JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2022MS003400

Machine-Learned Climate Model Corrections From a Global Storm-Resolving Model: Performance Across the Annual Cycle

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

[1]Allen Institute for Artificial Intelligence, Seattle, WA, USA, [2]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ, USA

Key Points:

- Machine-learned (ML) corrective tendencies and radiative fluxes for a coarse-grid global model are trained using a year-long global storm-resolving simulation
- The ML corrections reduce systematic errors of land surface temperature and precipitation in multiyear coarse-model simulations
- The ML corrections nevertheless induce systematic biases in the simulated Hadley circulation

Correspondence to:
A. Kwa,
annak@allenai.org

Citation:
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*, e2022MS003400. https://doi.org/10.1029/2022MS003400

Received 9 SEP 2022
Accepted 28 APR 2023

Abstract One approach to improving the accuracy of a coarse-grid global climate model is to add machine-learned (ML) state-dependent corrections to the prognosed model tendencies, such that the climate model evolves more like a reference fine-grid global storm-resolving model (GSRM). Our past work demonstrating this approach was trained with short (40-day) simulations of GFDL's X-SHiELD GSRM with 3 km global horizontal grid spacing. Here, we extend this approach to span the full annual cycle by training and testing our ML using a new year-long GSRM simulation. Our corrective ML models are trained by learning the state-dependent tendencies of temperature and humidity and surface radiative fluxes needed to nudge a closely related 200 km grid coarse model, FV3GFS, to the GSRM evolution. Coarse-grid simulations adding these learned ML corrections run stably for multiple years. Compared to a no-ML baseline, the time-mean spatial pattern errors with respect to the fine-grid target are reduced by 6%–26% for land surface temperature and 9%–25% for land surface precipitation. The ML-corrected simulations develop other biases in climate and circulation that differ from, but have comparable amplitude to, the no-ML baseline simulation.

Plain Language Summary A recent vein of research uses machine learning (ML) to try and improve the predictive accuracy of climate models. Fine-resolution global storm-resolving simulations make more accurate rainfall and temperature predictions than climate models, but computers take too long to finish many-year simulations. We use data from a year-long high quality reference simulation to train ML models, which are then used in a lower quality but faster climate model. The ML applies corrections continually during the faster simulation to make it act more like the slow, high quality model. This improves the faster model's predictions for rainfall and temperature over land but also has unintended side effects of drying out the atmosphere and changing its circulation and tropical rainfall patterns.

1. Introduction

Due to computational constraints, running global climate models (GCMs) for more than a few years requires a spatial grid ($\gtrsim$50 km) too coarse to resolve two key atmospheric physical processes: cumulonimbus convection and airflow over orography, coastlines and other land-surface heterogeneities. The subgrid variability of these processes are approximately represented in GCMs via expert-designed physical parameterizations. Subjective choices made within these parameterizations contribute significantly to uncertainty in GCM predictions of precipitation, cloud cover, etc. (Chen et al., 2007; Woelfle et al., 2018; Zhao, 2014).

Global storm-resolving models (GSRMs), defined following Stevens et al. (2019) as global models with horizontal grid spacings <5 km and at least 50 vertical levels, are able to better resolve these processes, but are currently too computationally demanding to be run for periods of more than a year or two. One way to improve the accuracy of GCMs while still maintaining the computational speedup from running at lower resolution is to train a machine learning (ML) model whose outputs can be applied to update the GCM state at each timestep, with the goal of making the GCM state evolve more like the coarsened state of an equivalent fine-grid GSRM model. Brenowitz and Bretherton (2019) and Yuval and O'Gorman (2020) trained ML models to predict the coarsened fine-grid apparent sources (Yanai et al., 1973) and apply these within coarse-grid aquaplanet simulations. Bretherton et al. (2022) (hereafter B22) trained ML models to predict corrections to the GCM subgrid parameterizations, which were then predicted and applied at runtime in a coarse-grid simulation with realistic topography.

Prior work has used the nudging framework as a means of generating training data and applying corrective ML tendencies within a free-running coarse-grid simulation (Bretherton et al., 2022; Clark et al., 2022; Watt-Meyer

et al., 2021). Training data is generated by running a coarse-grid model that is linearly relaxed (nudged) toward the reference simulation state at each timestep; this constrains the training simulation to follow the evolution of the target. ML models are trained to use the coarse-grid states in each grid column (79-level vertical profiles of temperature, humidity and optionally horizontal wind components) as input features and predict the corresponding vertical profiles of nudging tendencies. The nudging target varies across these works depending on the goal: nudging to observational reanalysis allows for effective bias correction without costly fine-grid simulations (Watt-Meyer et al., 2021), nudging to a higher-resolution (∼25 km grid) model with parameterized convection allows for longer nudged runs to explore model performance in a range of climates (Clark et al., 2022), and nudging to a fine-grid (∼3 km grid) GSRM allows the ML to benefit from a target data set which directly resolves convection (B22).

B22 found that corrective ML models trained with the nudge-to-fine approach improved the weather forecast skill and reduced errors in time-mean land precipitation and surface temperature and the diurnal cycle of land precipitation in coarse-grid simulations of the 40-day period of their GSRM reference data. In this paper, we extend B22's study by using a year-long GSRM reference simulation for training the corrective ML and evaluating simulations with the resulting ML-corrected coarse model on seasonal and multiyear timescales. In Section 2 we describe our coarse-grid and fine-grid models as well as the nudged coarse-grid simulation used as our training data set. In Section 3 we describe the training procedure for our corrective ML models. Section 4 presents the offline (or single timestep) skill of the trained ML models. Section 5 presents online results and discusses the impact of the corrective ML when coupled to a free-running yearlong coarse-grid FV3GFS simulation. Following B22, we use land precipitation and surface temperature as the primary metrics to gauge improvement over a free-running FV3GFS no-ML coarse simulation. We also discuss circulation changes in the ML-corrected simulations.

2. Simulations

2.1. Coarse-Grid Model

As in B22, the coarse-grid model is FV3GFS (Zhou et al., 2019) run at C48 (∼200 km) resolution with the same 79 vertical model levels as the fine-grid model. Coarse simulations are carried out using a python-wrapped version of FV3GFS, which allows for easy customization and setup of nudging and integration of ML models (McGibbon et al., 2021). The model (and physics) timestep is 15 min, with 6 dynamics substeps per physics timestep. Time-varying sea surface temperatures (SSTs) and sea ice fraction are prescribed to be identical to those used in the coarsened fine-grid reference.

The following improvements were made to the coarse-model configuration in B22:

1. Fast saturation adjustment of humidity is enabled during each dynamics substep to better simulate fast-evolving cloud formation and dissipation. This significantly improves the baseline simulation's surface radiation and precipitation climatology with respect to the fine-grid reference. Over a 40 day free-running coarse-grid simulation, the time-mean global-averaged surface downward shortwave and longwave radiative fluxes and precipitation root mean squared errors (RMSEs) are improved by 13%, 12%, and 30%, respectively.
2. The background vertical diffusion coefficient for heat and moisture is set to 2 m^2/s to match the value of this parameter over land in the reference GSRM. Without this, the default FV3GFS value of 1 m^2/s led to crashes in the initial timestep.

We perform a free-running year-long simulation initialized from the coarsened fine-grid reference state on 19 January 2020. This simulation has no ML corrections and is referred to hereafter as the “baseline” run.

2.2. Reference Simulation

Our reference GSRM simulation is similar but not identical to the configuration used by B22, which was run for a much shorter 40-day period. In both this work and B22, the GSRM simulation is made with the X-SHiELD model, a modified version of FV3GFS with a C3072 cubed-sphere grid (∼3 km spacing) and 79 vertical levels, run on NOAA's Gaea computing system by collaborators at the Geophysical Fluid Dynamics Laboratory. The FV3GFS deep cumulus parameterization is disabled, though the shallow cumulus convection scheme is left active. The convective gravity wave drag scheme is disabled.

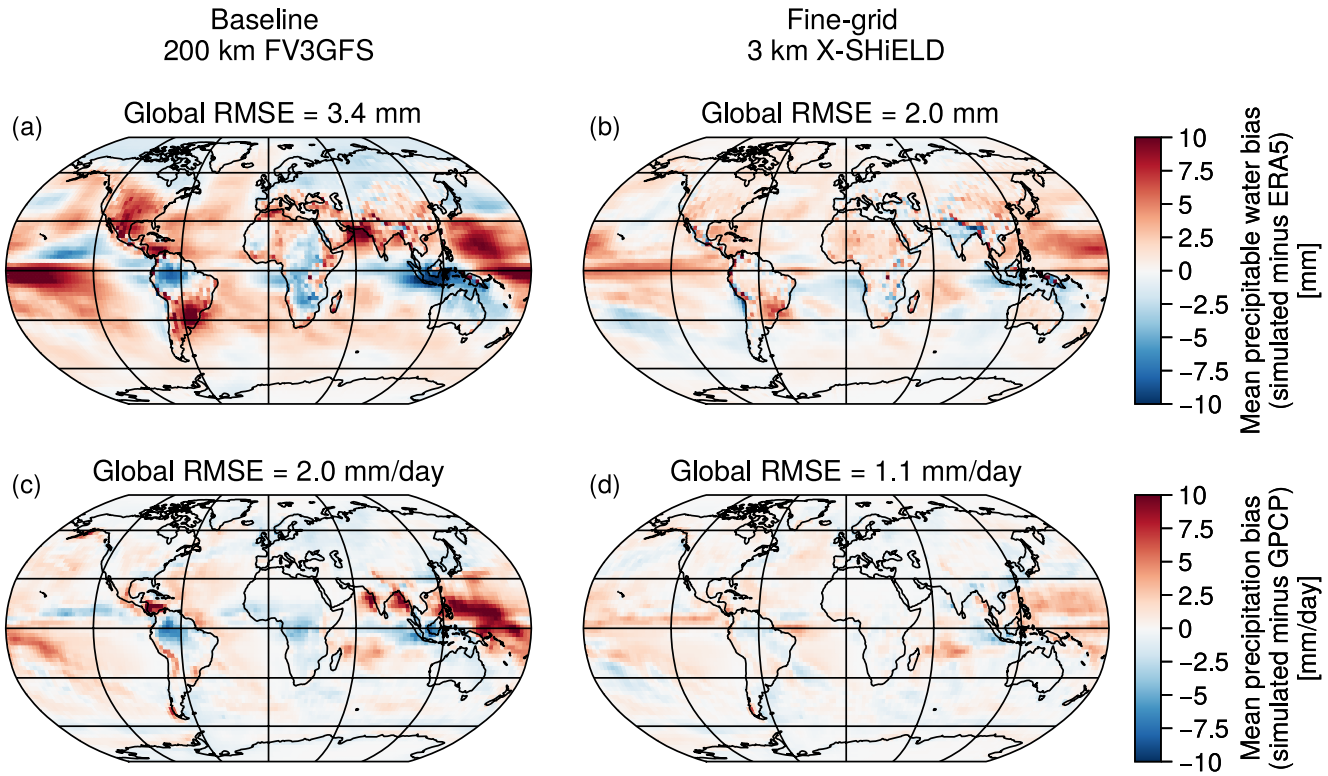

Figure 1. Top: Time-mean bias map of precipitable water over one simulated year in a coarse-grid FV3GFS run at C48 (∼200 km) resolution (a) and storm-resolving X-SHiELD run at C3072 (∼3 km) resolution (b). Bottom: Time-mean bias map of precipitation over 1 year in the coarse-grid FV3GFS (c) and fine-grid X-SHiELD (d) runs. The root mean squared error for each time-mean quantity is given in each subtitle.

Cheng et al. (2022) describes the exact X-SHiELD configuration used here and some general features of the simulation. The free-running simulation is initialized on 20 October 2019. The first three months are excluded as spin-up, resulting in a year-long reference data set spanning the remaining time from 19 January 2020 through 17 January 2021. The following are notable configuration differences in the reference fine-grid model used here with respect to the configuration in B22:

- There is no meteorological nudging of atmospheric fields to analysis.
- It uses the newer, inline version of the GFDL microphysics (Zhou et al., 2022).
- It uses a mixed-layer ocean between 45°S and 45°N, nudged with a 15-day timescale to ECMWF SSTs from analysis. SSTs outside of those latitudes are prescribed to ECMWF SSTs.
- Orographic gravity wave drag and mountain blocking schemes are enabled.

For use as the reference target in nudged training simulations, the X-SHiELD data is averaged from its native C3072 resolution down to the same C48 resolution as the coarse model, using the pressure-level coarsening procedure described in B22. All three-dimensional fields needed to restart the model, as well as numerous two-dimensional diagnostic fields, are coarsened inline and saved every 3 hr (B22 saved this output every 15 min, which is too expensive for our nine-fold longer training simulation).

Figure 1 compares coarse-grid FV3GFS and coarsened fine-grid X-SHiELD biases in 2020 annual-mean precipitable water with respect to ERA5 reanalysis (Hersbach et al., 2019) and precipitation biases with respect to GPCP observations (Adler et al., 2018). Both fields have a spatial RMSE with respect to the observations that is 40%–45% lower in the 3 km X-SHiELD simulation than the coarser 200 km FV3GFS simulation. This supports using X-SHiELD as a reference target for improving the accuracy of our coarse-grid simulation.

2.3. Nudged Training Simulation

We follow the nudging framework of B22. We initialize a 1-year coarse-grid FV3GFS simulation with the configuration described in Section 2.1 using the coarsened X-SHiELD state on 19 January 2020. Temperature,

humidity, model layer pressure thickness, and horizontal wind fields are nudged at each model timestep toward the fine-grid state with a 3-hr nudging timescale and the interval-averaged nudging tendencies at all coarse grid points are saved every three hours. The reference state is interpolated for timesteps that lie between the 3-hr intervals of the fine-grid data. The nudging tendencies are calculated as follows in Equation 1:

$$\Delta Q_a = -\frac{a^n - \bar{a}}{\tau} \tag{1}$$

where a^n is a prognostic field in the model, $\bar{a}$ is the coarsened value of that field in the reference fine-resolution data, and τ is a constant nudging timescale. As in B22, we use a nudging timescale τ of 3 hr. We interpret the nudging tendencies as corrections that would make the coarse model follow the evolution of the coarsened fine grid model and machine-learn these corrective tendencies as functions of the column state.

We use the following approach from B22 to correct for systematic biases in surface precipitation and downwelling radiative fluxes due to the coarse model physics producing less cloud and land precipitation than the fine-grid model. During the nudged training run we prescribe surface downwelling shortwave and longwave fluxes as well as precipitation from the coarsened fine-grid output to the land model. This prevents biases in land surface properties from feeding back into the atmosphere and affecting the temperature and humidity nudging tendencies. We train a ML model to predict downward longwave and shortwave surface fluxes for application in prognostic simulations. The radiative flux model training scheme is modified from B22 to improve its accuracy, as described below.

3. Machine-Learned Corrective Models

3.1. Training and Test Data

The nudged run data is divided into interleaved blocks of 2 weeks of training data followed by 1 week of validation and test data. The first and last two timesteps of each 2-week training data block are discarded to ensure that there are no consecutive timesteps in both the training and validation or test data sets.

One thousand three hundred sixty seven timesteps are selected out of the available training data. We make the assumption that the corrections predicted by the neural networks (NNs) are column-local and only depend on the state within a single grid column. At C48 grid resolution, each timestep contains 13,824 columns. Since nearby columns are strongly correlated, inclusion of all columns in a timestep does not provide significant benefit over using a reasonable subsample of the data. Initial experiments using a training data set of 200 timesteps showed that subsampling down to about 10% of the global set of columns in each timestep did not negatively impact the validation loss compared to using all columns. We use a subsample fraction of 15% of the total number of columns in each timestep in order to reduce memory usage and training time. Thus the training data set consists of $\sim 2.8 \times 10^6$ samples. The test data set used for offline evaluation excludes validation timesteps and is subsampled to 100 of the available test timesteps, with no column subsampling so as to allow for the creation of time-mean offline bias maps.

3.2. Neural Network Training

Following B22, we train two fully connected dense NNs to separately predict (a) vertical profiles of air temperature and specific humidity tendencies and (b) column shortwave transmissivity and downward longwave surface flux. We train four NNs for each set of outputs using different random seeds and also construct an ensemble model using all of the randomly seeded NNs which outputs the median prediction of the ensemble members for each field.

Choices of width, depth and learning rate were guided by a hyperparameter sweep in a randomized grid search (Biewald, 2020). To speed the hyperparameter sweep, we used the hyperband algorithm (Li et al., 2018) for early stopping, where training was terminated after 10 epochs if validation loss was not improved relative to previously tested sets of hyperparameters. Validation loss was primarily affected by the choice of learning rate; width and depth had much less impact on model skill. We first chose the value of learning rate that performed best in the sweep, and then set the network width and depth using the set of sweep parameters that performed best near that value of learning rate.

3.2.1. Air Temperature and Specific Humidity Tendencies NN

The NN trained to predict corrective air temperature and specific humidity tendencies is a fully-connected dense network with 3 hidden layers with 419 neurons per layer. It is trained for 500 epochs with early stopping and a learning rate of 1.4×10^{-4}. Its input features are the cosine of solar zenith angle, surface geopotential, latitude, and the vertical profiles of air temperature and specific humidity. The surface geopotential implicitly provides information about the surface type (land or sea/sea ice). Latitude is included as a new feature not used by B22, because it improves offline model skill at high latitudes, where there are extreme conditions but relatively few columns contributing to the loss function.

As in B22, inputs to the network are normalized via standard scaling each vertical level using the mean and standard deviation of the first 4×10^5 samples. Output profiles passed to the loss function are normalized by the standard deviation of each vertical level such that all vertical levels are equally weighted in the loss. We use the same mean absolute error loss and L2 regularization penalty of 10^{-4} as B22.

The following new ML configuration options helped ensure online stability:

- ML-predicted tendencies of heating and humidity are limited to magnitudes less than 0.002 K/s and 1.5×10^{-6} kg/kg/s, respectively. These limiters are applied as a layer within the dense NN such that the limited outputs are used during optimization. These ranges comfortably extend beyond the nudging tendency minima and maxima in the training data by a factor of 3, but prevent the NNs from making extreme predictions when undesirable feedback between the coarse-grid model and ML corrections lead to atmospheric input states outside the envelope of the training data.
- Following Clark et al. (2022), we exclude ("clip") the uppermost 25 vertical model levels ($\lesssim$150 hPa) of specific humidity and air temperature state inputs from the feature set. Without doing this, the models' Jacobian matrices showed that the output fields in the boundary layer were as sensitive to inputs from the uppermost 25 model levels as they were to input levels in their immediate locality (Brenowitz & Bretherton, 2019).
- The uppermost three vertical levels of temperature and humidity tendency outputs are excluded from the prediction. The ML model always applies a zero corrective tendency for these levels when used online. Differences in the sponge layer damping between FV3GFS and the X-SHiELD reference model lead to large nudging tendencies in these few levels with magnitudes similar to those in the boundary layer, but we do not consider these differences to be part of the coarse model physics that we wish to correct. The output clipping here is less extensive than the approach taken in Clark et al. (2022), where the output tendencies were tapered to zero at the top of the model in the uppermost 25 levels. We found that clipping just the uppermost three levels provided similar benefits in terms of online stability and performance.

3.2.2. Surface Radiative Flux NN

We train a NN to predict surface downward longwave radiative flux and column shortwave transmissivity to correct for the effect of systematic cloud biases on the land surface in the coarse run (Section 2.3). Its input features are the cosine of solar zenith angle, surface geopotential, latitude, and the vertical profiles of air temperature and specific humidity. Transmissivity is set to zero in the training data for nighttime columns with zero solar insolation. The predicted column transmissivity is multiplied by the top-of-atmosphere downward shortwave flux to infer the predicted downward shortwave flux at the surface. This approach, introduced by Clark et al. (2022), differs slightly from the radiative flux NN in B22, which directly predicted the downward and net surface shortwave radiative fluxes and did not include latitude as an input feature.

Like the tendency model, the surface radiative flux NN is a fully-connected dense network with 3 hidden layers of width 419, trained for 500 epochs with early stopping and mean absolute error loss. It uses a learning rate of 4.9×10^{-5} and L2 regularization penalty of 10^{-4}.

Longwave flux outputs are enforced to be positive or zero, and transmissivity outputs are limited to the range [0, 1]. As in the tendency NN, these limits are applied as a dense network layer such that the limited outputs are used in the loss function.

3.3. Sensitivity to Data Sampling and NN Configuration

We had to update our training methodology from B22 for the ML corrections to be stable and skillful over a year-long simulation. Early NN versions using similar training data set size and hyperparameters as in B22 developed

Table 1
Sensitivity of Offline Column-Integrated Machine Learned (ML)-Predicted Heating $\langle heating \rangle_{ML}$ R^2, Column-Integrated ML-Predicted Moistening $\langle moistening \rangle_{ML}$ R^2, and Mean 150–400 hPa ML Heating Bias at the Poles ($|lat| > 60°$) to Various Changes in the Data Sampling and Neural Network Configuration

Configuration	$\langle heating \rangle_{ML}$ R^2	$\langle moistening \rangle_{ML}$ R^2	Polar 150–400 hPa ML heating bias (K/s)
Base	0.17	0.15	1.1E−06
Base + lat	0.18	0.16	8.1E−07
Base + lat + sampling	0.18	0.15	7.0E−07
Base + lat + lower LR	0.28	0.22	−1.8E−07
Base + lat + sampling + lower LR	0.29	0.23	2.2E−07

large drifts over a year-long simulation, with some random seed variants crashing before a full year completed. A leading indicator of drift and instability was a steady warming of 200 hPa air temperature starting at the poles and growing to a ∼20 K warm bias that extended into the mid-latitudes within 90 days. This online bias was linked to a positive offline bias ML heating tendencies near the poles in the upper troposphere.

Several changes described in Section 3.2 improved our offline skill and subsequently reduced the online growth of stratospheric air temperature biases. In Table 1 we present the impact of those configuration choices on offline skill. The last column of Table 1 lists the polar regions' mean offline bias in 150–400 hPa heating tendencies; though this metric is calculated offline we found it to be a useful proxy for online air temperature drifts in the first few months of the simulations.

Updates are described relative to a starting "base" configuration, which used the same data set size as B22 (130 timesteps randomly chosen from the year of data, with no downsampling of columns) as well learning rate (0.002) and number of training gradient descent steps ($\sim 2.9 \times 10^7$). The input clipping, limiters on output magnitudes, and the clipping of the uppermost 3 vertical tendency levels (described in Section 3.2.1) are present in all the configurations tested below. Table 1 compares the following configuration choices, applied sequentially:

+ **lat** Latitude is included as an input feature.
+ **sampling** The number of timesteps in the training data set is increased >10× and subsample down to 15% of the grid columns in each timestep.
+ **lower LR** The learning rate is lowered from the value used in B22 from 0.002 to 0.0014 and the number of gradient descent steps is increased >30×.

Training with a lower learning rate over more gradient descent steps is the main contributor to the increased global offline R^2 of column-integrated tendencies.

4. Offline Performance

Here we discuss the ML models' "offline" skill in predicting their training data targets over a single time-step. Offline skill is a necessary but not sufficient condition for successful application of the corrective ML in the coarse model. However, we do not strictly aim to maximize offline skill, as our choice of L2 regularization in model training (which slightly lowers offline skill) is required for online stability.

As in Figure 4 of B22, the instantaneous nudging tendencies at any given timestep are quite noisy, so the ML model cannot be expected to have perfect skill. In Figure 2 we show the zonal and pressure-level mean coefficient of determination R^2 on the offline testing data for the ML-predicted corrective tendency fields. Both temperature and humidity tendency predictions are most skillful in the boundary layer and in the tropics, with zonal-mean R^2 values upwards of 0.8 in the tropical boundary layer and 0.5–0.8 in the tropical free troposphere and extratropical boundary layer. Model skill in the mid-to-upper troposphere degrades to 0.1–0.3 at higher latitudes. The global-mean vertical profiles of R^2 (not shown) are much improved relative to Figure 5 of B22, increasing up to a factor of two in the lower to mid-troposphere. This improvement in offline skill from B22 results from the updates to our training methodology and configuration described above.

As described in Section 3.2.2, the ML model for downwelling surface radiation is slightly different from B22. Time-mean offline biases for the downward radiative fluxes are shown in Table 2. Their global average biases are

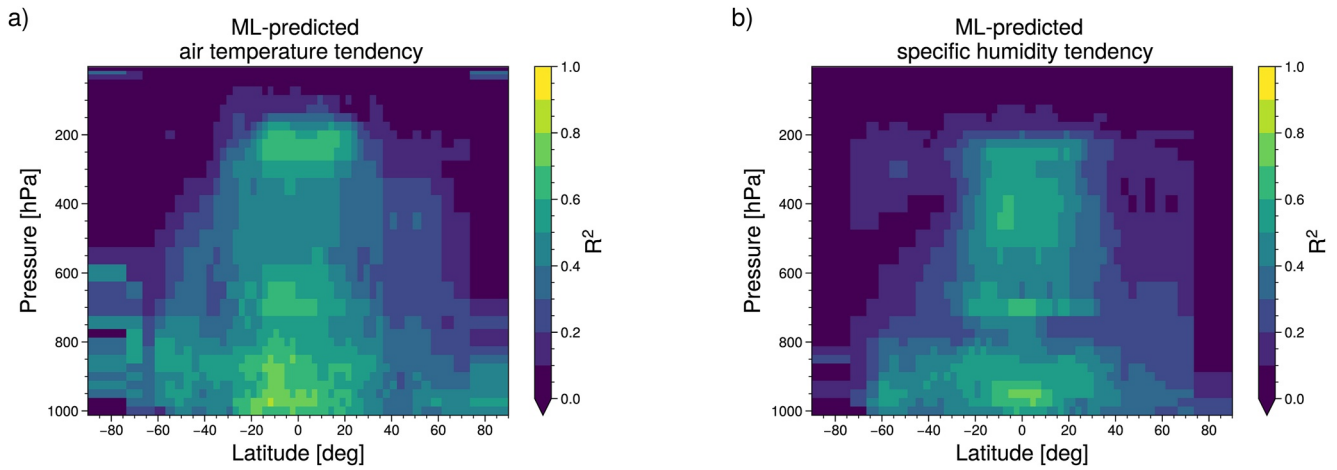

Figure 2. R^2 of offline predictions of air temperature and humidity tendencies, averaged over latitude and pressure. Predictions are generated using the neural network ensemble; results for individual seeds are similar.

small, regardless of neural-net random seed choice: 1.2 ± 0.7 W/m^2 for downward shortwave and -0.3 ± 0.2 W/m^2 for downward longwave surface radiative flux. The RMSE of the time-averaged total downward flux prediction is significantly reduced from 11.6 W/m^2 in B22 to 3.4 W/m^2 here.

5. Online Performance

The true test of the corrective ML comes by applying it online in a prognostic simulation of the coarse-grid FV3GFS model and measuring results over seasonal and yearlong timescales. Here, we present results from a suite of five such ML-corrected simulations. Four use independent random seeds to initialize the tendency and radiative flux NNs. A fifth uses the median of the corrective tendencies and surface fluxes over this four-member NN ensemble. All ML-corrected simulations ran stably for the full simulation length of 360 days.

For each 1-year ML-corrected simulation we assess the global-mean biases and seasonal-mean spatial pattern errors of the precipitation and land surface temperature with respect to the X-SHiELD reference run. These are compared to the baseline coarse-grid FV3GFS simulation. All coarse-grid FV3GFS simulations use the same namelist configuration, initial conditions, and prescribed sea ice and SSTs. Ideally, the biases and pattern errors of the ML-corrected simulations will be significantly smaller than those of the baseline simulation. We also examine the effects of including the ML correction on other measures of large scale circulation and climate drift in the coarse simulations.

We also tested two methods of applying non-state-dependent corrections at each timestep. Applying the spatially-smoothed, monthly-mean air temperature and specific humidity nudging tendencies in each column was unsuccessful and simulations crashed within 2 weeks. A different approach used the monthly-mean baseline bias to derive the corrective tendency in each column; this simulation ran stably for 1 year. Its time-mean land surface temperature and precipitation errors were on par with the baseline simulation, and larger than those of the ML-corrected simulations. It did have marginally improved precipitation skill over ocean compared to the ML-corrected simulations. More details (including bias maps) are given in Appendix A. In the remainder of this section we limit our comparisons to the baseline and ML-corrected simulations.

Table 2
Offline Metrics for Downward Surface Fluxes

Surface radiative flux field	RMSE (W/m^2)	Bias (W/m^2)	R^2
Downward shortwave	3.8	1.2	0.99
Downward longwave	1.2	−0.3	0.99
Total downward	3.4	0.9	0.99

Note. Downward shortwave and longwave fluxes are predicted by the radiative flux neural network; total downward flux is the sum of shortwave and longwave predictions.

5.1. Improvements in Precipitation and Surface Temperature

Figure 3 displays the RMSE and bias of time-averaged surface precipitation over the simulation year. The ML corrections from the individual NNs and the NN ensemble all improve the coarse model's time-mean precipitation skill over both land and ocean. As we only sample four randomly seeded models, we will cite the minimum and maximum relative improvement in

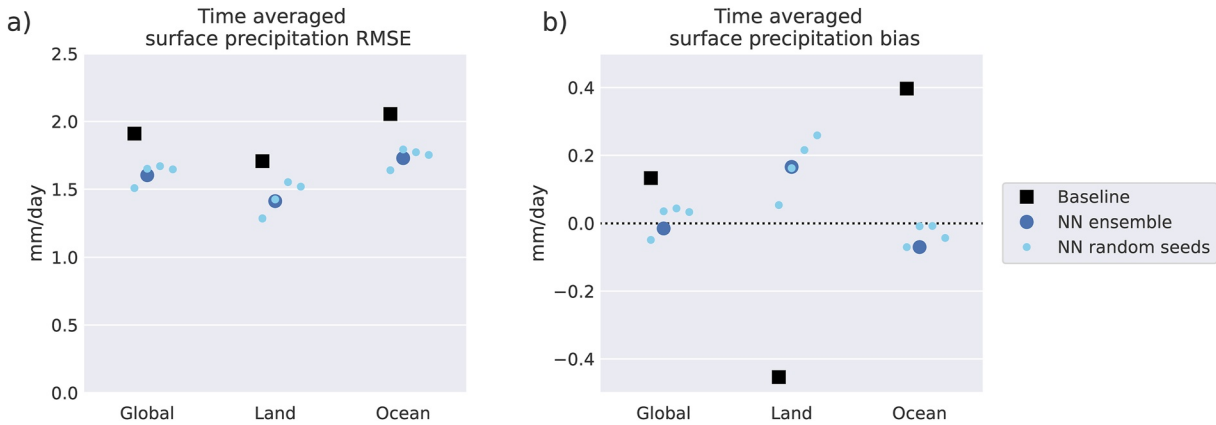

Figure 3. Time-mean root mean squared error and bias (with respect to the fine-grid reference) of precipitation in the baseline and machine learned-corrected coarse runs, shown for global, land, and sea domain averages.

each metric across the seeds instead of a standard deviation. ML-corrected models improve upon the baseline precipitation RMSE by 13%–21% globally, 9%–25% over land, and 13%–20% over ocean. The magnitude of the precipitation bias is strongly reduced by 63%–89% globally, 43%–88% over land, and 82%–98% over ocean.

The time-averaged precipitation error pattern is qualitatively similar across the various NNs. For conciseness we only show the ML-corrected simulation using the NN ensemble in Figure 4a. The baseline coarse run tends to have a large negative bias in precipitation over equatorial Africa and South America, and a large positive bias over

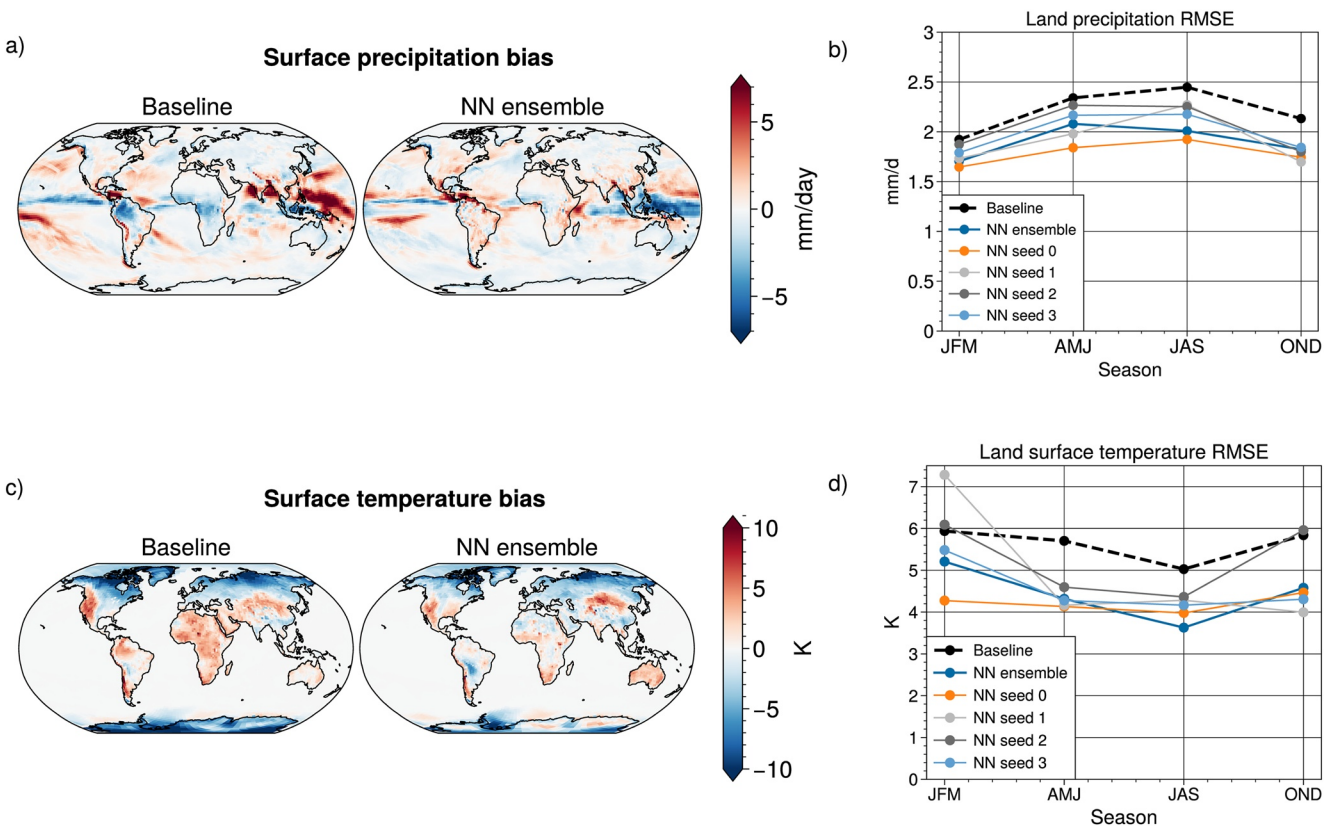

Figure 4. Top: (a) Time-mean error pattern map of surface precipitation relative to the fine-grid model. The machine learned (ML)-corrected run uses the neural network ensemble. (b) Seasonal root mean squared error of land surface precipitation for the coarse model baseline and ML-corrected simulations. Bottom: As above, but for land surface temperature. Surface temperature is only shown over land and sea ice, as the sea surface temperature is prescribed from the fine-grid reference.

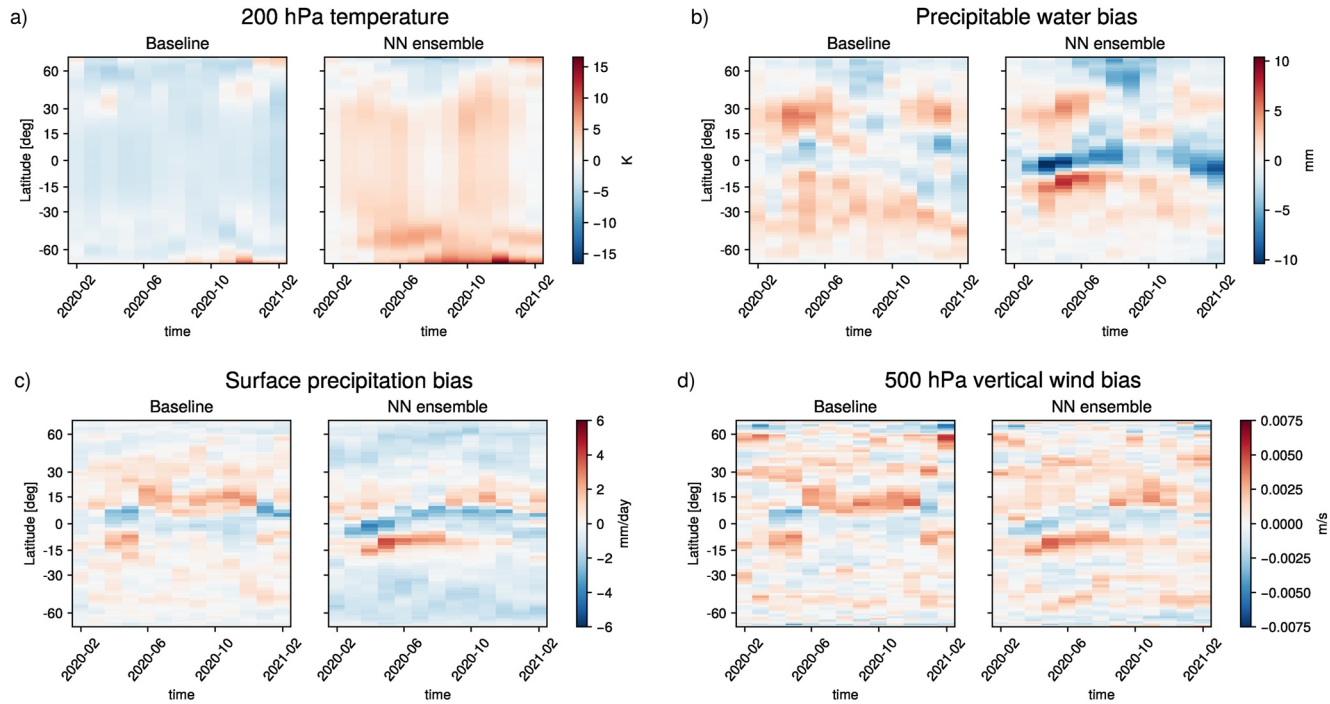

Figure 5. Zonal mean versus time plots for biases (with respect to the fine-grid model) in 200 hPa temperature, precipitable water, precipitation, and 500 hPa vertical wind in the baseline and machine learned-corrected prognostic runs.

the western Pacific warm pool. The ML-corrected model reduces the magnitude of these regional biases. The baseline's wet bias in the western Pacific is replaced with a dry bias of slightly smaller magnitude. The pattern error of precipitation in the ML runs in the tropics differs significantly from the baseline due to the changes in circulation brought about by the ML correction (see Section 5.2).

Figure 4b shows the seasonally averaged errors in precipitation over land in the baseline and ML-corrected runs. The improvement in land precipitation errors is robust across all seasons in all four NNs as well as the NN ensemble.

The 10%–20% relative improvement over the baseline in the RMSE of time-averaged surface precipitation is less than the ∼30% reduction found by B22. We discovered that enabling fast saturation adjustment on each of the six dynamics substeps within the dynamical core reduced the baseline model precipitation RMSE from 3.7 mm/day over the 40 day run in B22 down to 1.9 mm/day in annual average here. Thus, the absolute errors in the baseline as well as the ML-corrected runs are notably improved in this work over B22.

Land surface temperature errors are also reduced by 6%–26% in the ML-corrected runs. Figure 4c shows the annual-average land surface temperature pattern error in the baseline and NN ensemble runs. The ML-corrected runs reduce a warm bias across Africa and the western United States. Seasonal surface temperature RMSE is plotted in Figure 4d. While ML-corrected runs consistently improve land surface temperature from April through September, their behavior in boreal winter is less consistent across seeds, with NNs generated from two seeds having comparable or worse skill than the baseline in these months. Those simulations amplify a systematic cold bias during boreal winter at high northern latitudes ($\gtrsim$50°N) also present in the baseline, while ML-corrected runs with other seeds reduce this bias (not shown).

Both land surface temperature and precipitation show the largest seasonal skill improvements during April - September, with the best-performing model (seed 0) reducing both land temperature and precipitation seasonal RMSEs by over 20% in these seasons.

5.2. Other Climate Biases in ML-Corrected Simulations

B22 found that 40-day coarse-grid simulations with and without corrective nudging both developed mean-state biases in the latitudinal structure of upper-tropospheric (200 hPa) temperature which could be 5 K or more in parts of the extratropics (their Figure 13d). The full year of reference X-SHiELD simulation data allows us to

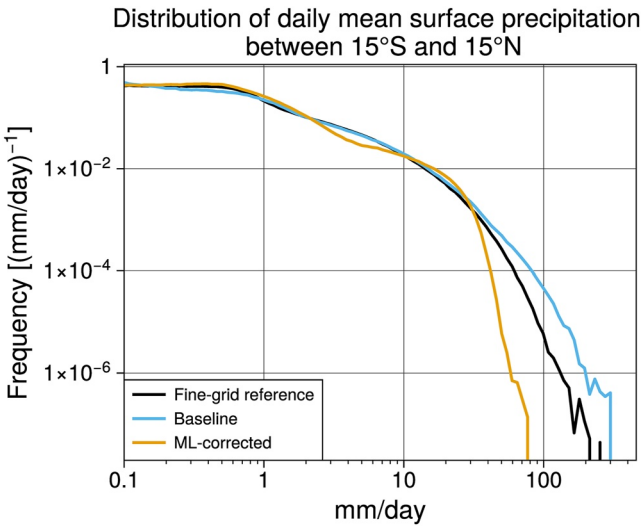

Figure 6. Distribution of tropical (15°S–15°N) daily mean surface precipitation in the fine-grid, baseline, and ML-corrected simulations.

test how these and other biases develop over longer timescales. We compare just the NN ensemble model to the baseline, as the biases discussed here are robust across all ML-corrected simulations.

Our ML-corrected simulations develop a warm bias in 200 hPa air temperature (Figure 5a). This bias is no more than ∼5 K over most of the globe, but is substantially larger at high southern latitudes. In contrast to B22, all the NNs in this work drift similarly over the year-long run. The baseline run develops a smaller cold bias at most latitudes.

The ML-corrected runs also develop robust bias patterns in precipitable water (Figure 5b). A strong dry bias of up to −10 mm develops over the seasonally shifting GSRM-simulated tropical ocean ITCZs, as well as a northern summertime dry bias at high latitudes northwards of ∼50°. The Northern Hemisphere dry bias is also evident in the baseline run, albeit to a lesser magnitude. The baseline model does not share the dry ITCZ bias, but is too moist in most other parts of the subtropics.

A related bias of the ML-corrected runs is a weakened Hadley circulation that is also shifted southward during boreal spring and summer. Figures 5c and 5d show the biases in precipitation and 500 hPa vertical wind in the baseline and ML-corrected runs with respect to the coarsened fine-grid reference. The baseline model has a northward shift of vertical wind and precipitation in the tropics relative to the fine-grid reference during Jun-Aug. In contrast, the ML-corrected run displays southward-displaced upward motion and precipitation during Jun-Aug.

We do not yet have a remedy for these ML-induced climate biases. Like B22, we tried including a NN that learned corrective tendencies for the zonal and meridional winds. B22 found that including ML-corrected wind tendencies led to large time-mean 200 hPa air temperature biases developing over their 40-day run. In the present study, we similarly found that including corrective wind tendencies caused errors in 200 hPa temperature that grew quickly within the first 7 days of the run and ultimately led to numerical instability. In a forthcoming publication, we will explore the use of additional online ML models to restrict the application of ML corrections to states within the training data envelope defined by the range of atmospheric column states in the reference simulation. The goal is to keep the model from pushing farther outside this envelope, inhibiting instability and perhaps leading to a more accurate simulated climatology.

5.3. Tropical Variability

This section documents that the space-time variability of tropical precipitation is a weak point of our ML-corrected simulations, even though they improve aspects of the seasonal mean precipitation. We believe this is mainly caused by training using nudging tendencies. Because the coarse-grid convective parameterization includes a trigger that is highly sensitive to the column thermodynamic state, the precipitation predicted by physical parameterizations of the nudged coarse simulation is intermittent and poorly synchronized with the more smoothly varying precipitation in the reference fine-grid model. The temperature and humidity nudging tendencies mediate this disagreement by damping the temperature and humidity tendencies in the coarse-grid physics; this is learned by the ML correction. Hence, the ML correction tends to reduce the sensitivity of the physical parameterizations to column thermodynamic state, reducing extreme precipitation and tropical variability of precipitation.

Figure 6 shows the PDF of daily precipitation between 15°S and 15°N. The baseline run has excessive extreme precipitation but the ML-corrected simulation has too little extreme precipitation.

Figure 7 plots tropical (15°S–15°N average) precipitation across longitude and time in the fine-grid and free-running coarse-grid simulations. Eastward-propagating Kelvin waves and hints of a Madden-Julian Oscillation stand out in the fine-grid reference but are largely absent in both coarse-grid simulations. The baseline run has overly-strong precipitation in the western Pacific (∼150°E–200°E) and too much power in westward-propagating waves; in contrast, the ML-corrected model largely damps out zonally propagating waves in both directions and underpredicts western Pacific precipitation.

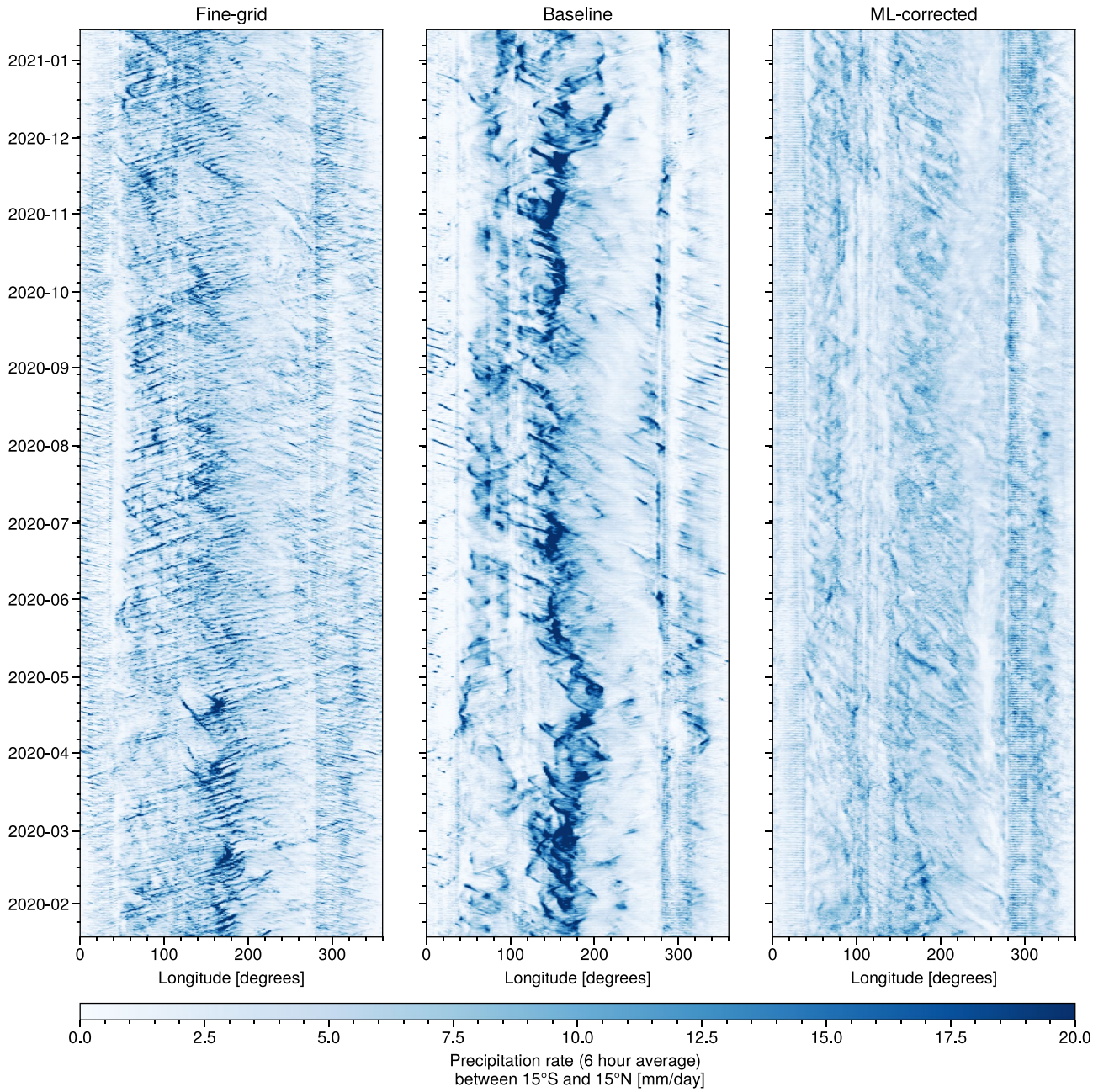

Figure 7. Longitude-time plots of 6-hourly averaged precipitation rate between ±15° latitude in the fine-grid, baseline, and ML-corrected simulations.

5.4. Stability Over Multiple Years

Would the climate in ML-corrected runs continue to drift, with larger biases developing over multiple years of simulation, or are its biases largely repeatable in subsequent annual cycles, as in the no-ML baseline? This inter-annual stability under constant SST forcing is a prerequisite for use in climate-length simulations.

To address this question, we used the NN with the lowest surface temperature and precipitation errors (seed 0) for a 5-year coarse-grid run. All prognostic runs so far used the SSTs from the yearlong reference X-SHiELD simulation, which did not span the full length of the extended 5 year simulation and did not exactly repeat at the end of the annual cycle. To enable a smoothly-forced 5-year simulation, we instead used a climatological annual cycle of SSTs.

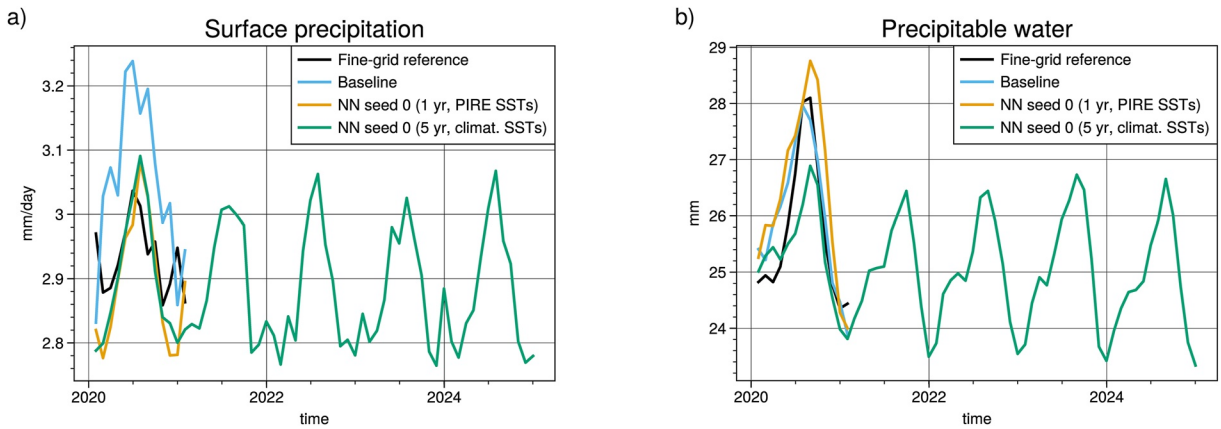

Figure 8. Globally-averaged time series of monthly-mean surface precipitation and precipitable water from the reference fine-grid, baseline coarse-grid, and machine learning (ML)-corrected coarse-grid simulations. Unlike the yearlong baseline and ML runs, the 5 year ML run uses climatological sea surface temperatures and is thus not limited to the time range of the reference run.

Figure 8 shows time series of precipitable water and precipitation in this extended ML-corrected run as well as the yearlong fine-grid reference, baseline, and ML-corrected runs with SSTs prescribed from reference data. As desired, the multiyear ML-corrected run maintains consistent seasonal bias patterns across years that match up well in its first year to the fine-grid X-SHiELD reference simulation.

6. Conclusions

This study extends previous work done in B22 in which corrective ML models trained using fine-resolution data were applied within coarse-grid climate models. The novelty of this work lies in the training and evaluation of the corrective ML over the entire annual cycle. This advance was enabled by the use of the year-long X-SHiELD simulation reference data set. The longer-term goal is to use GSRM simulations in multiple climates (Cheng et al., 2022) to train corrective ML that can be used in climate change simulations, following the template in Clark et al. (2022), who used 25-km grid reference simulations in multiple climates to this end.

We show results from multiple 360-day-long coarse-grid prognostic runs using four randomly seeded pairs of NN models for temperature and humidity tendencies and radiative surface fluxes as well as an ensemble of the four pairs of NNs. Bias and RMSE are reported with respect to the fine-grid X-SHiELD reference model. We observe robust improvements across all NNs tested in time-mean land surface precipitation (9%–25% lower RMSE) and land surface temperature (6%–26% lower RMSE) with respect to a non-ML-corrected baseline simulation. Seasonally averaged land surface precipitation RMSE is also robustly improved across all seasons for all ML-corrected runs. Seasonally averaged land surface temperature RMSE is consistently improved during boreal summer across ML models, and two of four also reduce the boreal winter cold biases at high northern latitudes.

All ML-corrected simulations ran to completion without any crashes or runaway drifts. We tested our best-performing model (seed 0) over a 5 year simulation. It maintains a stable climate throughout the run, with consistent seasonal cycles year over year that closely repeat its first-year behavior.

Like the baseline model, the ML-corrected prognostic simulations develop significant seasonal biases in precipitable water, precipitation and vertical motion. The ML models somewhat overcorrect many of the tropical circulation biases of the baseline model.

Tropical precipitation variability is significantly reduced in the ML-corrected runs. While the baseline run produces too much extreme precipitation, the ML-corrected runs produce too little. Both the baseline and ML-corrected runs have weaker MJO and Kelvin wave propagation than the fine-grid reference, with the ML-corrected run having the weakest wave propagation of the three.

Our results are encouraging for the prospect that ML trained on GSRM simulations can improve coarser-grid climate models. To realize this prospect, future work is still needed to improve the climate drift, circulation biases, and precipitation variability of ML-corrected runs.

We focused on demonstrating the viability of the corrective ML approach to improving coarse-grid simulations over the annual cycle. Because our reference fine-grid simulation covered 1 year using observed SSTs, it was not exactly comparable with coarse simulations with annual repeating SST patterns. Hence, we focused on comparing this reference with single year simulations using a suite of five ML-corrected simulations. Longer reference and coarse simulations would enable more precise climatological comparisons. It might also be interesting to look at how corrective ML affects the spread of an ensemble of global forecasts with perturbed initial conditions.

Another longer-term goal is to run an ML-corrected atmospheric model coupled simulation with an no-ML (or ML-enhanced) ocean model, as recently explored by Arcomano et al. (2022). For this purpose, the ML correction should improve (relative to reference data) the surface fluxes of heat, fresh water, and momentum over the world oceans.

Appendix A: Non-State-Dependent Corrective Tendencies

In addition to the free-running baseline control experiment, it is also useful to compare the performance of the machine learned (ML)-corrected simulations to one using a simpler method of coarse-resolution corrections. Do the ML state-dependent corrections perform significantly better than non-state-dependent, localized corrections? We tested two variations of the non-state-dependent approach by applying corrective tendencies at each timestep derived from either (a) the time-averaged nudging tendencies or (b) the baseline simulation biases.

To test (a) we performed a free-running simulation with the monthly-averaged nudging tendencies at each grid point applied at each timestep. Tendencies were linearly interpolated in between monthly averages to avoid sharp discontinuities between months. We tried horizontally smoothing the nudging tendencies with kernels of various sizes, in case grid-scale noise was causing instabilities. However, all variations of this approach crashed within 2 weeks of the January initialization.

We then tested a gentler approach (b) of using the monthly-mean baseline run bias at each grid point, calculated with respect to the fine-grid reference. The corrective tendency applied at each timestep in this case was the inverse monthly mean baseline bias divided by a timescale of 1 month. The magnitudes of these tendencies are up to an order of magnitude smaller than the monthly averaged nudging tendencies (which used a much shorter 3-hr nudging timescale). This simulation ran stably for a year.

Figures A1 and A2 show online results of this approach compared to the baseline and ML-corrected runs. The global time-mean precipitation root mean squared error (RMSE) is improved at a level comparable to values in the ML-corrected simulations. This is almost entirely due to improved precipitation skill over ocean. The RMSE of land surface precipitation in the bias-corrected run is slightly worse than in the baseline, and its bias is almost twice as large.

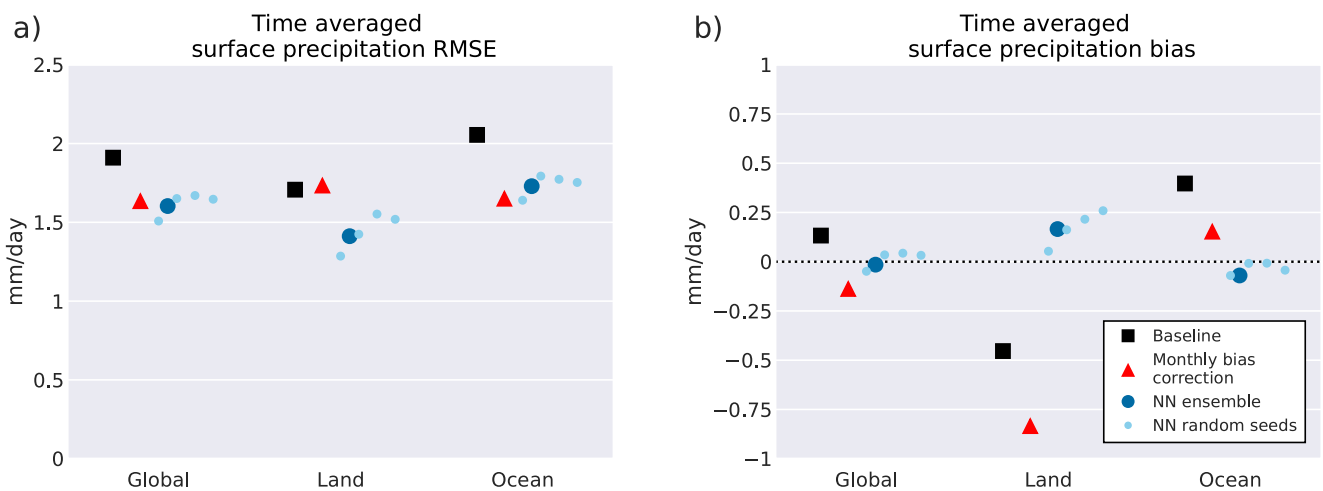

Figure A1. Similar to Figure 3 showing surface precipitation metrics, but with the month-averaged bias-correction approach included.

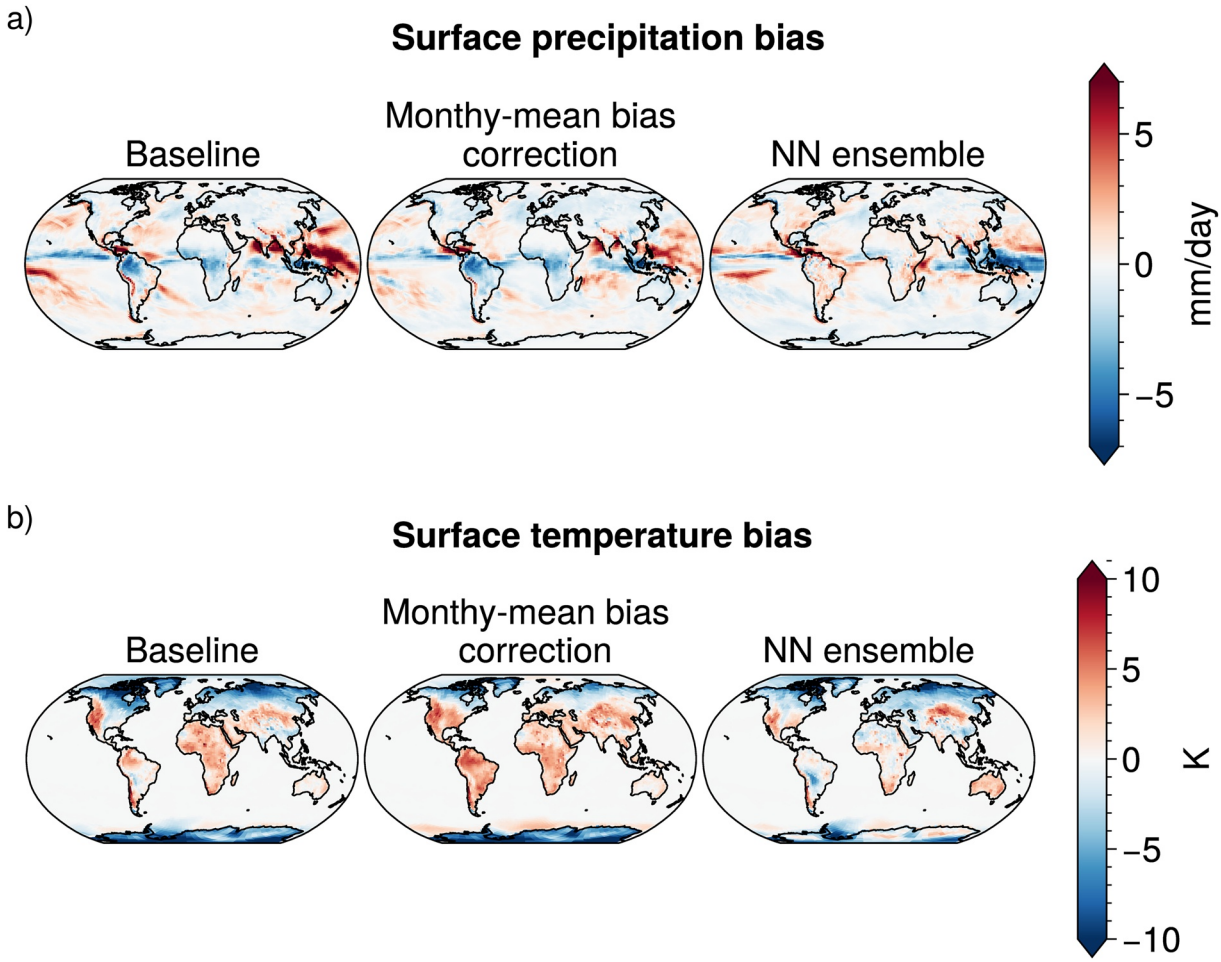

Figure A2. Similar to Figure 4 showing time-mean errors in surface precipitation (a) and land surface temperature (b), with the month-averaged bias-correction approach included.

Land surface temperature RMSE is reduced by 7% relative to the baseline. For comparison, the simulation using the neural network ensemble reduced the land surface temperature RMSE by 23% over the baseline, and the range of improvement across the simulations using individual seeds was 6%–26%. This approach reduced the magnitude of the cold surface temperature bias in the higher northern latitudes, but led to increased warm biases elsewhere, notably in North and South America and central Asia.

Data Availability Statement

The code and experiment configurations needed to reproduce this work are available at the Github repository https://github.com/ai2cm/nudge-to-3km-PIRE-1yr-workflow which is archived at Zenodo (https://doi.org/10.5281/zenodo.7063087). The coarsened fine-grid data used for initial conditions and in the nudged coarse-grid simulation is available upon request through a Google Cloud Storage "requester pays" bucket. The ERA5 data was obtained at https://doi.org/10.24381/cds.f17050d7. The GPCP data was obtained at https://psl.noaa.gov/data/gridded/data.gpcp.html.

Acknowledgments
We thank the Allen Institute for Artificial Intelligence for supporting this work and NOAA-GFDL for running the 1-year X-SHiELD simulation on which our ML is trained using the Gaea computing system. We also acknowledge NOAA-GFDL, NOAA-EMC, and the UFS community for making code and software packages publicly available.

References

Adler, R. F., Sapiano, M. R. P., Huffman, G. J., Wang, J.-J., Gu, G., Bolvin, D., et al. (2018). The Global Precipitation Climatology Project (GPCP) monthly analysis (new version 2.3) and a review of 2017 global precipitation. *Atmosphere*, *9*(4), 138. https://doi.org/10.3390/atmos9040138

Arcomano, T., Szunyogh, I., Wikner, A., Hunt, B. R., & Ott, E. (2022). A hybrid atmospheric model incorporating machine learning can capture dynamical processes not captured by its physics-based component. https://doi.org/10.22541/essoar.167214579.97903618/v1

Biewald, L. (2020). Experiment tracking with weights and biases. Retrieved from https://www.wandb.com/

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *Journal of Advances in Modeling Earth Systems*, *11*(8), 2728–2744. https://doi.org/10.1029/2019MS001711

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

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

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz-Sabater, A., et al. (2019). *ERA5 monthly averaged data on single levels from 1979 to present*. Copernicus Climate Change Service (C3S) Climate Data Store (CDS).

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., & Talwalkar, A. (2018). Hyperband: A novel bandit-based approach to hyperparameter optimization. *Journal of Machine Learning Research*, *18*(185), 1–52. Retrieved from http://jmlr.org/papers/v18/16-558.html

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

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

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. (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

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., Harris, L., & Chen, J.-H. (2022). *The GFDL cloud microphysics parameterization (technical report)*. Geophysical Fluid Dynamics Laboratory (U.S.). https://doi.org/10.25923/pz3c-8b96

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