Correcting a coarse-grid climate model in multiple climates by machine learning from global 25-km resolution simulations

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

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

Key Points:

- Machine learning models trained from fine-grid outputs correct the evolution of coarse-grid models in four climates
- Ablating upper level inputs and outputs of machine learning models robustly stabilizes multi-year simulations
- Trained models reduce rainfall and surface temperature errors over land in five-year simulations in each climate

Corresponding author: Spencer K. Clark, `spencerc@allenai.org`

Abstract

Bretherton et al. (2022, `https://doi.org/10.1029/2021MS002794`) demonstrated a successful approach for using machine learning (ML) to help a coarse-resolution global atmosphere model with real geography (a ∼200 km version of NOAA's FV3GFS) evolve more like a fine-resolution model. This study extends that work for application in multiple climates and multi-year ML-corrected simulations. Here four fine-resolution (∼25 km) two-year reference simulations are run using FV3GFS with climatological sea surface temperatures perturbed uniformly by −4 K, 0 K, +4 K, and +8 K. A dataset of state-dependent corrective tendencies is then derived through nudging the ∼200 km model to the coarsened state of the fine-resolution simulations in each climate. Along with the surface radiative fluxes, the nudging tendencies of temperature and specific humidity are machine-learned as functions of the column state. ML predictions for the fluxes and corrective tendencies are applied in 5.25 year ∼200 km resolution simulations in each climate, and improve the spatial pattern errors of land precipitation by 17 % to 30 % and land surface temperature by 20 % to 23 % across the four climates. The ML has a neutral impact on the pattern error of oceanic precipitation.

Plain Language Summary

Previous work demonstrated how to use machine learning to help a computationally efficient coarse-grid climate model behave like a more realistic, but expensive, fine-grid reference simulation that we could only afford to run for 40 days. The machine learning was interpreted as correcting errors in the representation of uncertain small-scale cloud, precipitation, and turbulence processes on the model simulations. By using a fine-grid model with a grid spacing eight times as large as our previous reference that runs tens of times faster, we extend that approach to multi-year coarse-grid simulations of a range of climates, both warmer and colder than the present day. Different random starting guesses ('seeds') lead to slightly different machine learning corrections even with exactly the same training protocol. When applied interactively in one-year coarse-grid simulations, the machine learning corrections consistently improve the time-mean pattern of rainfall and surface temperature over land vs. fine-grid reference simulations in each of the climates we trained against. These machine learning models can be used successfully to enhance the accuracy of five-year simulations in all climates.

1 Introduction

To make accurate and precise predictions of climate change, global climate models (GCMs) should realistically include and resolve as many physical processes as possible. However, computational power is an important constraint, so trade-offs must be considered, e. g. between grid spacing and subgrid parameterization. Current GCMs with grid spacings of 50 km or more can be affordably run for thousands of years, using physical parameterizations for subgrid-scale processes such as cumulus convection and gravity wave drag. However, these parameterizations are a major source of uncertainty (Shepherd, 2014), and as a result, even the same model, when run at finer resolution, might project different regional patterns of climate change (van der Wiel et al., 2016). Furthermore, spatial resolution trade-offs mean coarse-grid simulations often cannot represent important processes like rainfall as well as finer grid runs (e.g., Stevens et al., 2020; Caldwell et al., 2021).

Through the use of machine learning (ML), it may be possible to improve affordable coarse-grid model simulations by leveraging output from finer-grid runs. This has been demonstrated in idealized settings by Brenowitz and Bretherton (2019), Yuval and O'Gorman (2020), Yuval et al. (2021), and Yuval and O'Gorman (2021), and recently in a real-geography setting in Bretherton et al. (2022), hereafter referred to as "B22." In Brenowitz and Bretherton (2019), Yuval and O'Gorman (2020), Yuval et al. (2021), and Yuval and O'Gorman (2021), ML models were trained using coarse-grained outputs of fine resolution reference runs to fully represent the apparent sources (Yanai et al., 1973) of temperature, specific humidity, or horizontal momentum of the coarse model, while in B22 corrections to the parameterized apparent sources were learned. In each of these studies, when run with these ML tendencies included, aspects of the coarse simulations behaved more like the coarsened fine resolution model.

In this study we extend the corrective ML approach introduced in Watt-Meyer et al. (2021), hereafter "W21," and B22, to multi-year simulations in multiple climates. Based on the output of coarse-grid simulations that were nudged to observational analysis or the coarsened state of a fine-grid model, W21 and B22 trained machine learning models to predict corrections to the physical parameterization tendencies of a full-geography coarse-grid model in the present-day climate. When applied in otherwise free-running prognostic simulations, these corrections, among other things, brought the precipitation

climatology of the coarse model closer to that of observations or a fine-grid reference. We apply similarly obtained ML corrections in free-running prognostic simulations in multiple climates, and quantitatively evaluate their impact on improving selected climate metrics compared to baseline simulations without ML corrections. Biases are calculated with respect to the fine-grid reference simulations. Because this ML approach optimizes only the single timestep evolution versus the fine-grid reference simulations, it is not guaranteed to yield stable simulations with smaller long term mean biases in all climates than for the baseline model.

To keep the scope manageable, our simulations use specified sea-surface temperature (SST) distributions to which globally uniform offsets are added to generate colder and warmer climates. We use a $\sim$25 km grid version of our climate model as our fine-grid reference, and a $\sim$200 km grid version of the same model with the same set of parameterizations serves as the coarse-grid model whose baseline (no-ML) simulations are to be improved using the ML. Eventually, like B22, we would like to use a global storm resolving model with a 3 km or finer horizontal grid as the reference model, but it is still too computationally expensive to make the multi-year simulations over multiple climates that would entail.

To develop an effective multi-climate scheme, we build upon earlier findings that ML models perform best when making predictions within the bounds of their training data (O'Gorman & Dwyer, 2018; Rasp et al., 2018). New offline results suggest that it may be possible to develop ML parameterization (Beucler et al., 2021) or classification (Molina et al., 2021) schemes that generalize to climates outside their training range. However, to minimize any changes to the method this work is based upon (B22), we choose to focus our offline and online tests on the range of climates present in our training data, since application of the methods of B22 in multiple climates is novel in and of itself.

Our goal is to deploy ML that improves coarse resolution climate simulations of indefinite duration. Recently, an analogous study used the output of a present-climate superparameterized GCM to train deep neural nets to emulate the apparent sources of temperature and humidity generated by the cloud-resolving models running within each GCM grid column (Y. Han et al., 2020; Wang et al., 2021). With an extensive trial-and-error approach, they found an ML configuration that ran stably for five years with time-mean biases in temperature and precipitation relative to the superparameterized refer-

ence simulation that were comparable to a conventional GCM. Here, we also test the approach using five-year ML-corrected runs – significantly longer than those attempted in W21 or B22 – to see how the method performs, not just on the current climate, but also with SSTs ranging from 4 K colder to 8 K warmer.

Section 2 presents our simulation, training, ML, and evaluation methods. Section 3 presents results for both offline and online skill across the selected range of climates. Section 4 presents a discussion and conclusions.

2 Methods

As in W21 and B22, the coarse model we aim to improve is a 79-level C48 (∼200 km) resolution version of NOAA's FV3GFS (`https://github.com/ai2cm/fv3gfs-fortran`), a full-complexity atmosphere model typically used for numerical weather prediction (UFS Community, 2020). It is based on the FV3 dynamical core (Putman & Lin, 2007; Harris et al., 2021) and contains a configurable suite of physics parameterizations. The dynamical core uses the same number of vertical remappings (1) per physics timestep and dynamical substeps per vertical remapping (6) as in W21 and B22. For this work, in terms of physical parameterizations, the model is configured to use the hybrid eddy-diffusivity mass flux turbulence scheme (J. Han et al., 2016), the GFDL microphysics (Zhou et al., 2019), the scale-aware mass flux shallow and deep convection schemes (J. Han & Pan, 2011), the Rapid Radiative Transfer Model for GCMs (Iacono et al., 2008), a gravity wave drag scheme (Alpert et al., 1988), a mountain blocking scheme (Lott & Miller, 1997), and the Noah land surface model (Ek et al., 2003).

These are the same schemes as those used in W21, but there are two configuration differences. The first is that we reduce the physics timestep to 450 s, which is needed to stabilize runs in warmer climates. The second is that we configure the model to be run with some microphysical processes occurring in the vertical remapping loop of the dynamical core in addition to in the physics. This is consistent with our fine-resolution simulations. These are run with 7 vertical remappings per physics timestep, since frequent application of microphysical adjustments leads to a more accurate representation of precipitation (Zhou et al., 2019). Although the coarse-resolution simulations use only one vertical remapping per physics timestep, configuring the microphysics in a consistent way

improves the climatology of precipitation and surface radiative fluxes in baseline runs relative to the fine-resolution reference runs.

Our reference fine grid model is a C384 (∼25 km) version of FV3GFS. It uses the the same vertical levels, physics timestep, and physics configuration as the coarse-grid model, making the fine and coarse model versions identical except for their grid resolution and dynamical substepping frequency, in this case 7 vertical remappings per physics timestep and 8 dynamical substeps per vertical remapping. Thus, the corrective ML is purely accounting for systematic effects of the additional spatial variability captured by the fine-grid simulation but not the coarse simulation. In a practical application, better results might be obtainable by combining corrective ML with tuning of the coarse-model namelist parameters, but we choose to forgo this step for simplicity and clarity of comparison. Our fine-grid reference model resolution differs from B22, who used a C3072 (∼3 km) resolution simulation completed using the NOAA Geophysical Fluid Dynamics Laboratory's SHiELD model (Harris et al., 2020). This choice made it computationally practical to produce years of training/testing data for multiple climates.

Table 1 summarizes the configuration and duration of all the simulations we complete for this study. We describe these runs in more detail in the following subsections.

2.1 Reference simulations

To produce an ML scheme calibrated across the annual cycle in multiple climates, we need at least one full year of training data from a reference fine-grid simulation in each such climate. We include an additional independent year to validate the predictions of the ML models we train offline, and to compare with simulations where we apply the ML predictions online. Accordingly, we run two-year C384 (25 km grid) FV3GFS reference simulations with climatological sea surface temperatures (SSTs) perturbed uniformly by -4 K, 0 K (control climate), $+4$ K, and $+8$ K. From these two year reference simulations, every 15 minutes we output restart files and diagnostics containing the state of the model, which is coarse-grained online following the methodology described in B22 to C48 resolution.

Table 1. The configuration of the simulations used in this study. Their durations in months are shown in each climate in the final four columns.

					Duration (months)			
Description	Initial condition	Resolution	T, q ML	$\mathcal{T}$, $L_{\text{sfc}}^{\text{down}}$ ML	−4 K	0 K	+4 K	+8 K
Spin-up	GFS analysis	C48	-	-	12	12	12	12
Reference	End of spin-up[a]	C384	-	-	24	24	24	24
Nudged	Start of reference[b]	C48	-	-	24	24	24	24
Baseline	Midpoint of reference[b]	C48	-	-	63	63	63	63
ML-corrected	Midpoint of reference[b]	C48	Seed 0 NN	RF	15	15	15	15
ML-corrected	Midpoint of reference[b]	C48	Seed 1 NN	RF	15	15	15	15
ML-corrected	Midpoint of reference[b]	C48	Seed 2 NN	RF	63	63	63	63
ML-corrected	Midpoint of reference[b]	C48	Seed 3 NN	RF	15	15	15	15

[a]Upsampled to C384 resolution using the `chgres_cube` tool.

[b]Coarsened to C48 resolution using method outlined in B22.

169 ### *2.1.1 Control climate reference simulation*

170 The control-climate simulation is forced with historical SST and sea ice conditions.
171 The SSTs are derived from the 1/12° resolution Real Time Global Sea Surface Temper-
172 ature (RTGSST) dataset (Thiébaux et al., 2003), averaged into climatological monthly
173 means across the period 1982 – 2012. For each simulation, SSTs are then interpolated
174 in space and time to the model's grid and the day of the year, repeating annually. The
175 sea ice distribution is derived from 1982-2012 monthly means of the 0.5° resolution Cli-
176 mate Forecast System Reanalysis (Saha et al., 2014). While it was initially intended that
177 the sea ice distribution would vary with the annual cycle, instead, due to a configura-
178 tion error, the sea ice distribution is held fixed to its August climatological pattern in
179 both the reference fine-resolution and coarse-resolution simulations. Ideally the sea ice
180 would be consistent with the annual cycle, but since this error occurs in both our ref-
181 erence and coarse-resolution simulations, it should not have an impact on our conclu-
182 sions regarding the ability of the ML to make a coarse-resolution simulation evolve more
183 like a fine-resolution one.

184 In the control climate, the climatological biases in precipitable water and precip-
185 itation are substantially reduced with a ∼25 km grid vs. a ∼200 km grid. Figure 1 shows
186 maps of these biases in annual-mean precipitable water and precipitation compared to
187 1982 – 2012 averages for ERA5 reanalysis (Hersbach et al., 2019) and Global Precipi-
188 tation Climatology Project (GPCP) (Adler et al., 2003) observations, the same years used
189 to form the SST climatology used in our simulations. In both simulations, the spatial
190 patterns of the precipitable water and precipitation biases are highly correlated, reflect-
191 ing the strong observed relationship between the two fields (Bretherton et al., 2004). The
192 finer grid results in smaller biases in mountainous terrain such as the Andes and Himalayas,
193 as well as improved simulation of tropical rain belts, e.g., over northwest South Amer-
194 ica and central Africa. Overall, by increasing the resolution, the global root mean square
195 error (RMSE) in time-mean precipitable water is reduced by 48 % and that of precip-
196 itation is reduced by 30 %. This motivates using the 25 km simulation as a reference across
197 the control and perturbed climates.

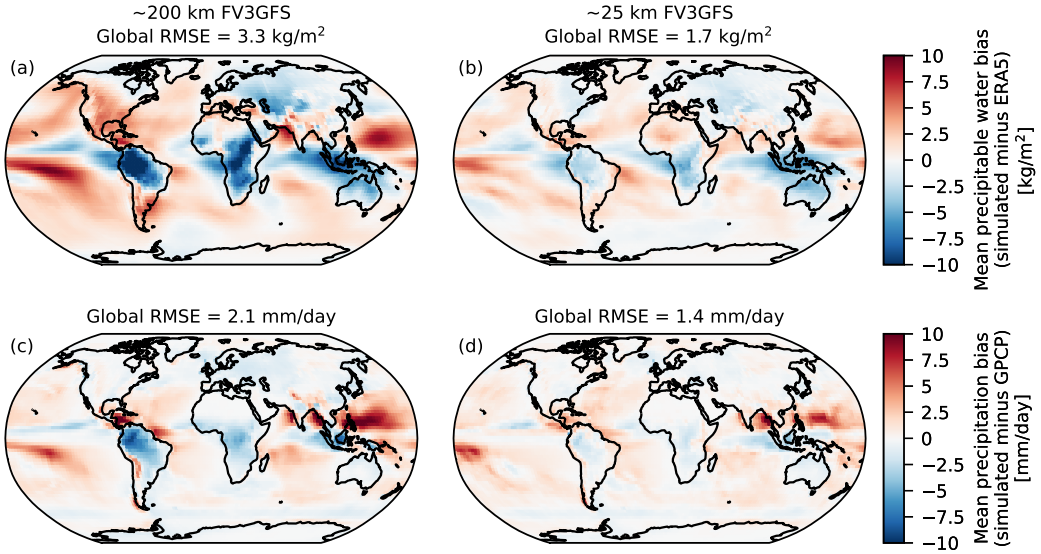

Figure 1. Top row: Time-mean precipitable water bias compared to ERA5 reanalysis for the (a) ∼200 km baseline and (b) ∼25 km reference simulations in the control climate. Bottom row: As in top row, but time-mean precipitation bias compared to GPCP observations. The time means are taken over the five post-spinup years of the baseline simulation, the second year of the reference simulation, and years 1982 – 2012 of the ERA5 reanalysis or GPCP observations.

2.1.2 Perturbed climate reference simulations

For the perturbed-climate simulations, a uniform offset is added to the specified climatogical SST. We assume the prescribed climatological distribution of sea ice, defined as a fraction of area covered by sea ice in each grid cell, remains the same across all climates, a commonly-made but unrealistic simplification. An additional simplification we make is that we do not perturb the carbon dioxide concentration; instead it is prescribed to its present-day value in each simulation.

To efficiently spin up the land surface and atmosphere to the perturbed-climate SSTs, we initialize all C384 runs—including the control climate case for consistency—using restart files from the end of year-long C48 simulations with the same SST perturbations (the "spin-up" simulation listed in Table 1). We upsample the restart files from C48 to C384 resolution using the `chgres_cube` tool developed at the NOAA Environmental Modeling Center (EMC), included in the `UFS_UTILS` GitHub repository (Gayno et al., 2020). The C48 spin-up simulations are started from initial conditions derived from the Global Forecasting System analysis (NCEI, 2020) for the date 2016-08-01 at 00Z, with SSTs per-

turbed uniformly depending on the climate. The coarse-grid spin-up and fine-grid reference simulations are run on NOAA's Gaea supercomputer using the pure Fortran version of FV3GFS maintained by our group linked to earlier.

2.2 Baseline coarse-resolution simulations

For comparison, we run 5.25 year baseline simulations with FV3GFS at C48 resolution in each climate and discard the first three months as a pre-analysis spinup period. Each simulation is initialized using a coarsened set of restart files from the end of the first year of the fine resolution reference simulations and uses the same sea ice and climate-specific SST climatology. These no-ML baseline simulations provide a skill benchmark for evaluating our ML-corrected simulations. The baseline and subsequently discussed nudged and ML-corrected simulations are run with cloud computing resources within a Python-wrapped version (McGibbon et al., 2021) of the pure Fortran version of FV3GFS.

2.3 Generating a training dataset

To derive a training and testing dataset of corrective tendencies for the coarse model's temperature and specific humidity, we extend the nudging approach described in B22. We run two-year C48 simulations with FV3GFS in which we nudge the temperature, specific humidity, zonal wind, meridional wind, and pressure thickness to the coarsened state of the C384 reference runs, in each climate (the "nudged" simulations in Table 1). As in B22, "nudging" is defined as the relaxation of a prognostic field in the model, a^n, to its coarsened value in a reference fine-grid dataset, $\overline{a}$, with a uniform timescale, τ, here chosen to be 3 h. This involves adding a tendency of the form

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

to the governing equations of the nudged variables in the model, constraining the nudged coarse model fields to approximately track the reference.

Ideally, this nudging approach smoothly changes the atmospheric state such that the tendencies due to physical parameterizations and dynamics respond smoothly on timescales much longer than the nudging timescale. However, in practice, this is often not the case, especially in the atmospheric boundary layer or around parameterized moist convection. That can lead to undesirable sensitivity of the nudging tendencies and the division of

work between parameterized physics and nudging to the somewhat arbitrarily chosen nudging timescale (Kruse et al., 2022, submitted to *JAMES*).

Also following B22, we prescribe the downward shortwave, net shortwave, and downward longwave radiative fluxes and precipitation rate seen by the land surface model from the coarsened fine resolution reference, as these have significant time-mean biases in our nudged coarse runs, and otherwise feed back to alter the temperature and specific humidity nudging tendencies.

In the un-nudged baseline coarse-grid simulations, the net surface radiative flux into the land surface in the coarse model, defined as:

$$R_{sfc}^{net} = S_{sfc}^{down} + L_{sfc}^{down} - S_{sfc}^{up} - L_{sfc}^{up} \tag{2}$$

has a mean bias between $-10\,\mathrm{W\,m^{-2}}$ and $-18\,\mathrm{W\,m^{-2}}$, depending on the climate. Here S_{sfc} is the downward or upward shortwave component of the radiative flux at the surface and L_{sfc} is the downward or upward longwave component of the radiative flux at the surface. This bias is primarily due to too much cloud and too little downward shortwave radiative flux at the surface compared to the fine-grid reference. It has the opposite sign to that found by B22, mainly due to our aforementioned inclusion of microphysical adjustment in the dynamical core remapping step, which increases simulated cloud formation over land. The coarse-model bias in downwelling surface radiative flux is a good target to correct via machine learning because it induces climatically important biases in the land surface skin temperature, hereafter referred to as "surface temperature," and latent heat flux.

As in B22, our machine learning targets from these simulations are the column-wise 79-level vertical profiles of nudging tendencies time-averaged over 3 h intervals, with time labels at the interval centers, and the instantaneous downwelling surface radiative fluxes. In addition to outputting the targets as diagnostics, we output the features used by our ML. These are the instantaneous profiles of model temperature and specific humidity at the time the nudging tendencies are defined, as well as some scalar quantities, which are the surface geopotential (which can act in part as a continuous-valued proxy for discriminating land from ocean and sea ice), the cosine of the solar zenith angle (computed from the time of day, longitude, and latitude following Monteiro et al. (2018)), the land surface type (an integer-valued field which is zero in ocean grid cells, one in land grid cells, and two in sea ice grid cells), and the surface albedo.

2.4 Predicting the nudging tendencies

Using the profiles of temperature and specific humidity, surface geopotential, and cosine of the solar zenith angle as inputs, we learn the column temperature and specific humidity nudging tendencies. B22 found that using ML correction of winds produces large mean state drifts in upper atmospheric temperature, so we choose not to do that here. In addition, B22 compared the use of a random forest or an ensemble of neural networks to predict the nudging tendencies, finding comparably skillful results. We choose to focus on using neural networks, because they require less memory to store and are computationally efficient in a variety of hardware settings, including on GPUs (Yuval et al., 2021). In addition, the random seed used in their training—a parameter used in setting the initial weights of the model, and the order of the shuffling of the samples in a training batch—introduces some variability in online performance for similar offline skill, allowing selection of an ML model to minimize climate bias.

2.5 Predicting the surface radiative fluxes

We make ML-based predictions for the radiative flux inputs to the land surface model. These inputs are the downward shortwave, net shortwave, and downward longwave radiative fluxes at the surface. For predicting the shortwave fluxes, B22 used the cosine of solar zenith angle as a proxy feature for top-of-atmosphere downward solar flux. This does not account for the 7% variation of insolation with time of year due to the eccentricity of the Earth's orbit. That had negligible impact during the 40 d simulations of B22, but is relevant in our simulations which span the full annual cycle.

Thus we use a slightly different ML approach for shortwave radiative fluxes than in B22. It is based on the shortwave transmissivity of the atmospheric column, $\mathcal{T}$, defined as the ratio of the downward shortwave radiative flux incident on the surface (S_{sfc}^{down}) to the downward shortwave radiative flux at the top of the atmosphere (S_{toa}^{down}):

$$\mathcal{T} = \frac{S_{sfc}^{down}}{S_{toa}^{down}}. \tag{3}$$

If we train the ML model to predict $\mathcal{T}$, we can then compute the downward and net (S_{sfc}^{net}) shortwave radiative fluxes at the surface using FV3GFS's values for the downward shortwave radiative flux at the top of the atmosphere and the surface albedo (α):

$$S_{sfc}^{down} = \mathcal{T} S_{toa}^{down} \tag{4}$$

$$S_{sfc}^{net} = (1 - \alpha)\, S_{sfc}^{down}. \tag{5}$$

Explicitly computing the net shortwave radiative flux at the surface using the coarse model's surface albedo provides a less biased prediction than forcing the ML to learn this relationship, particularly over high-albedo regions like the Sahara and Arabian deserts or polar ice-covered regions.

To predict the shortwave transmissivity and downward longwave radiative flux at the surface, we use a random forest with the column temperature, column specific humidity, surface geopotential, surface type (ocean, land, or sea ice), cosine of the solar zenith angle, and surface albedo as input features. When predicting the full values for all the surface radiative flux inputs to the land surface model, B22 demonstrated that a random forest (RF) and a neural network (NN) with outputs appropriately rectified to be greater than or equal to zero, performed comparably in terms of offline skill. We use a random forest because it automatically constrains the predicted transmissivity to be between 0 and 1; with an appropriate activation function this constraint could also be applied to a neural network.

2.6 ML training

When training the neural networks and random forests, we use data from the first year of the nudged simulations in all climates. We follow a similar time-sampling approach to that of W21, who also trained models across the annual cycle. We randomly select 160 of the 2920 available times to sample both the annual and diurnal cycles to enable efficient training (early tests indicated that training on more data did not make a material difference when models were used online). These times are then separated into 16 batches of 10 each. Within each batch, data from each of the times is loaded from each of the climates, forming a two-dimensional array with "sample" and "feature" dimensions. Since the machine learning problems are column-based, the sample dimension has a length corresponding with the total number of columns in the batch: (6 x 48 x 48 = 13 824 columns per time) x (10 times per batch) x (4 climates) = 552 960 columns, while the length of the "feature" dimension depends on the inputs we are using for the model. This array is then randomly shuffled along the "sample" dimension. Since we train on a sequence of 16 batches, in total our models are trained on 16 x 552 960 = 8 847 360 samples.

To train neural networks for the temperature and moisture nudging tendencies, the gradient is updated every 512 samples within each batch, and the full set of batches is repeatedly iterated over in 24 training epochs. We use the same implementation in `keras` (Chollet et al., 2015), and the same hyperparameters for the temperature and specific humidity nudging tendency network as in B22, i.e. a mean absolute error loss function, two hidden layers with a width of 128, a learning rate of 2×10^{-3}, and an L2 regularization penalty of 1×10^{-4}.

Inputs and outputs of the neural networks are normalized or de-normalized following similar procedures to those in B22. Specifically, we normalize a scalar input or output $x \in \mathbb{R}$ (e.g. temperature at a single level, cosine of the solar zenith angle, etc.) with $(x - \bar{x})(\bar{\sigma}_x + 10^{-7})^{-1}$, where $\bar{x}$ and $\bar{\sigma}_x$ are the sample mean and standard deviation. The ML then predicts a normalized value $\tilde{y} \in \mathbb{R}$, and $y := \tilde{y}\bar{\sigma}_y + \bar{y}$ is the ML prediction in physical units. These may seem like standard methods for working with neural networks, but there are many small differences in this recipe across the ML parameterization literature, which, in our experience, can alter both offline and online performance.

For reproducibility, the random seed for all elements of randomness during the training process is a parameter in our training workflow. We train neural networks with four random seeds, labeled 0-3. These neural networks have similar offline skill, but produce different outcomes when applied online. This phenomenon was illustrated in a more extreme way in Wang et al. (2021), where they trained 50 ML models with comparable offline skill, but found only a small subset that could support stable long-term simulations.

To train a random forest model to predict the shortwave transmissivity and downward longwave radiative flux at the surface, like B22, we use the `scikit-learn` (Pedregosa et al., 2011) implementation with a mean square error loss function and a maximum depth of 13. The ensemble consists of 16 trees where each tree is trained on a batch of 10 timesteps. Like in W21, no transformations are applied to the inputs of the RF, but similar to the case of the NNs in this study, the ML predicts a normalized value $\tilde{y} \in \mathbb{R}$ and the predictions are de-normalized to be placed in physical units, in the case of the RF using $y := \tilde{y}\left(\bar{\sigma}_y + 10^{-12}\right) + \bar{y}$. While there is an element of randomness to training an RF, in previous work we have found empirically that this does not have a significant impact on offline or online results.

For offline testing, for computational efficiency, we randomly select 90 times from the second year of the nudged simulations, and combine all the columns associated with those times into a testing dataset. These provide a set of samples that we can test our models against that is independent from the data the models were trained with. We compute offline skill both aggregated across all climates and separated into different climates to evaluate each model's overall skill and to ensure that the models are indeed skillful in each of the climates we train on and not subtly optimizing for a specific climate.

2.7 Input ablation and output tapering of vertically resolved fields

For handling model inputs and outputs of the nudging tendency NN, we initially followed B22. For every vertically resolved input, like temperature, we provided its values at all 79 vertical levels in the column, and for every vertically resolved output, like the temperature nudging tendency, we predicted its full target value at each vertical level. Such models worked reasonably well in 40 d simulations, but were prone to cause online drift and/or crashes in simulations longer than a few months, due to problematic behavior of the ML in the uppermost 25 model levels.

As an example, the left panels in Figure 2 illustrate the time series of temperature and ML-predicted heating rate at a representative column in a five-year simulation using an ML configuration similar to that used in B22. A high-amplitude wave-like pattern in temperature develops in the upper model levels for the first year of the run, driven by ML-predicted heating. As the temperature near the tropopause starts to drift cold, this signal disappears. However, once the temperature sinks below the training range, indicated by the purple regions in Figure 2c, the ML-predicted heating rate spikes in magnitude, leading to further temperature drift, even near the surface and in the mid-troposphere.

Past authors have encountered similar problems when including upper-atmospheric inputs in column-based machine-learning parameterizations. Coarsening in space and time creates simultaneous correlations between inputs (e. g. high upper-tropospheric humidity) and outputs (e.g. strong mid-tropospheric latent heating) that the ML unphysically encodes into causal predictions. Brenowitz and Bretherton (2019) showed that the ML-predicted precipitation was spuriously sensitive to stratospheric moisture offline. They stabilized online runs by excluding (ablating) that input. Brenowitz et al. (2020) further found analytically that upper atmospheric temperature and moisture inputs can de-

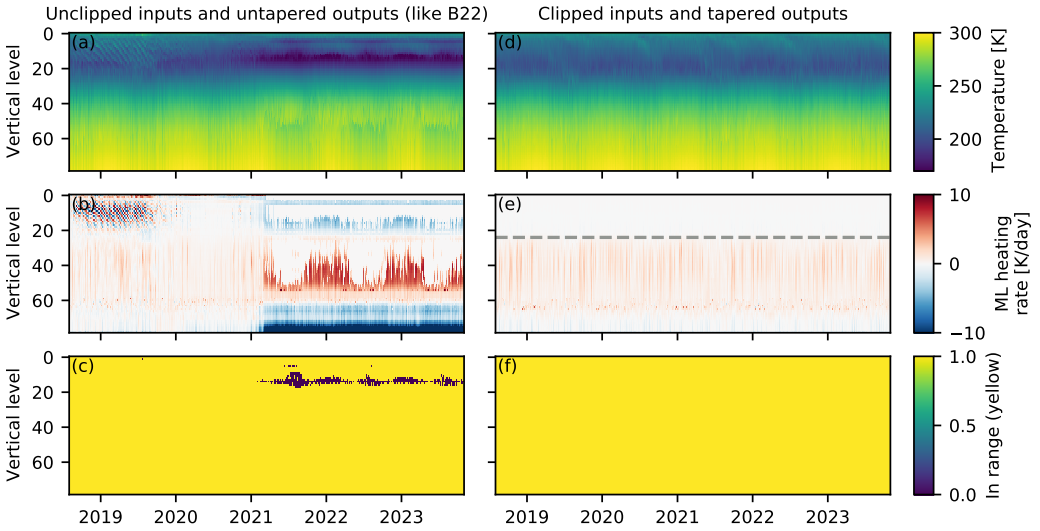

Figure 2. Time series of temperature, ML-predicted heating rate, and flag denoting whether the temperature is inside or outside the range of temperatures in the training data, at a single column in a control-climate ML-corrected simulation without input ablation or output tapering (left) and a control-climate ML-corrected simulation with input ablation and output tapering (right). The grey dashed line in panel (e) indicates the 25th level from the top, the level at which ablation and tapering begins.

velop unstable feedbacks with gravity-wave modes. Ablation is also used implicitly—even if not emphasized—by other related works, e.g., O'Gorman and Dwyer (2018); Yuval et al. (2021). Together with empirical experimentation, this motivated us to ablate the inputs from the uppermost top 25 model levels.

We also see large predicted corrective temperature tendencies in the uppermost atmospheric levels, illustrated by Figure 2b. Parameterized physical processes provide only weak thermal damping in this region. Thus these corrective tendencies may derive from the training data, but they get amplified and distorted by dynamical feedbacks, doing more harm than good. A natural solution is to reduce the magnitude of the predicted tendencies our ML models, so the weak damping provided by the model physics adequately stabilizes the system. We do this by multiplying the target corrective tendencies by a tapering factor that exponentially decreases from 1 down to near zero in the top 25 model levels:

$$f(k) = \begin{cases} e^{\frac{k-25}{5}} & k < 25 \\ 1 & k \geq 25, \end{cases} \tag{6}$$

where k is the integer-valued model level index, following FV3GFS's internal convention that $k = 0$ corresponds to the level closest to the model top, and $k = 78$ corresponds to the level closest to the surface. This tapering factor decreases by a multiple of e every five levels above level 25, reducing to $e^{-5} \approx 0.007$ in the uppermost level. (Yuval & O'Gorman, 2020) did something similar in that they omitted using ML to predict the radiative heating rate in vertical levels above 11.8 km.

The combination of ablating inputs and tapering outputs in this fashion results in ML models that reliably lead to stable and non-drifting ML-corrected simulations (for comparison see the column time series plots in Figure 2d-f). However, a more careful ablation study would be useful to determine whether input ablation and output tapering are both necessary, or whether doing just one or the other could have a similar effect.

2.8 ML-corrected online simulations

While we test our machine learning models offline using independent test data, the most important test comes in using them to correct the temperature and specific humidity tendencies and surface radiative fluxes during each timestep in free-running FV3GFS simulations. To do this, we run a suite of simulations in each of the four climates using

four ML configurations, one for each of the neural networks trained with the four random seeds, keeping the surface radiative flux model the same across all configurations. This is a total of 16 ML-corrected simulations. In these runs, as in B22, the ML predictions of the tendency corrections and radiative flux overrides are integrated into the time loop of the model using a Python-wrapped version of FV3GFS (McGibbon et al., 2021) that we run (along with the full ML workflow) on Google Cloud. To assess the configurations' performance before running longer simulations, we run each ML-corrected simulation for 1.25 years, and extend the simulations of the best-performing configuration to 5.25 years to generate five full post-spinup years of statistics. This is an analogous approach to that of Wang et al. (2021), though we tried far fewer candidate configurations.

2.9 Evaluation of skill

To determine how the ML corrections impact the quality of coarse-grid simulations, we compute error metrics for the climate statistics of the ML-corrected runs using the fine-grid runs as a reference, and compare these to the same error metrics computed using the baseline runs. To allow the baseline and ML-corrected coarse-grid simulations to sufficiently diverge from their initial conditions, which are derived from the fine-grid reference simulations, we begin our analysis after a three month spin-up period. Starting in month four, we partition each coarse simulation into as many complete non-overlapping twelve-month periods as possible. Each such period serves as an approximately independent sample year of coarse-model climate statistics; initial ML-corrected runs therefore have one year of climatological data, while baseline runs have five. Regardless of the year in the coarse runs, error metrics are always computed relative to the second year of the corresponding fine-resolution run in each climate. This is appropriate since the sea ice and SST lower boundary conditions for the fine and coarse runs follow the same repeating annual cycle for all years. Qualitatively our results are not sensitive to this choice. We have computed the error metrics with respect to the first years of the fine-resolution simulations in each climate and find them to be similar to those we report here.

We focus on a limited set of societally relevant and climatically important metrics that we hope will be improved by the corrective ML:

1. The root mean square error (RMSE) of the time mean spatial pattern of precipitation.

2. The time and spatial mean bias of the precipitation rate over land.
3. The RMSE of the diurnal cycle of precipitation over land with the mean bias removed.
4. The RMSE of the time mean spatial pattern of the surface temperature over land.
5. The time and spatial mean bias of the net radiative flux into the land surface.

Precipitation is affected by ML-predicted atmosphere drying. Surface temperature is affected by ML surface radiative flux predictions and near-surface temperature tendency corrections. The net radiative flux into the land surface depends on ML-predicted values for the net shortwave and downward longwave radiative fluxes at the surface. We will also document the vertical structure of zonal mean biases of temperature, specific humidity, and the mass streamfunction.

3 Results

3.1 Biases in the nudged simulations

The primary goal of the machine learning is to bring the weather variability and the resulting climate statistics of coarse resolution simulations closer to those of fine resolution runs. Accordingly, our "truth" dataset – i.e. the dataset that we will compute our biases against – consists of the second year of output of the ∼25 km simulations in each of the climates, coarsened to ∼200 km resolution. The ML can only be as good for this purpose as its training methodology, which is based on the nudging tendencies diagnosed from the nudged runs. As B22 noted, that methodology is a compromise between keeping the coarse model state as close as possible to the fine-grid reference state, while also evolving smoothly in a dynamically balanced way with a minimum of small-scale vertical velocity transients. While most aspects of the nudged simulations, such as temperature and humidity fields, remain close to the coarsened fine-grid reference data on which it is based, there are important aspects of the nudged simulations, notably time-mean precipitation, that prove more sensitive to this methodology. That is, the nudged training dataset does not have the same statistics as does the reference, potentially building biases into the ML training even if the ML itself were perfect.

With this in mind, in this section we will present some biases of the nudged and baseline runs related to the metrics described in Section 2.9 for comparison with results of the ML-corrected runs presented later. For each metric, we are hoping that the nudged

run bias is much smaller that the baseline run bias, so that the ML has a chance to correct most of the baseline bias despite possible shortcomings of the training approach.

3.1.1 Precipitation

Precipitation in the nudged and ML-corrected simulations is computed as a budget-implied precipitation rate. This is a concept discussed in W21 and B22 and is an estimate of the precipitation rate that takes into account contributions from the model physics as well as the specific humidity nudging or ML-predicted tendency in the column. In the context of nudged runs it is computed following

$$P_{nudged} = P^p - \langle \Delta Q_q \rangle \,, \tag{7}$$

where P^p is the precipitation rate predicted by the model physics, and ΔQ_q is the nudging tendency of specific humidity, with the angle brackets denoting a mass-weighted vertical integral. In ML-corrected runs we take the additional step of rectifying this quantity such that it is always greater than or equal to zero:

$$P_{ML-corrected} = \max\left(0, P^p - \langle \Delta Q_q \rangle\right). \tag{8}$$

We do this differently in the nudged and ML-corrected simulations because high-frequency fluctuations in the nudging tendencies can rectify into a large high bias in implied precipitation. In addition we do not need this precipitation estimate to be positive in the nudged run, in which it is not used to force the land surface model. The rectification bias is unavoidable but much less important in ML-corrected prognostic runs (less than $0.05\,\mathrm{mm\,d^{-1}}$ in all simulated climates) since the ML correction is less prone to such fluctuations.

Figure 3a shows a time-mean map of precipitation biases in the nudged run with respect to the fine resolution run. They are reassuringly small over most of the oceans. A dipole pattern in the vicinity of the Intertropical Convergence Zone (ITCZ) over the Eastern Pacific suggests a slight southward shift of the ITCZ in the nudged run compared to the fine resolution run, while over the Atlantic a tripole pattern is present suggesting a slight widening of the ITCZ. There are larger grid-scale biases over land, with regional dry biases over sub-Saharan Africa and the Rocky Mountains. These land biases contribute to a land root mean square error (RMSE) of $1.4\,\mathrm{mm\,d^{-1}}$ in the control climate.

Bias patterns are similar in the other climates, increasing in magnitude and grid-scale noisiness with increased SST (the RMSE metrics for the nudged runs in each of the climates are plotted for reference in Figure 9c and d as blue dots). The mean precipitation rate over land has only a slight negative bias in all climates, with values around $-0.2\,\mathrm{mm\,d^{-1}}$, much smaller than that for the baseline model, which has values around $-0.8\,\mathrm{mm\,d^{-1}}$, mainly due to dry biases over tropical South America and Africa (Figure 3b). This suggests that our specific humidity nudging and radiative flux prescription has the desired effect of creating a training dataset with biases versus the reference simulation that are much smaller than those of the baseline model.

The mean diurnal cycle of precipitation over land regions between 60°S and 60°N as a function of local solar time is plotted in Figure 3c. The latitudinal limits are imposed to make a fair comparison with the spatial extent of available observations, which are derived from year 2016 of the Integrated Multi-SatellitE Retrievals for GPM (IMERG) (Huffman et al., 2019). The black curve shows the control-climate ∼25 km reference, which peaks around 14:30 local solar time, about two hours earlier and with a slightly lower amplitude than the IMERG observations, the dashed black curve. The orange curve shows the baseline run, which has a peak at a similar time to the ∼25 km run, but too low an amplitude, a common problem in coarse resolution climate models (Christopoulos & Schneider, 2021). In the nudged run, the amplitude of the afternoon peak is improved, but the budget-inferred precipitation rate decreases too sharply in the evening and is too large in the late morning hours; in a qualitative sense, however, this is more in line with the behavior of the fine-resolution reference simulation than the baseline. This bias is qualitatively similar in analogous nudged runs in the other climates, and will be discussed further in Section 3.6.1.

Overall, this analysis suggests that ML that seeks to learn the nudging tendencies and surface radiative fluxes has potential to make improvements to the precipitation climatology.

3.1.2 Surface temperature

Nudging greatly reduces surface temperature bias over land. The time-mean surface temperature bias in the second year of the control climate nudged run is shown in Figure 3d. Since the SSTs are prescribed, the bias in surface temperature over ocean is

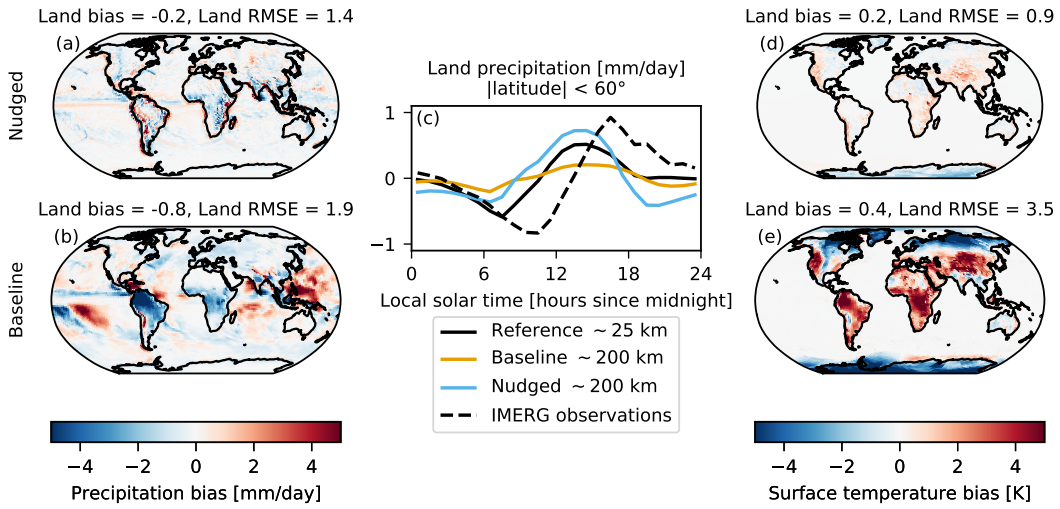

Figure 3. Time-mean precipitation bias in the control climate nudged (a) and baseline (b) simulations, diurnal cycle of precipitation over land with the mean removed in the reference, nudged, and baseline simulations, as well as IMERG observations (c), and time-mean surface temperature bias in the nudged (d) and baseline (e) simulations. Biases are computed as coarse-grid run statistics minus fine-grid run statistics.

trivially zero. Over land, the surface temperature is influenced by the net radiative flux into the surface, which is largely prescribed in our nudging procedure, but also depends on the partitioning between latent and sensible heat fluxes by the land surface model, which can differ between the nudged coarse and reference fine simulations. The biases are generally much smaller than those of the baseline simulation (Figure 3e), which has predominantly warm biases in the tropics and mid-latitudes and cold biases in the polar regions. The spatial pattern and amplitude of the surface temperature biases in the nudged and baseline runs are similar across climates. As with land precipitation, this suggests that if corrective ML can retain the bias reduction in the nudged training data, ML correction could reduce the land surface temperature biases of the baseline run.

3.2 Nudging tendencies

Despite using reference simulations with different configurations, both in terms of spatial resolution and some physical parameterizations, and different time periods, the time-mean nudging tendencies that emerge from the nudged simulations are similar to those shown in B22. Figure 4a and Figure 4c show the mean column-integrated heat-

ing, $\langle \Delta Q_T \rangle$, and moistening, $\langle \Delta Q_q \rangle$ over the test dataset in the control climate. In these spatial plots, as in B22, we can see that the nudging tendencies are largely associated with making up for missing precipitation and latent heating in the nudged coarse simulation; the column-integrated temperature nudging tendency is generally positive, and largest in regions of greatest column-integrated drying. The panels in the right column of Figure 4 show the global-mean vertical profile of the nudging tendencies in each climate.

The magnitudes of the nudging tendencies increase with warmer SSTs. In a column-integrated sense, for both temperature and specific humidity, this increase is approximately at a rate of $3\,\%\,\mathrm{K}^{-1}$ to $5\,\%\,\mathrm{K}^{-1}$ increase in SST, somewhat less than the rate of increase of the column-integrated parameterized temperature and specific humidity physics tendencies (5 % to 6 %), or the $\sim 7\,\%\,\mathrm{K}^{-1}$ Clausius-Clapeyron scaling for water vapor with warming (Held & Soden, 2006). The spatial patterns of the column-integrated nudging tendencies do not differ significantly with climate (not shown). While we do not predict them in this work, for reference the mean horizontal wind nudging tendencies are plotted in Figure S1, which have a similar spatial pattern to those in B22, but a slightly weaker magnitude.

3.3 Offline skill in predicting the nudging tendencies

In individual samples, the nudging tendencies are noisy. Figures 5a and c show the target temperature and specific humidity tendencies for a representative evening in August in the control climate of the test dataset for a vertical cross section along 0°E. These tendencies and predictions are illustrative of their character in other climates and at other times. The nudging tendencies are typically largest near the top of the boundary layer, and in regions of deep convection. The seed 2 neural network makes a prediction that is smoother than the targets for both ΔQ_T (Figure 5b) and ΔQ_q (Figure 5d). The other NNs make qualitatively similar predictions. Because of the noisiness of the target tendencies, it is difficult for the neural networks to capture all of their variance.

Figure 6 shows the coefficient of determination (R^2) for the temperature and specific humidity nudging tendencies computed offline across the 90 times of the test dataset in all climates binned by latitude and pressure. For the temperature nudging tendency, skill is highest in the tropical boundary layer and upper troposphere, where values reach

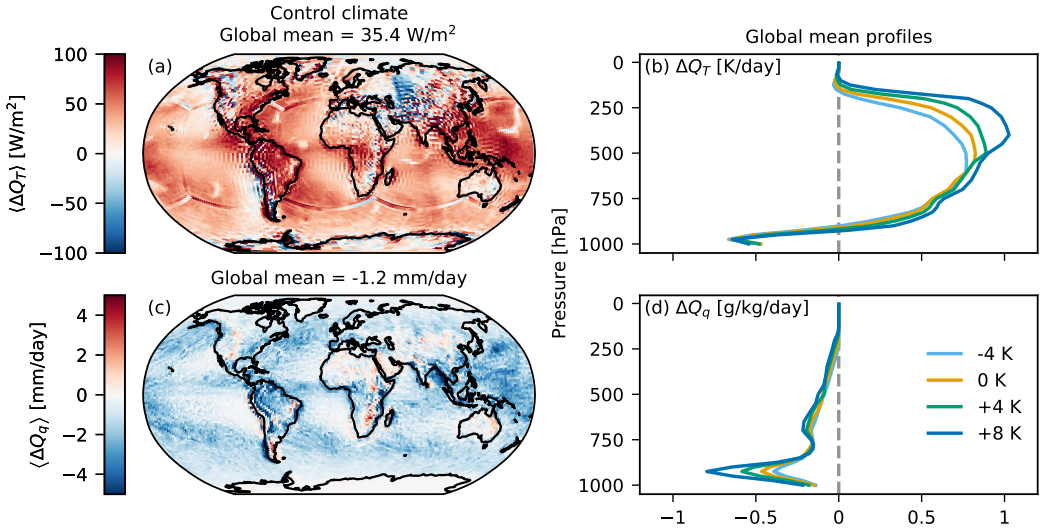

Figure 4. Column-integrated temperature (a) and specific humidity (c) nudging tendencies in the control climate, averaged over the test dataset, and global mean vertical profiles of the temperature (b) and specific humidity (d) nudging tendencies averaged over the test data in each climate. In each case the tapering of the tendencies in the upper 25 model levels described in Section 2.7 has been applied. The x-axis scale is the same for panels (b) and (d) despite representing different units.

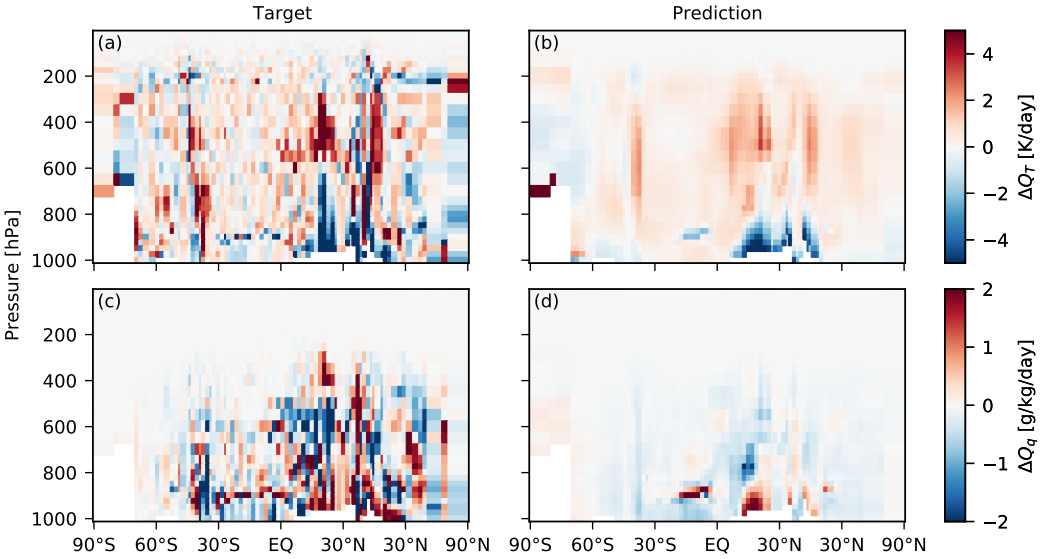

Figure 5. Samples of the target and offline-predicted nudging tendencies at 2018-08-07 20:30:00 along 0°E in the control climate. (a) and (c) are the target temperature and specific humidity tendencies, respectively, and (b) and (d) are the predicted temperature and specific humidity tendencies using the neural network trained with seed 2, respectively. For plotting purposes only, all fields are interpolated to surfaces of constant pressure after being computed.

0.2 – 0.3, and decreases as one moves poleward. For the specific humidity nudging tendency, skill is most concentrated in the tropical boundary layer where similar to the skill for the temperature nudging tendencies, R^2 maximizes around 0.25. If one were to make a plot aggregating data over all atmospheric columns instead of binning by latitude, the result would look similar to that of the "TquvR-NN" curve in Figures 5a and b of B22, but would be slightly smoother in the vertical and generally have lower values, here peaking around 0.2 while in B22 values peak around 0.3. In Figure 6 the skill is aggregated across all climates, but if one were to look at the skill in any one climate, it would look qualitatively similar, though skill in predicting either the temperature or specific humidity nudging tendency in the upper troposphere tends to be higher in the cooler climates.

3.4 Offline skill in predicting the radiative fluxes

The random forest trained to predict the surface radiative fluxes is quite accurate when evaluated offline. When evaluated globally at each of the 90 times in the test dataset, depending on the climate, the root mean square error of the time-mean pattern glob-

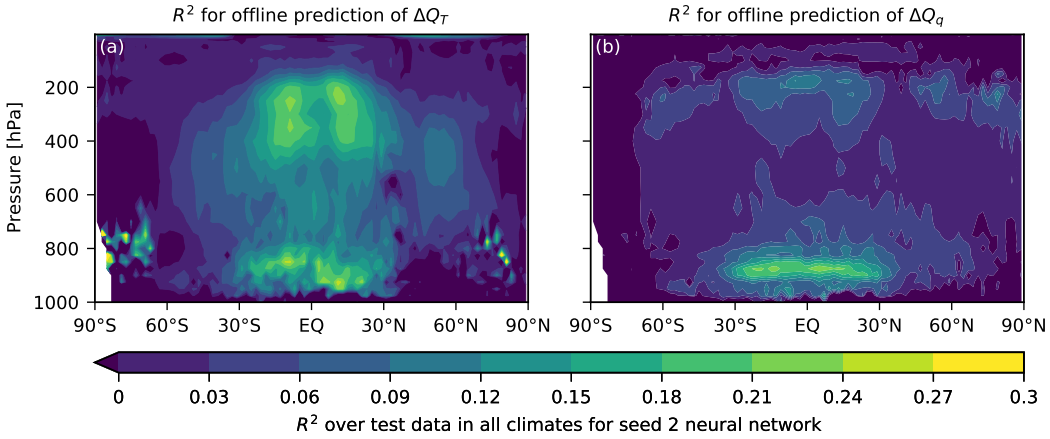

Figure 6. Coefficient of determination as a function of latitude and pressure for the offline prediction of the temperature (a) and specific humidity (b) nudging tendencies across the 90 times of the test dataset in all four climates. The values here are for the neural network trained with seed 2, but the plots look qualitatively similar with neural networks trained with other seeds.

ally is between $4\,\mathrm{W\,m^{-2}}$ to $5\,\mathrm{W\,m^{-2}}$ for the downward longwave radiative flux, $10\,\mathrm{W\,m^{-2}}$ to $12\,\mathrm{W\,m^{-2}}$ for the downward shortwave radiative flux, and $9\,\mathrm{W\,m^{-2}}$ to $11\,\mathrm{W\,m^{-2}}$ for the net shortwave radiative flux. For the control climate, the statistics broken down into land and ocean/sea-ice regions can be found in the panel titles of Figure 7.

Figure 7 shows the time mean spatial pattern of the offline prediction bias for each surface radiative flux component in the control climate. In the spatial mean, these are reassuringly small. Because downwelling clear-sky radiation is a smooth function of atmospheric temperature and humidity profiles (and solar zenith angle for shortwave radiation) we interpret these biases as due to the RF not fully learning the radiative effects of clouds in the fine-grid reference run. More cloud leads to less daytime downwelling shortwave and somewhat more downwelling longwave radiation. This bias is prominent over stratocumulus regions in the subtropical oceans in Figure 7b and 7c and (to a lesser extent) Figure 7a. Since SST is specified, surface radiative biases over ocean regions do not feed back on our simulations, so this is not an immediate concern. Similar weaker but broad-scale biases are seen over the Southern Ocean and (more importantly) a land region, Siberia. These suggest the fine-grid reference supports more cloud in these regions than radiatively accounted for by the ML scheme. The reverse bias, only weaker,

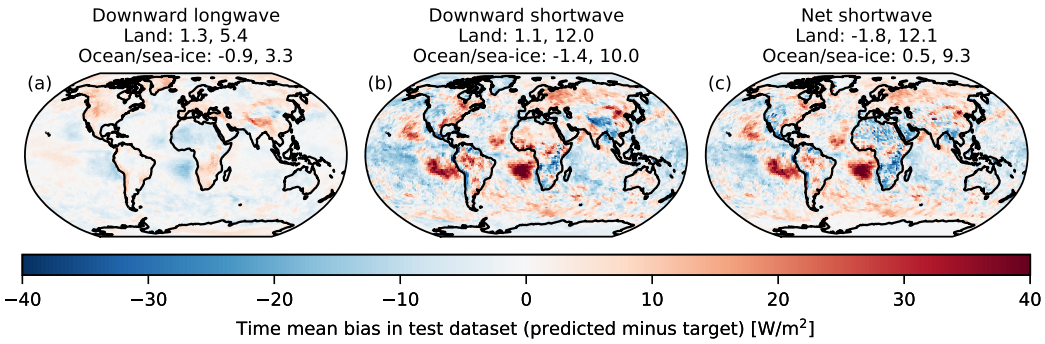

Figure 7. Time mean bias in the offline prediction of the downward longwave radiative flux at the surface (a), downward shortwave radiative flux at the surface (b), and net shortwave radiation flux at the surface (c) in the control climate in the test dataset. Comma-separated spatial mean bias and spatial RMSE statistics for the time-mean pattern over land and ocean/sea-ice are reported in the panel titles with units of $\mathrm{W\,m^{-2}}$.

is seen in the subtropical oceanic shallow cumulus regimes. We interpret this as the ML overpredicting cloud-induced surface shortwave and longwave radiative effects. The strong radiative biases over the Himalayas may also involve the RF inadequately accounting for the effects of extreme surface elevation on clear-sky surface radiation.

Through the surface albedo, downward shortwave (Figure 7b) and net shortwave radiation (Figure 7c) are directly correlated (Equation 5). In most regions, the learned biases in net shortwave radiation correlate tightly with biases in downward shortwave radiation, as physically expected. An exception is over the Sahara and Arabian deserts, where we underpredict time-mean net shortwave radiation. In these regions we generally see a mild positive bias in downward shortwave radiative flux, which suggests a slight mismatch between the surface albedo in the coarse and fine-resolution simulations in these regions. Such a mismatch might result from how we coarsen different properties of the land surface that factor into its surface albedo.

3.5 Results of initial ML-corrected simulations

The strongest test for the machine learning approach is to see whether it improves the simulation of climate when used online. As discussed in Section 2.8, we start by briefly analyzing the results of 1.25 year simulations in each climate using neural networks trained with four different random seeds. Figure 8 shows "swarmplots" (Waskom, 2021) of the

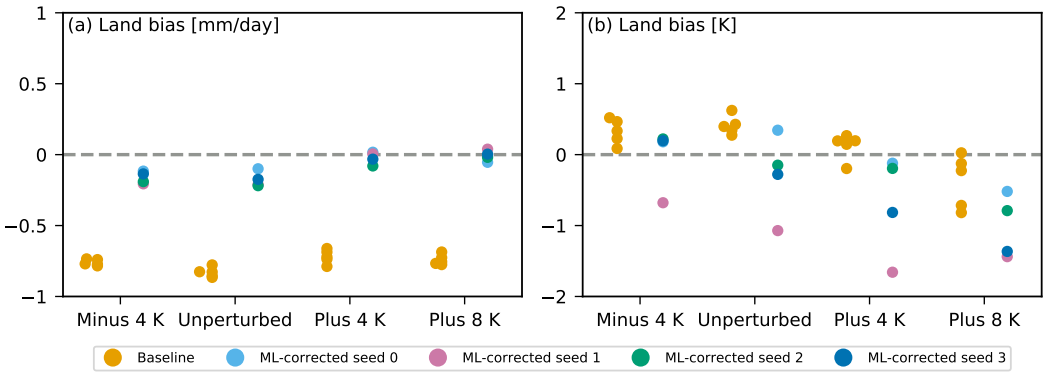

Figure 8. The annual mean bias in land precipitation (a) and land surface temperature (b) in individual post-spinup years of the baseline (yellow dots) and first post-spinup year of ML-corrected simulations with neural networks trained with different random seeds (colored dots) in each climate.

land-mean bias in precipitation rate and surface temperature in the five post-spinup years of each baseline simulation in climate compared to the same biases during each of the post-spinup years of the 1.25 year ML-corrected simulations with each random seed. With the ablation and tapering approach described in Section 2.7, all of the NNs led to stable non-drifting 1.25 year simulations in each climate, so no NNs are immediately disqualified from being selected for extended runs.

Figure 8a shows that land precipitation bias is generally not sensitive to the random seed used to train the neural network. All ML-corrected simulations in all climates exhibit an improvement over the baseline simulations, which all have a large negative land-mean precipitation bias.

Land surface temperature bias, shown in Figure 8b, on the other hand, is sensitive to the random seed of the neural network. While the baseline simulations generally have only a small net land surface temperature bias, some ML-corrected simulations, e.g. the seed 1 simulations in the +4 K and +8 K climates, exhibit large negative biases of over 1 K in magnitude. It is notable that in this case the ordering of the land bias by NN random seed tends to be similar across climates, suggesting that characteristics of the NNs when applied in one climate tend to be consistent with those characteristics when applied in another climate. The seed 1 NN leads to the most negative surface temperature biases, followed by seed 3, seed 2, and finally seed 0. The differences in surface temper-

ature biases between the different ML-corrected runs can largely by attributed to differences in the low-level heating rate predictions in the polar regions (not shown).

We do a five-year simulation using the seed that leads to the smallest biases. We start by eliminating seed 1, which leads to the largest negative temperature biases of all the seeds tried. Next, we eliminate seed 0, which while it appears does the best in reducing temperature errors both near the surface and higher in the atmosphere, does the worst in terms of specific humidity errors, leading to large positive biases in the tropics in all climates. This leaves seeds 2 and 3 which lead to similar results. Ultimately we focus on the results of 5.25 year simulations with seed 2, since it leads to slightly less biased surface temperatures than seed 3. Five-year simulations with seed 3 performed comparably well (not shown).

3.6 Results of multi-year ML-corrected simulations

In this section we will more comprehensively present the results of 5.25 year simulations completed with the seed 2 neural network. Table 2 summarizes our primary metrics for the baseline and seed-2 simulations. Ideally the corrective ML would improve these quantities without harming other aspects of the simulations; we now discuss them one by one.

3.6.1 Precipitation rate

Figure 9 illustrates the character of the annual-mean precipitation errors in the nudged, baseline, and ML-corrected coarse-resolution simulations. Recall that precipitation is computed following Equations 7 and 8 described earlier. The maps show the annual mean precipitation bias patterns in the control climate baseline and ML-corrected runs. These are averaged over the five post-spinup years of the runs. The swarmplots on the right treat individual years as individual samples, meaning that there are five datapoints per climate for the baseline and ML-corrected cases, and one datapoint per climate in the nudged run case. The precipitation rate in the baseline simulation is predicted purely by the model physics, P^p.

The precipitation bias pattern in the baseline run (Figure 9a) features large dry biases over land in the deep tropics, particularly in South America, a large wet bias over the Western Pacific Warm Pool, and an eastward shift in the South Pacific Convergence

Table 2. Summary of the mean metrics across individual years in baseline (no-ML) and ML-corrected simulations completed with seed 2. Mean metrics written in bold are considered to be robustly better for a particular configuration if values from each of the five years used to compute the mean are better than any of the years used to compute the mean in the other like-climate configuration. Percent difference ranges between the baseline and ML-corrected runs are reported only in cases where the difference in metrics is robust in all climates and has the same sign.

		Climate								
		−4 K		0 K		+4 K		+8 K		
Metric	Region	No-ML	ML	No-ML	ML	No-ML	ML	No-ML	ML	Range
P RMSE [mm d^{-1}]	Land	1.7	**1.2**	1.9	**1.4**	1.8	**1.4**	1.7	**1.5**	17 % to 30 %
	Ocean/sea-ice	1.2	1.2	1.6	1.6	**1.8**	2.0	2.3	2.2	-
P mean bias [mm d^{-1}]	Land	-0.8	**-0.2**	-0.8	**-0.2**	-0.7	**0.0**	-0.7	**0.0**	76 % to 100 %
	Ocean/sea-ice	0.1	-0.1	0.1	-0.1	**0.1**	-0.3	**0.2**	-0.4	-
T_s RMSE [K]	Land	3.6	**2.8**	3.5	**2.7**	3.6	**2.8**	3.4	**2.7**	20 % to 23 %
T_s mean bias [K]	Land	0.3	0.2	0.4	**-0.1**	0.1	-0.5	-0.4	-0.9	-
R_{sfc}^{net} RMSE [W m^{-2}]	Land	21.7	**15.3**	24.8	**14.4**	25.6	**13.1**	27.8	**13.6**	30 % to 51 %
	Ocean/sea-ice	**8.9**	9.6	9.6	10.2	**9.3**	10.5	10.8	10.4	-
R_{sfc}^{net} mean bias [W m^{-2}]	Land	-10.2	**-6.0**	-12.9	**-5.5**	-15.0	**-3.2**	-18.0	**-2.3**	41 % to 87 %
	Ocean/sea-ice	**-1.3**	-3.2	**-1.7**	-3.2	-1.7	-1.9	**-1.6**	-2.6	-

manuscript submitted to *Journal of Advances in Modeling Earth Systems (JAMES)*

Zone (SPCZ), indicated by the dipole pattern in precipitation bias in the southwest Pacific. Each of these biases is substantially corrected or (over the Western Pacific Warm Pool) slightly overcorrected, by the ML (Figure 9b). However, new precipitation biases emerge in the Indian Ocean off the east coast of central Africa and in the Bay of Bengal. The spatial pattern of the biases is similar in the other climates in the baseline configuration, though the error magnitude somewhat increases in mid-latitude ocean regions (not shown). Something similar can be said for the ML-corrected cases.

Figure 9c shows that the land root mean square error of the annual mean spatial pattern of precipitation is improved over the baseline in every year of the ML-corrected simulations in each climate. This improvement is on average between 17 % to 30 % depending on the climate (shown in last column of Table 2). Surprisingly, the RMSE over land of the implied precipitation in the nudged runs worsens faster as the SSTs warm than that of the baseline or ML-corrected runs, eventually becoming larger than in the baseline run in the +8 K climate. This is because as the climate warms, grid-scale noise in the column integrated drying tendency due to nudging over land (Fig. 3d) increases; however, broad-scale precipitation biases remain small in the nudging runs for all four climates. The ML correction learned from the humidity nudging tendencies smooths out the grid-scale noise when making predictions, allowing it to reduce this pattern error in the +8 K climate.

Figure 9d shows that the ML-corrected simulation almost eliminates the 0.7 mm/d land time-mean dry bias of the baseline simulation in all climates. As in B22, we attribute this primarily to the ML surface radiation correction.

Figure 9e depicts the RMSE of the annual mean spatial pattern of precipitation computed over ocean and sea ice. Unlike over land, the error magnitudes increase as the SSTs warm. The RMSEs of the baseline and ML-corrected runs are not robustly different (Table 2), indicating that the ML does not help or hurt ocean/sea-ice precipitation estimates. Over the oceans, the precipitation biases of the nudged runs are smaller and less affected by grid-scale noise than over land, and their precipitation pattern RMSE remains much smaller than for the baseline or ML-corrected simulations. In other words, despite the cleaner improved precipitation signal over ocean/sea ice in the nudged runs, we have a more challenging time improving the precipitation climatology over that region with ML.

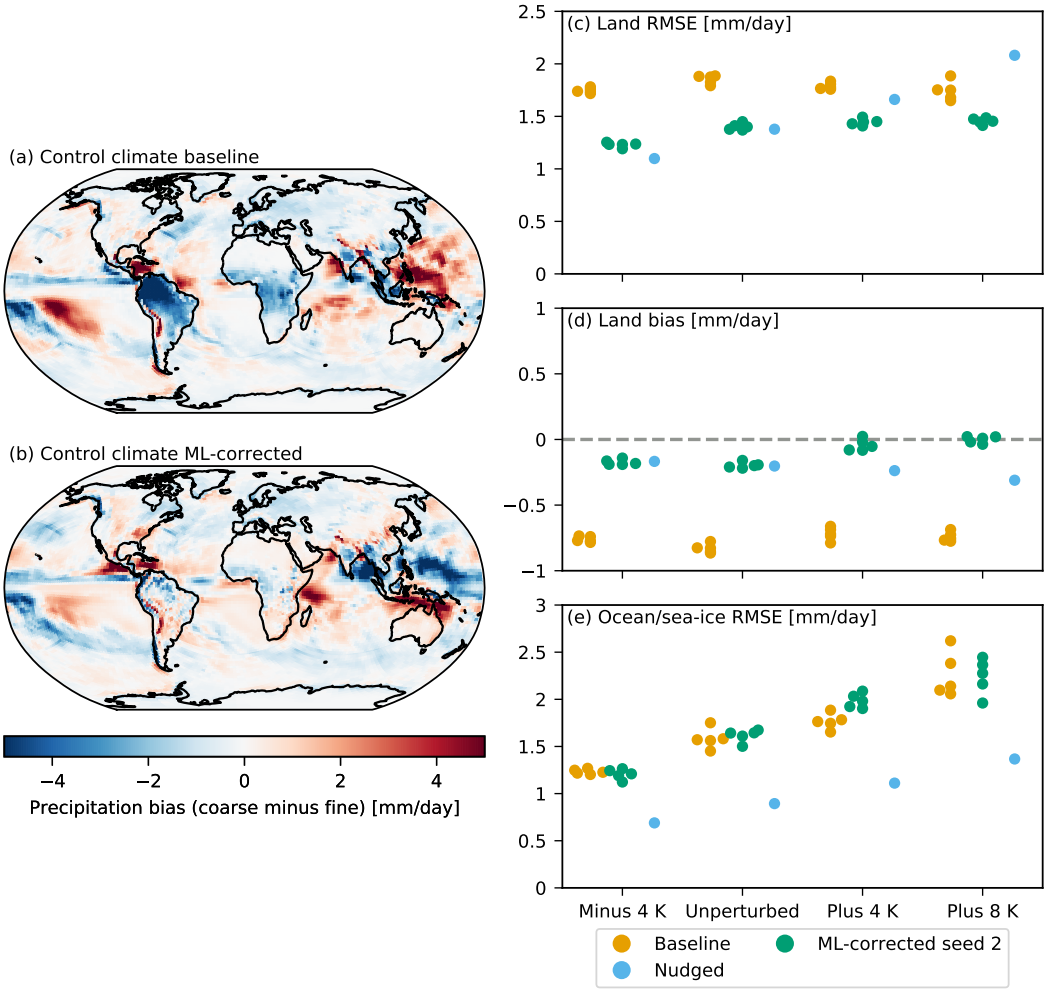

Figure 9. Time-mean spatial pattern of the precipitation bias in the baseline (a) and ML-corrected seed 2 (b) control climate simulations. Land root mean square error (RMSE) in the time-mean spatial pattern of the precipitation rate during each year of baseline (orange dots), and ML-corrected seed 2 (green dots), and nudged (blue dots) simulations in each climate (c). Panels (d) and (e) are structured similarly, but depict the mean bias over land and the RMSE over ocean/sea-ice, respectively.

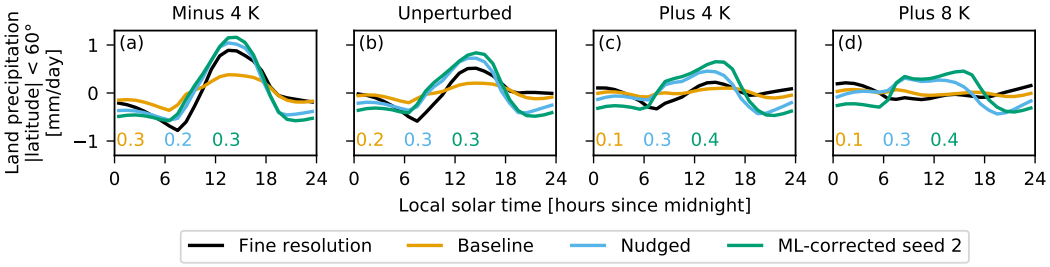

Figure 10. Diurnal cycle of precipitation over land in the ∼25 km reference (black curve), ∼200 km baseline (orange curve), ∼200 km nudged (blue curve), and ∼200 km ML-corrected (green curve) simulations in each climate, with the overall time-and-land mean removed. While the observations are not shown in these plots, for consistency, as in Figure 3c, the diurnal cycle is computed over land regions where the latitude is between 60°S and 60°N. The values in the lower left corner of each panel represent the root mean square error relative to the fine-resolution curve for the baseline, nudged, and ML-corrected simulations respectively.

.

Figure 10 shows the diurnal cycle of precipitation in the fine-grid, baseline, nudged, and seed 2 ML-corrected simulations in each climate. For each model configuration, the mean precipitation rate over land stays relatively constant across climates. However, the amplitude of the diurnal cycle over land in the reference simulation is largest in the coldest climate (−4 K) and absent in the warmest (+8 K). The baseline runs capture this trend but with much-reduced amplitude in all climates. The nudged and seed 2 ML-corrected runs capture some, but not all, of this amplitude decrease with warming SSTs. Due to the amplitude overestimation, if we compute an RMSE of the diurnal cycle of land precipitation vs. the 25 km reference simulations, we find that the ML-corrected simulations do slightly worse than the baseline ones; however they closely match the nudged simulations used to train the ML. Thus these diurnal cycle errors derive mainly from the nudging approach, not lack of ML skill.

3.6.2 Surface temperature

Surface temperature over land, which is an emergent property of the simulations not directly modified by our ML, is robustly improved in the seed 2 ML corrected run (the prescribed sea surface temperatures are trivially bias-free). Figure 11 shows the time-mean bias in surface temperature in the baseline and seed 2 ML-corrected runs in the

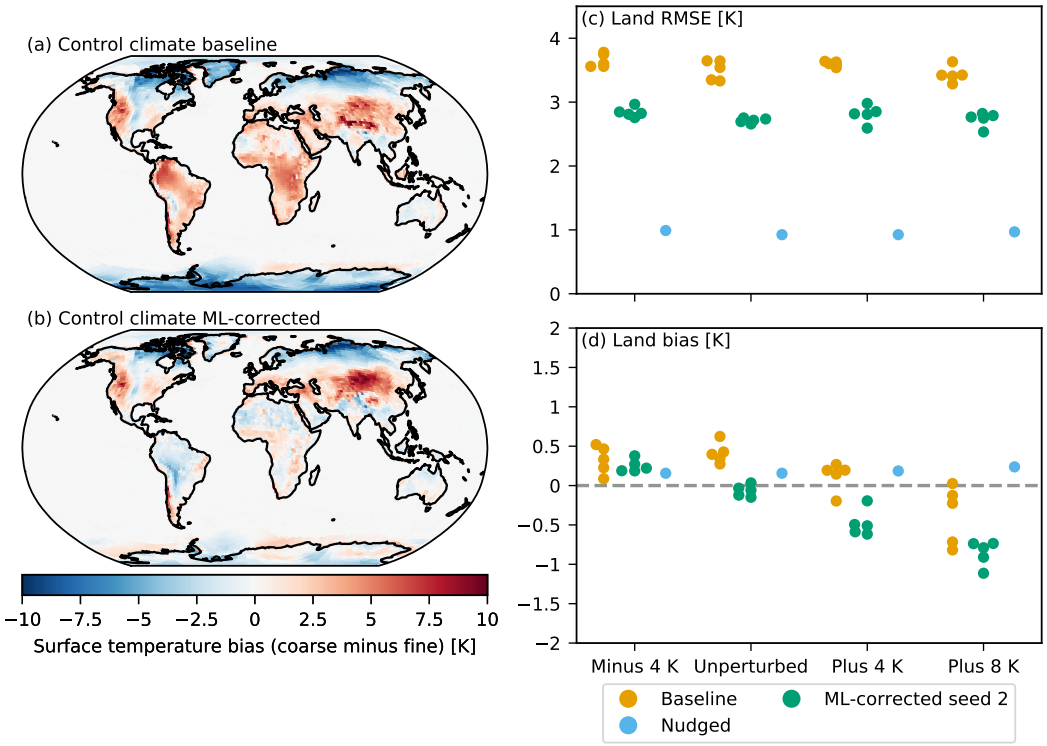

Figure 11. As in Figure 9 (excluding ocean/sea-ice RMSE), for the time-mean spatial pattern of the surface temperature bias.

control climate. The baseline run (Figure 11a) has 2 K to 5 K warm biases over much of the tropics and mid-latitudes, and cold biases in the polar regions that intensify poleward to as much as −7.5 K. Like the baseline precipitation biases, the baseline land surface temperature biases have a similar spatial pattern and RMSEs in the other climates, though the positive bias over tropical land regions in the baseline begins to become more over-corrected in the +4 K and +8 K ML-corrected simulations.

Figures 11c and d summarize the RMSE and mean bias of surface temperature over land in each year of the baseline and ML-corrected simulations versus the reference fine grid simulation in each climate. In all four climates, surface temperature RMSE over land is improved over the baseline by the seed 2 ML-corrected runs by 20 % to 23 %; as in the land precipitation RMSE case, this result is robust across years (Table 2).

Positive mean biases in the tropics and mid-latitudes offset negative mean biases in the polar regions in the baseline simulations to result in largely unbiased baseline simulations in each climate. In the ML-corrected runs, there is more variability depending

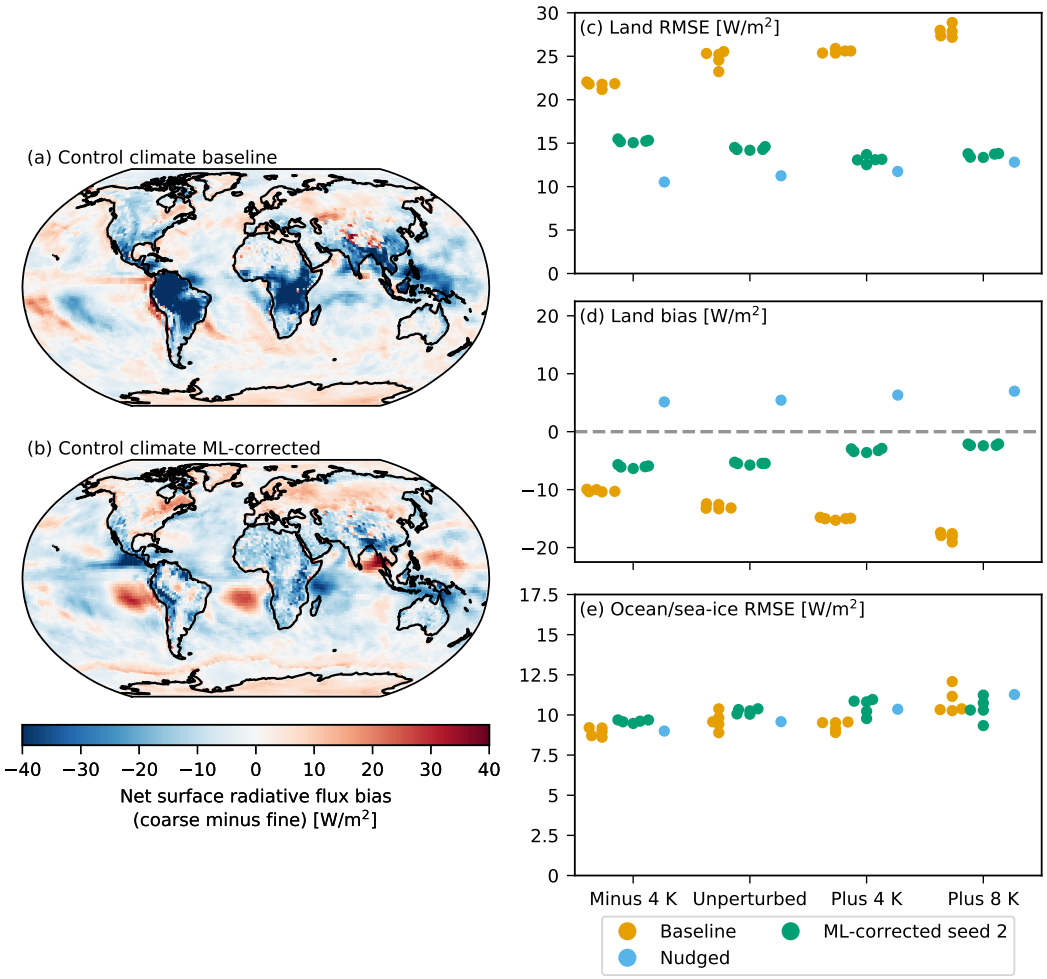

Figure 12. As in Figure 11, for the time-mean spatial pattern of the net surface radiative flux bias.

on the climate. In the −4 K and control climates, the land-mean surface temperature bias is near zero; however, in the +4 K and +8 K climates, cool biases over tropical land regions drive an overall negative land surface temperature bias.

3.6.3 Net surface radiative flux

As mentioned in Section 2.3, we use ML to correct the downwelling radiative fluxes used to force the underlying surface. Figures 12a and b compare the time mean bias in net surface radiative flux in the control climate in the baseline and seed 2 ML-corrected run. The baseline model has large negative biases in the baseline over tropical land regions, contributing to RMSEs over land of over 20 W m^{-2} in all climates. In the ML-corrected

runs this RMSE is cut by 30 % to 51 % (Figure 12c and Table 2), and the mean negative bias is greatly reduced (Figure 12d), indicating that the strong offline skill of the ML surface radiative flux model over land, illustrated in Section 3.4, translates well into online simulations. The large offline bias in downward shortwave radiation in the ocean stratocumulus regions noted in Section 3.4 persists in online simulations, and moderate negative biases in net surface shortwave radiation emerge online throughout the non-stratocumulus ocean regions. These biases would be of concern if we coupled the ML-corrected atmosphere model to a dynamical ocean model, but they have no impact on our prescribed-SST simulations.

3.6.4 Temperature, specific humidity, and circulation biases

While we predict tendency corrections to the temperature and specific humidity at each level of the atmosphere in ML-corrected runs, these predictions do not necessarily improve the zonal mean climatological biases in these fields over those in the baseline simulations.

Figures 13a and b show the zonal mean temperature biases in the baseline and seed 2 ML-corrected simulations in the control climate. The baseline simulation has a roughly 1 K warm bias in the boundary layer in all but the polar regions, where there is a larger cold bias. and has a mid-tropospheric cold bias of about 1 K at all latitudes. The largest temperature bias is a vertical dipole pattern of magnitude 2–3 K in the polar stratosphere. In the ML-corrected simulation, the bias is reduced near the surface but is more severe in the polar mid-troposphere. Above 200 hPa, we are intentionally tapering the corrective tendencies, so we might expect the ML-corrected simulation to have similar temperature biases as the baseline. However, large warm biases develop, locally exceeding 5 K. These may be associated with circulation changes induced by ML predictions lower in the atmosphere.

Specific humidity biases are shown in Figures 13 c and d). ML again helps reduce biases of the baseline model in some regions but not others. The baseline model has negative specific humidity biases around $-0.2\,\mathrm{g\,kg^{-1}}$ near the surface in the polar regions, positive biases in the mid-latitude troposphere around $0.3\,\mathrm{g\,kg^{-1}}$, negative biases in the deep tropics between $-0.1\,\mathrm{g\,kg^{-1}}$ to $-0.4\,\mathrm{g\,kg^{-1}}$. The ML-corrected run reduces the surface bias near the South Pole as well as the mid-latitude positive mid-tropospheric bi-

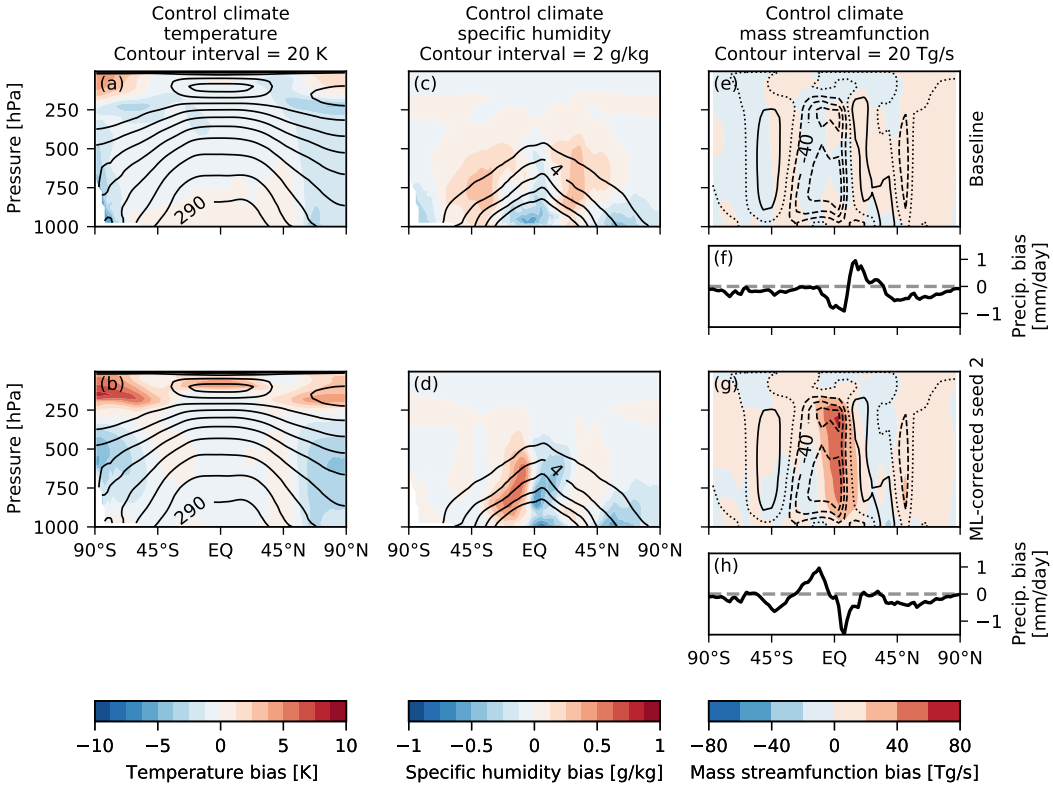

Figure 13. Time and zonal mean biases of temperature, specific humidity, and the mass streamfunction in the baseline (top row) and ML-corrected seed 2 simulations (bottom row) relative to the fine resolution reference in the control climate (filled contours). The line contours represent the reference values of the fields in the fine-resolution reference dataset, with contour intervals shown in the column titles. Panels (f) and (h) in the third column show the zonal mean bias in precipitation for the baseline and seed 2 ML-corrected simulations, respectively.

754 ases, but introduces a dipole bias pattern in the tropics, anomalously moist just south
755 of the equator and anomalously dry just north by up to $\pm 5\,\mathrm{g\,kg^{-1}}$.

756 The humidity biases in the ML-corrected simulation are consistent with a change
757 in the zonal mean overturning simulation. Figure 13g shows the bias in the zonal mean
758 mass streamfunction for the ML-corrected case. It depicts a southward shift in the up-
759 ward branch of the overturning circulation, also evident as a dipole bias in zonal mean
760 precipitation (Figure 13h). Figure 9b suggests the precipitation bias is mainly coming
761 from the West Pacific/Bay of Bengal region and off the east coast of Africa. The zonal
762 mean precipitation biases in the baseline simulation (Figure 13f), though comparable in
763 magnitude to those in the ML-corrected run, cannot be so easily explained by the rel-
764 atively small and unfeatured biases in the streamfunction (Figure 13e).

765 4 Discussion and Conclusion

766 In this study we extended the approach described in B22 to train ML models for
767 application in multiple climates and around the annual cycle. The ML adds state-dependent
768 corrections to the temperature and specific humidity tendencies, and predicts surface ra-
769 diative fluxes, to optimally correct single timestep tendencies of the coarse model (in-
770 cluding land-atmosphere interaction) to match those of a fine-grid reference simulation.
771 Although this method does not guarantee good longer-term skill, we showed that with
772 ablation and tapering of ML inputs and outputs in the uppermost 25 model levels, we
773 were able to obtain robustly stable ML-corrected simulations. The annual mean climate
774 biases in ML-corrected runs depend somewhat on the random seeds used to initialize the
775 neural networks before training. However, each of the four NNs we tested online improve
776 the land RMSE of the annual mean spatial pattern of precipitation, and three out of the
777 four improve the surface temperature climate.

778 We presented five-year prognostic simulations with the seed 2 NN, selected because
779 they had the smallest overall surface temperature and humidity biases over the first sim-
780 ulated year across the four climates. Depending on the climate, ML improved the land
781 RMSE of precipitation by 17 – 30%, and the land RMSE of surface temperature by 20
782 – 23%. The ML corrections also improved the amplitude of the diurnal cycle of precip-
783 itation over land in the $-4\,\mathrm{K}$ and control climates, but slightly exaggerated it in the $+4\,\mathrm{K}$
784 and $+8\,\mathrm{K}$ climates. In contrast to the land-surface-level metrics, ML tendency correc-

tions generally did not improve the precipitation or net surface radiative flux RMSE over ocean/sea-ice, or the zonal mean bias pattern of temperature or specific humidity, and through dynamical feedbacks actually introduced errors into the zonal mean overturning circulation. Although not shown here, we also performed 5-year simulations with seed 3 NN with comparable results.

While we obtain robust improvements in precipitation and surface temperature over the baseline in the ML-corrected runs in the individual climates, the differences between simulated climates are generally not significantly improved or worsened (not shown). A better ML correction which made larger improvements in the individual climates would be more likely to translate to improvements in the difference between climates.

While encouraging, the relative improvements in precipitation RMSE are not as large as the 25–30% obtained by B22. Our baseline 200 km simulation in the control climate has a much lower RMSE versus the fine-grid reference, $1.6\,\mathrm{mm\,d^{-1}}$ globally compared with $3.7\,\mathrm{mm\,d^{-1}}$ in B22, making it more difficult to improve upon. Three contributors to the improved baseline skill were: first, using the same microphysics configuration (including saturation adjustment within the dynamical core) as in the reference model, second, using a coarser resolution "fine" resolution target model ($\sim$25 km resolution versus $\sim$3 km resolution), which we assess skill against, and third, computing an RMSE for the time-mean over the full annual cycle rather than a single 40 d period.

There are still substantial differences in the surface downwelling radiation predicted by the physical parameterizations of the baseline and reference models; these differences can feed back on the land surface. As in B22, overriding the physical parameterization's predictions of these fluxes with the ML's greatly reduces surface radiation bias in prognostic runs, and helps to remove land-mean precipitation biases and significantly reduce land surface temperature biases.

Future work on a number of aspects of the problem might improve on these results; three are discussed more below. First, as mentioned in B22, it would be beneficial to find a way to re-introduce ML corrections of the horizontal wind tendencies. This currently is an inconsistency in our approach; when producing the training data we nudge the horizontal winds, but we only train models to predict the temperature and specific humidity nudging tendencies, because we found in B22 (and verified in the setting of the present study) that the nudging-trained approach for predicting wind tendency corrections leads

to large temperature biases through circulation feedbacks. If we can find a way to re-introduce these in a way that does not lead to these large temperature biases, it might reduce the circulation biases noted in this study.

Second, a corrective approach similar to the way we handle temperature and moisture might improve the skill of the ML for predicting the downwelling radiative fluxes. We currently attribute the fine-coarse surface radiation differences mainly to cloud differences. If the coarse-grid clouds are more skillful predictors of the fine-grid clouds than are the column temperature and humidity profiles, then a corrective approach might add skill. Figure 12 of the present study (and a similar figure in B22) suggests this might hold in the subtropical marine stratocumulus regions. This might enable skill improvements over the baseline in predicting the net surface radiative flux over ocean, which the current approach does not achieve (Table 2). This would become important if this ML approach were used as part of an ocean-coupled model.

Third, we showed that an NN trained with one random seed systematically produced different climate biases compared to networks trained with other seeds across all climates. It would be useful to develop a more systematic way of optimizing the ML models to not only reduce single timestep errors, but also reduce errors in climate statistics. For instance, Balogh et al. (2022) used a targeted set of online simulations to tune embedded parameters within an ML model to optimize climate-like statistics in an idealized model problem.

In future work it could also be interesting to address questions related to how well this existing ML approach might apply to an interpolation-type problem, e.g. correcting a coarse-resolution simulation in a $+2\,\mathrm{K}$ climate, or potentially modify the approach such that it could be applied in an extrapolation context, e.g. in a climate not within the bounds of the training data, something explored at least in an offline context on a different ML problem in Beucler et al. (2021). We also acknowledge that despite its success, the nudging method for generating an ML target has fundamental limitations when and where physical processes are adjusting to changing conditions faster than the nudging timescale (Kruse et al., 2022, submitted to *JAMES*) and will need to be improved upon. One manifestation documented here was a distorted diurnal cycle of precipitation over land.

In conclusion, the results presented here are an important step toward applying corrective ML through coarse-graining in a model with realistic topography across the full annual cycle, and in multiple different climates. Substantial further improvements should be achievable using the best possible reference models, ML methodologies and training approaches.

Acknowledgments

We thank Vulcan Inc., the Allen Institute for Artificial Intelligence, and NOAA-GFDL for supporting this work. The one-year ∼200 km resolution spin-up and two-year ∼25 km resolution reference simulations in each of the four climates were completed on NOAA's Gaea supercomputer. Additionally, we appreciate Kai-Yuan Cheng for permitting us to use his version of the `UFS_UTILS` package compiled on Gaea, as well as Linjiong Zhou for discussion regarding the configuration of the microphysics in the simulations described in this study. More broadly, we acknowledge NOAA-EMC, NOAA-GFDL, and the UFS Community for publicly hosting source code for the FV3GFS model and the `UFS_UTILS` package, and NOAA-EMC for providing the necessary forcing data to run FV3GFS. The specific code and configuration files used to run and analyze the results of the experiments in this study are contained in a GitHub repository, `https://github.com/ai2cm/nudge-to-fine-25km-manuscript-workflow`, which is archived at Zenodo (`https://doi.org/10.5281/zenodo.6584122`). The monthly mean GPCP precipitation dataset was obtained through `https://psl.noaa.gov/data/gridded/data.gpcp.html`, the monthly mean ERA5 precipitable water dataset was obtained through `https://doi.org/10.24381/cds.f17050d7`, and the half-hourly IMERG precipitation dataset was obtained through `https://doi.org/10.5067/GPM/IMERG/3B-HH/06`.

References

Adler, R. F., Huffman, G. J., Chang, A., Ferraro, R., Xie, P.-P., Janowiak, J., ... Nelkin, E. (2003, December). The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present). *Journal of Hydrometeorology*, *4*(6), 1147–1167. doi: 10.1175/1525-7541(2003)004⟨1147:TVGPCP⟩2.0.CO;2

Alpert, J., Kanamitsu, M., Caplan, P. M., Sela, J. G., White, G. H., & Kalnay, E. (1988). Mountain induced gravity wave drag parameterization in the NMC

879 medium-range forecast model. In *Eighth Conference on Numerical Weather*
880 *Prediction* (pp. 726–733). Baltimore, MD: American Meteorological Society.
881 Balogh, B., Saint-Martin, D., & Ribes, A. (2022). How to Calibrate a Dynamical
882 System With Neural Network Based Physics? *Geophysical Research Letters*,
883 *49*(8), e2022GL097872. doi: 10.1029/2022GL097872
884 Beucler, T., Pritchard, M., Yuval, J., Gupta, A., Peng, L., Rasp, S., ... Gentine, P.
885 (2021, December). Climate-Invariant Machine Learning. *arXiv:2112.08440*
886 *[physics]*.
887 Brenowitz, N. D., Beucler, T., Pritchard, M., & Bretherton, C. S. (2020, Decem-
888 ber). Interpreting and Stabilizing Machine-Learning Parametrizations of
889 Convection. *Journal of the Atmospheric Sciences*, *77*(12), 4357–4375. doi:
890 10.1175/JAS-D-20-0082.1
891 Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially Extended Tests of a Neural
892 Network Parametrization Trained by Coarse-Graining. *Journal of Advances in*
893 *Modeling Earth Systems*, *11*(8), 2728–2744. doi: 10.1029/2019MS001711
894 Bretherton, C. S., Henn, B., Kwa, A., Brenowitz, N. D., Watt-Meyer, O., McGib-
895 bon, J., ... Harris, L. (2022). Correcting Coarse-Grid Weather and Climate
896 Models by Machine Learning From Global Storm-Resolving Simulations. *Jour-*
897 *nal of Advances in Modeling Earth Systems*, *14*(2), e2021MS002794. doi:
898 10.1029/2021MS002794
899 Bretherton, C. S., Peters, M. E., & Back, L. E. (2004, April). Relationships between
900 Water Vapor Path and Precipitation over the Tropical Oceans. *Journal of Cli-*
901 *mate*, *17*(7), 1517–1528. doi: 10.1175/1520-0442(2004)017⟨1517:RBWVPA⟩2.0
902 .CO;2
903 Caldwell, P. M., Terai, C. R., Hillman, B., Keen, N. D., Bogenschutz, P., Lin, W.,
904 ... Zender, C. S. (2021). Convection-Permitting Simulations With the E3SM
905 Global Atmosphere Model. *Journal of Advances in Modeling Earth Systems*,
906 *13*(11), e2021MS002544. doi: 10.1029/2021MS002544
907 Chollet et al., F. (2015). *Keras*.
908 Christopoulos, C., & Schneider, T. (2021). Assessing Biases and Climate Impli-
909 cations of the Diurnal Precipitation Cycle in Climate Models. *Geophysical Re-*
910 *search Letters*, *48*(13), e2021GL093017. doi: 10.1029/2021GL093017
911 Ek, M. B., Mitchell, K. E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., ... Tarp-

ley, J. D. (2003). Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *Journal of Geophysical Research: Atmospheres*, *108*(D22). doi: 10.1029/2002JD003296

Gayno, G., Beck, J., Reames, L., programmer, G., Wright, D., Hu, M., ... Gerheiser, K. (2020, August). *UFS_UTILS.* Unified Forecast System (UFS).

Han, J., & Pan, H.-L. (2011, August). Revision of Convection and Vertical Diffusion Schemes in the NCEP Global Forecast System. *Weather and Forecasting*, *26*(4), 520–533. doi: 10.1175/WAF-D-10-05038.1

Han, J., Witek, M. L., Teixeira, J., Sun, R., Pan, H.-L., Fletcher, J. K., & Bretherton, C. S. (2016, February). Implementation in the NCEP GFS of a Hybrid Eddy-Diffusivity Mass-Flux (EDMF) Boundary Layer Parameterization with Dissipative Heating and Modified Stable Boundary Layer Mixing. *Weather and Forecasting*, *31*(1), 341–352. doi: 10.1175/WAF-D-15-0053.1

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A Moist Physics Parameterization Based on Deep Learning. *Journal of Advances in Modeling Earth Systems*, *12*(9), e2020MS002076. doi: 10.1029/2020MS002076

Harris, L., Chen, X., Putman, W., Zhou, L., & Chen, J.-H. (2021). A Scientific Description of the GFDL Finite-Volume Cubed-Sphere Dynamical Core. doi: 10.25923/6NHS-5897

Harris, L., Zhou, L., Lin, S.-J., Chen, J.-H., Chen, X., Gao, K., ... Stern, W. (2020). GFDL SHiELD: A Unified System for Weather-to-Seasonal Prediction. *Journal of Advances in Modeling Earth Systems*, *12*(10), e2020MS002223. doi: 10.1029/2020MS002223

Held, I. M., & Soden, B. J. (2006, November). Robust Responses of the Hydrological Cycle to Global Warming. *Journal of Climate*, *19*(21), 5686–5699. doi: 10.1175/JCLI3990.1

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

Huffman, G. J., Stocker, E. F., Bolvin, D. T., Nelkin, E. J., & Tan, J. (2019). GPM IMERG Final Precipitation L3 Half Hourly 0.1 degree x 0.1 degree V06.

945 doi: 10.5067/GPM/IMERG/3B-HH/06

946 Iacono, M. J., Delamere, J. S., Mlawer, E. J., Shephard, M. W., Clough, S. A., &
947 Collins, W. D. (2008). Radiative forcing by long-lived greenhouse gases: Cal-
948 culations with the AER radiative transfer models. *Journal of Geophysical*
949 *Research: Atmospheres*, *113*(D13). doi: 10.1029/2008JD009944

950 Kruse, C. G., Bacmeister, J. T., Zarzycki, C. M., Larson, V. E., & Thayer-
951 Calder, K. (2022, December). *Do Nudging Tendencies Depend on*
952 *the Nudging Timescale Chosen in Atmospheric Models?* [Preprint].
953 http://www.essoar.org/doi/10.1002/essoar.10510369.1. Earth and Space
954 Science Open Archive. doi: 10.1002/essoar.10510369.1

955 Lott, F., & Miller, M. J. (1997). A new subgrid-scale orographic drag parametriza-
956 tion: Its formulation and testing. *Quarterly Journal of the Royal Meteorologi-*
957 *cal Society*, *123*(537), 101–127. doi: 10.1002/qj.49712353704

958 McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J. P. S.,
959 Davis, E. C., ... Fuhrer, O. (2021, July). Fv3gfs-wrapper: A Python wrapper
960 of the FV3GFS atmospheric model. *Geoscientific Model Development*, *14*(7),
961 4401–4409. doi: 10.5194/gmd-14-4401-2021

962 Molina, M. J., Gagne, D. J., & Prein, A. F. (2021). A Benchmark to Test Gen-
963 eralization Capabilities of Deep Learning Methods to Classify Severe Con-
964 vective Storms in a Changing Climate. *Earth and Space Science*, *8*(9),
965 e2020EA001490. doi: 10.1029/2020EA001490

966 Monteiro, J. M., McGibbon, J., & Caballero, R. (2018, September). Sympl (v. 0.4.0)
967 and climt (v. 0.15.3) – towards a flexible framework for building model hier-
968 archies in Python. *Geoscientific Model Development*, *11*(9), 3781–3794. doi:
969 10.5194/gmd-11-3781-2018

970 NCEI. (2020, August). *Global Forecast System.* http://www.ncei.noaa.gov/products/weather-
971 climate-models/global-forecast.

972 O'Gorman, P. A., & Dwyer, J. G. (2018). Using Machine Learning to Parameterize
973 Moist Convection: Potential for Modeling of Climate, Climate Change, and
974 Extreme Events. *Journal of Advances in Modeling Earth Systems*, *10*(10),
975 2548–2563. doi: 10.1029/2018MS001351

976 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ...
977 Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of*

Machine Learning Research, *12*(85), 2825–2830.

Putman, W. M., & Lin, S.-J. (2007, November). Finite-volume transport on various cubed-sphere grids. *Journal of Computational Physics*, *227*(1), 55–78. doi: 10.1016/j.jcp.2007.07.022

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

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., ... Becker, E. (2014, March). The NCEP Climate Forecast System Version 2. *Journal of Climate*, *27*(6), 2185–2208. doi: 10.1175/JCLI-D-12-00823.1

Shepherd, T. G. (2014, October). Atmospheric circulation as a source of uncertainty in climate change projections. *Nature Geoscience*, *7*(10), 703–708. doi: 10.1038/ngeo2253

Stevens, B., Acquistapace, C., Hansen, A., Heinze, R., Klinger, C., Klocke, D., ... Zängl, G. (2020). The Added Value of Large-eddy and Storm-resolving Models for Simulating Clouds and Precipitation. *Journal of the Meteorological Society of Japan. Ser. II*, *98*(2), 395–435. doi: 10.2151/jmsj.2020-021

Thiébaux, J., Rogers, E., Wang, W., & Katz, B. (2003, May). A New High-Resolution Blended Real-Time Global Sea Surface Temperature Analysis. *Bulletin of the American Meteorological Society*, *84*(5), 645–656. doi: 10.1175/BAMS-84-5-645

UFS Community. (2020, October). *UFS Weather Model.* Zenodo. doi: 10.5281/zenodo.4460292

van der Wiel, K., Kapnick, S. B., Vecchi, G. A., Cooke, W. F., Delworth, T. L., Jia, L., ... Zeng, F. (2016, November). The Resolution Dependence of Contiguous U.S. Precipitation Extremes in Response to CO2 Forcing. *Journal of Climate*, *29*(22), 7991–8012. doi: 10.1175/JCLI-D-16-0307.1

Wang, X., Han, Y., Xue, W., Yang, G., & Zhang, G. J. (2021, September). Stable climate simulations using a realistic GCM with neural network parameterizations for atmospheric moist physics and radiation processes. *Geoscientific Model Development Discussions*, 1–35. doi: 10.5194/gmd-2021-299

Waskom, M. L. (2021, April). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021. doi: 10.21105/joss.03021

Watt-Meyer, O., Brenowitz, N. D., Clark, S. K., Henn, B., Kwa, A., McGibbon, J., ... Bretherton, C. S. (2021). Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations. *Geophysical Research Letters*, *48*(15), e2021GL092555. doi: 10.1029/2021GL092555

Yanai, M., Esbensen, S., & Chu, J.-H. (1973, May). Determination of Bulk Properties of Tropical Cloud Clusters from Large-Scale Heat and Moisture Budgets. *Journal of the Atmospheric Sciences*, *30*(4), 611–627. doi: 10.1175/1520-0469(1973)030⟨0611:DOBPOT⟩2.0.CO;2

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

Yuval, J., & O'Gorman, P. A. (2021, July). *Neural-network parameterization of subgrid momentum transport in the atmosphere* [Preprint]. http://www.essoar.org/doi/10.1002/essoar.10507557.1. Earth and Space Science Open Archive. doi: 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. doi: 10.1029/2020GL091363

Zhou, L., Lin, S.-J., Chen, J.-H., Harris, L. M., Chen, X., & Rees, S. L. (2019, July). Toward Convective-Scale Prediction within the Next Generation Global Prediction System. *Bulletin of the American Meteorological Society*, *100*(7), 1225–1243. doi: 10.1175/BAMS-D-17-0246.1