194/egusphere-2022-943

Derbyshire, S. H., Beau, I., Bechtold, P., Grandpeix, J.-Y., Piriou, J.-M., Redelsperger, J.-L., & Soares, P. M. M. (2004). Sensitivity of moist convection to environmental humidity. *Quarterly Journal of the Royal Meteorological Society*, *130*(604), 3055–3079. https://doi.org/10.1256/qj.03.130

Durran, D. R. (2010). *Numerical methods for fluid dynamics*. Springer. https://doi.org/10.1007/978-1-4419-6412-0

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

Grabowski, W. W. (2019). Separating physical impacts from natural variability using piggybacking technique. *Advances in Geosciences*, *49*, 105–111. https://doi.org/10.5194/adgeo-49-105-2019

Guichard, F., Petch, J. C., Redelsperger, J.-L., Bechtold, P., Chaboureau, J.-P., Cheinet, S., et al. (2004). Modelling the diurnal cycle of deep precipitating convection over land with cloud-resolving models and single-column models. *Quarterly Journal of the Royal Meteorological Society*, *130*(604), 3139–3172. https://doi.org/10.1256/qj.03.145

Han, J., & Pan, H.-L. (2011). Revision of convection and vertical diffusion schemes in the NCEP global forecast system. *Weather and Forecasting*, *26*(4), 520–533. https://doi.org/10.1175/WAF-D-10-05038.1

Han, J., Wang, W., Kwon, Y. C., Hong, S.-Y., Tallapragada, V., & Yang, F. (2017). Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Weather and Forecasting*, *32*(5), 2005–2017. https://doi.org/10.1175/WAF-D-17-0046.1

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). https://doi.org/10.1029/2020ms002076

Han, Y., Zhang, G. J., & Wang, Y. (2023). An ensemble of neural networks for moist physics processes, its generalizability and stable integration. *Journal of Advances in Modeling Earth Systems*, *15*(10), e2022MS003508. https://doi.org/10.1029/2022ms003508

Harris, L., Chen, X., Putman, W., Zhou, L., & Chen, J.-H. (2021). A scientific description of the GFDL finite-volume cubed-sphere dynamical core. https://doi.org/10.25923/6nhs-5897

Harris, L., Zhou, L., Kaltenbaugh, A., Clark, S., Cheng, K.-Y., & Bretherton, C. (2023). A global survey of rotating convective updrafts in the GFDL x-shield 2021 global storm resolving model. *Journal of Geophysical Research: Atmospheres*, *128*(10), e2022JD037823. https://doi.org/10.1029/2022JD037823

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. arXiv. Retrieved from https://arxiv.org/abs/1412.6980

Kwa, A., Clark, S. K., Henn, B., Brenowitz, N. D., McGibbon, J., Watt-Meyer, O., et al. (2023). Machine-learned climate model corrections from a global storm-resolving model: Performance across the annual cycle. *Journal of Advances in Modeling Earth Systems*, *15*(5), e2022MS003400. http://doi.org/10.1029/2022MS003400

McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J., Davis, E., et al. (2021). FV3GFS-wrapper: A python wrapper of the FV3GFS atmospheric model. *Geoscientific Model Development Discussions*, *14*(7), 4401–4409. https://doi.org/10.5194/gmd-14-4401-2021

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentine, P. (2021). Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *Journal of Advances in Modeling Earth Systems*, *13*(5), e2020MS002385. https://doi.org/10.1029/2020MS002385

Pithan, F., Shepherd, T. G., Zappa, G., & Sandu, I. (2016). Climate model biases in jet streams, blocking and storm tracks resulting from missing orographic drag. *Geophysical Research Letters*, *43*(13), 7231–7240. https://doi.org/10.1002/2016GL069551

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

Sanford, C., Kwa, A., Watt-Meyer, O., Clark, S. K., Brenowitz, N., McGibbon, J., & Bretherton, C. (2023). Improving the reliability of ML-corrected climate models with novelty detection. *Journal of Advances in Modeling Earth Systems*, *15*(11), e2023MS003809. https://doi.org/10.1029/2023ms003809

Song, X., & Zhang, G. J. (2018). The roles of convection parameterization in the formation of double ITCZ syndrome in the NCAR CESM: I. Atmospheric processes. *Journal of Advances in Modeling Earth Systems*, *10*(3), 842–866. https://doi.org/10.1002/2017MS001191

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

UFS Community. (2020). Unified forecast system (UFS) [Software]. Zenodo. https://doi.org/10.5281/zenodo.4460292

van Niekerk, A., Shepherd, T. G., Vosper, S. B., & Webster, S. (2016). Sensitivity of resolved and parametrized surface drag to changes in resolution and parametrization. *Quarterly Journal of the Royal Meteorological Society*, *142*(699), 2300–2313. https://doi.org/10.1002/qj.2821

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*(10), e2022MS002984. https://doi.org/10.1029/2022MS002984

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2022). Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development*, *15*(9), 3923–3940. https://doi.org/10.5194/gmd-15-3923-2022

Watt-Meyer, O. (2023). Experiment configuration and analysis code for "Neural network parameterization of subgrid-scale1 physics from a realistic geography global storm-resolving simulation" [Software]. Zenodo. https://doi.org/10.5281/zenodo.10138793

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., et al. (2021). Correcting weather and climate models by machine learning nudged historical simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. https://doi.org/10.1029/2021GL092555

Woelfle, M. D., Yu, S., Bretherton, C. S., & Pritchard, M. S. (2018). Sensitivity of coupled tropical pacific model biases to convective parameterization in CESM1. *Journal of Advances in Modeling Earth Systems*, *10*(1), 126–144. https://doi.org/10.1002/2017MS001176

Yanai, M., Esbensen, S., & Chu, J.-H. (1973). Determination of bulk properties of tropical cloud clusters from large-scale heat and moisture budgets. *Journal of the Atmospheric Sciences*, *30*(4), 611–627. https://doi.org/10.1175/1520-0469(1973)030⟨0611:DOBPOT⟩2.0.CO;2

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

Yuval, J., & O'Gorman, P. A. (2021). Neural-network parameterization of subgrid momentum transport in the atmosphere. *ESS 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

Zhao, M. (2014). An investigation of the connections among convection, clouds, and climate sensitivity in a global climate model. *Journal of Climate*, *27*(5), 1845–1862. https://doi.org/10.1175/JCLI-D-13-00145.1

Zhou, L., Lin, S.-J., Chen, J.-H., Harris, L. M., Chen, X., & Rees, S. L. (2019). Toward convective-scale prediction within the next generation global prediction system. *Bulletin of the American Meteorological Society*, *100*(7), 1225–1243. https://doi.org/10.1175/bams-d-17-0246.1