Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations

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

[1]Vulcan Inc. and Allen Institute for Artificial Intelligence, Seattle, WA
[2]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ

Key Points:

- With a state-dependent machine learning correction, a coarse-grid global atmospheric model evolves more like a fine-grid model version
- The skill of coarse-grid weather forecasts and time-mean rainfall are improved compared to the fine-grid storm-resolving reference model
- Accounting for cloud biases by correcting surface downwelling radiation improves surface turbulent fluxes and precipitation over land

Corresponding author: Chris Bretherton, `christopherb@allenai.org`

Abstract

Global atmospheric 'storm-resolving' models with horizontal grid spacing of less than 5 km resolve deep cumulus convection and flow in complex terrain. They promise to be reference models that could be used to improve computationally affordable coarse-grid global climate models across a range of climates, reducing uncertainties in regional precipitation and temperature trends. Here, machine learning of nudging tendencies as functions of column state is used to correct the physical parameterization tendencies of temperature, humidity, and optionally winds, in a real-geography coarse-grid model (FV3GFS with a 200 km grid) to be closer to those of a 40-day reference simulation using X-SHiELD, a modified version of FV3GFS with a 3 km grid. Both simulations specify the same historical sea-surface temperature fields. This methodology builds on a prior study using a global observational analysis as the reference. The coarse-grid model without machine learning corrections has too little cloud, causing too much daytime heating of land surfaces that creates excessive surface latent heat flux and rainfall. This bias is avoided by learning downwelling radiative flux from the fine-grid model. The best configuration uses learned nudging tendencies for temperature and humidity but not winds. Neural nets slightly outperform random forests. Forecasts of 850 hPa temperature gain 18 hours of skill at 3–7 day leads and time-mean precipitation patterns are improved 30% by applying the ML correction. Adding machine-learned wind tendencies improves 500 hPa height skill for the first five days of forecasts but degrades time-mean upper tropospheric temperature and zonal wind patterns thereafter.

Plain Language Summary

Global weather and climate models can be made more realistic by using a finer computational grid, but this is too expensive for routine use. We design a machine-learned correction to make a more economical coarse-grid model better track a fine-grid reference version of this model. The correction is trained using a limited, computationally affordable, period of fine-grid model output. It is applied interactively during the coarse-grid simulation. As desired, adding the correction substantially improves how well weather forecasts and time-mean rainfall patterns with the coarse-grid model match the fine-grid reference.

1 Introduction

In the last few years, global atmospheric 'storm-resolving' models (GSRMs) with horizontal grid spacing of less than 5 km have become computationally feasible for simulations of months and longer (Tomita et al., 2005; Stevens et al., 2019). Their fine grids enable these models to resolve vertical motions within deep convective cloud systems, rather than relying on assumption-laden cumulus parameterizations. They also more fully resolve circulations in complex orography that modulate precipitation and create drag. Attractively, because they resolve more fine-scale atmospheric circulations, many of their subgrid parameterization assumptions can be simpler and more testable than for coarser-grid climate models.

The recent DYAMOND intercomparison of nine such GSRMs (Stevens et al., 2019) shows their potential for more accurately simulating severe weather and global climate, especially as they become observationally-calibrated backbones for global weather forecasts. But computational constraints make it unlikely that GSRMs will soon be practical for the atmospheric part of century-long climate simulations. How, then, can the climate projection enterprise benefit from this exciting new class of global models?

This paper will explore coarse-graining (hereafter coarsening) of GSRMs for training machine learning (ML) parameterizations for use in coarse-grid global atmospheric models. Past studies have addressed aspects of this problem but not provided an end-to-end solution. A key challenge is training the ML to make stable, accurate online simulations in which it is coupled to the dynamical core (numerical flow solver) and other components of the model.

Some studies have demonstrated online skill in simplified settings such as zonally-symmetric aquaplanets and/or superparameterized models that include artificial scale-separation assumptions (Rasp et al., 2018; Brenowitz & Bretherton, 2019; Yuval & O'Gorman, 2020, 2021). This makes the ML training problem easier to precisely formulate, but also sidesteps important real-world complications such as orography, land surface heterogeneity, complex coastlines, etc. Other studies have tackled real-world geography but only demonstrated offline or single-column ML skill (Han et al., 2020; Mooers et al., 2021).

Here, we present a coarsening-based ML approach that is formulated to work within a state-of-the-art global atmospheric model, FV3GFS, used operationally by the National

Oceanic and Atmospheric Administration (NOAA) in the U. S. for weather forecasting on daily to seasonal timescales. We show that this approach has attractive on-line skill across these timescales and especially improves the coarse-grid simulation of time-mean precipitation over land, a challenge for many conventional climate models.

Our new 'nudge-to-fine' approach is a variant of the 'nudge-to-observations' methodology recently introduced by Watt-Meyer et al. (2021), hereafter WM21. Both approaches are forms of state-dependent bias correction, (e.g. Leith, 1978; DelSole et al., 2008); see WM21 for more historical context. WM21 add a 'corrective ML' scheme to the physical parameterizations that makes the coarse-grid model evolve similarly to a reference data set. For nudging-based ML training, a linear relaxation term is added to the coarse model that 'nudges' the temperature, humidity and wind at every grid point toward the reference data set; the ML learns these 'nudging tendencies'. In WM21, the reference was six-hourly global observational analyses. Here, the reference is coarsened output from a fine-grid GSRM. Nudging to a reference has also been used to facilitate parameter estimation within climate model parameterizations (Lyu et al., 2018).

It is conceptually helpful if the dynamical core and shared parameterizations of the coarse and fine-grid models are similar, so that the learned correction reflects the coarser grid spacing rather than the different model formulations. We use FV3GFS with a C48 cubed-sphere grid ($\sim$ 200 km grid spacing) and 79 vertical terrain-following coordinate levels as the coarse model and a modified version of FV3GFS, X-SHiELD, which has a C3072 grid (approximately 3 km spacing), as the fine model. X-SHiELD (Harris et al., 2020) is developed at NOAA's Geophysical Fluid Dynamics Laboratory, or GFDL. Both FV3GFS and X-SHiELD use D-grid horizontal staggering of wind components relative to cell centers. We use a basket of five weather forecast and time-mean bias metrics to choose between candidate ML configurations.

Section 2 describes the models and the coarsening method. Section 3 presents the nudged training simulation approach, including forcing of the land surface. Section 4 describes the ML methods used. Section 5 discusses ML-relevant aspects of the nudging tendency fields and the offline ML skill. Section 6 discusses prognostic skill for weather forecasts and time-mean biases. Section 7 interprets the nudging tendencies as the sum of a physics component due to fine-coarse parameterization tendency differences and a residual 'dynamics' component mainly due to fine-coarse vertical motion differences. Con-

clusions follow in Section 8. The acknowledgments include a statement of data and software availability.

2 3 km reference simulation and coarsening method

2.1 Reference simulation

We performed a 40-day fine-grid reference simulation using X-SHiELD. Our X-SHiELD implementation was configured similarly to the GFDL submission to the DYAMOND intercomparison of GSRMs (Stevens et al., 2019), with 79 vertical levels. The FV3GFS gravity wave drag and deep cumulus and parameterizations were disabled in X-SHiELD, but we retained the FV3GFS shallow cumulus convection because it improved cloud cover in the global simulations. The FV3 dynamical core employed an updated scalar advection scheme and included microphysical adjustments during each of seven sub-steps taken within the 225-second time step used for the parameterized physics. The reference simulation was initialized at 0 UTC on 1 Aug. 2016 from GFS operational analysis interpolated to the C3072 native grid, and ran through Sept. 9, 2016 on 13824 cores of NOAA's GAEA computing system.

One important change we made to the DYAMOND implementation was to lightly nudge the fields of temperature T, horizontal wind components u, v, and surface pressure p_s, but not humidity, toward GFS reanalysis with a 24-hour timescale. We call this 'meteorological nudging', following Zhou et al. (2021); it should not be confused with other applications of nudging for ML training in this paper. It kept the large-scale meteorology of X-SHiELD very similar to the reanalysis, with a 99.5% correlation between the simulated and reanalyzed patterns of 500 hPa height, while allowing meaningful comparison of the humidity, cloud and precipitation fields with observations. It had reassuringly little impact on the 40-day time-mean cloud and radiation biases of X-SHiELD compared to the original free-running DYAMOND simulations, with global-mean TOA outgoing longwave and shortwave fluxes matching within 1 W/m^2.

A second change was that many internal variables of X-SHiELD were 'coarsened' in-line to a C384 79-level grid of approximately 25 km horizontal resolution, as described in the next section, and output on this grid every 15 minutes. The coarsened output was exported to Google Cloud Storage for use in a custom ML workflow written in Python, described below.

2.2 Coarse-grid model, Python wrapper and cloud-based ML workflow

As in WM21, our goal is to use ML to improve a coarse-grid version of FV3GFS, with the same 79 vertical levels as X-SHiELD. The results presented here use a C48 horizontal grid with approximately 200 km horizontal resolution, a 15 minute physics timestep, and 6 dynamics substeps per physics timestep. Our version of FV3GFS, described in McGibbon et al. (2021), is built from portions of NOAA's Unified Forecast System (`https://ufscommunity.org`; code repository at `https://doi.org/10.5281/zenodo.4460292`). We disable microphysical updates within the dynamical core of the coarse-grid model to cleanly separate tendencies due to model dynamics and physics. Unlike X-SHiELD, the coarse model uses the FV3GFS deep cumulus and gravity-wave drag parameterizations. Other smaller differences include the choice of scalar advection scheme and version differences in the microphysics and land-surface models.

Because of the wealth of powerful machine-learning packages available in Python, major units of the FV3GFS Fortran code were wrapped in Python (McGibbon et al., 2021). The ML and FV3GFS workflows were executed as containerized steps on Google Cloud Platform, similar to WM21. We have shared the code to do this in a documented, open-source repository (`https://doi.org/10.5281/zenodo.5211066`).

2.3 Is the fine-grid reference more skillful than the coarse-grid model?

The nudge-to-fine approach can only be useful if the fine-grid reference simulation is substantially more skillful than the 'baseline' coarse-grid model with no ML correction. Because the fine simulation is meteorologically nudged, it has good weather forecast skill by construction. However, since its humidity is not nudged, it is still meaningful to compare precipitation from the fine simulation and a similar meteorologically nudged no-ML C48 coarse FV3GFS simulation with observations. Fig. 1 compares Day 3–40 mean precipitation maps from these simulations with observational estimates for the same period from the Global Precipitation Climatology Project or GPCP (Huffman et al., 2001) over land, where precipitation most immediately impacts human and natural systems. The fine reference model has 35% smaller spatial pattern errors of precipitation vs. GPCP than the coarse model, with negligible bias in land-mean precipitation. Although 38 days is a short comparison period, the meteorological nudging makes this comparison meaningful by removing weather variability as a source of uncertainty.

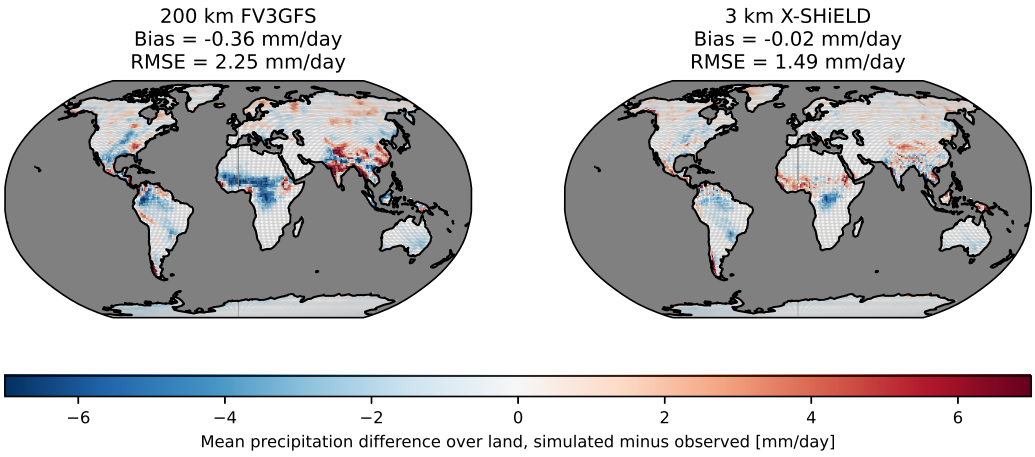

Figure 1. Error in Day 3–40 mean precipitation vs. GPCP of meteorologically nudged (a) FV3GFS with C48 (200 km) horizontal grid resolution, and (b) X-SHiELD with C3072 (3 km) horizontal resolution. In the plot titles, RMSE is the root mean squared spatial pattern error and bias is the time-mean error averaged over all land.

2.4 Pressure-level (p) coarsening

We coarsened the C3072 simulation outputs to C48 grid columns as follows. To conserve mass, the hydrostatic pressure thickness of each coarse grid layer was calculated as an area-average over fine grid cells with the same layer index. Surface and other two-dimensional fields were coarsened using area-weighted averaging. Three-dimensional atmospheric scalar fields were pressure-level (p) coarsened using area-weighted averaging of C3072 data vertically interpolated to the pressure levels of the parent C48 grid cell. In mountainous regions, this average was only over those C3072 grid columns for which the C48 pressure level in question was above ground. On a D-grid, the tangential horizontal velocity component are specified at the centers of each grid cell edge. Thus, tangential wind components are coarsened by pressure-level averaging over the fine-grid cell edges comprising each coarse-grid cell edge.

The attraction of p-coarsening is clearest in the special case of an atmosphere at rest over orography. For this case, neglecting virtual effects, temperature must be purely a function of pressure. Since p-coarsening preserves this relationship, it efficiently capture the stratified vertical structure of such an atmosphere, while coarsening along terrain-following model levels would average temperature across a range of heights. Similarly,

188 p-coarsening preserves thermal wind balance above orography, unlike model-level coars-
189 ening.

190 As described in Appendix A1, area-weighted coarsening of hydrostatic pressure lev-
191 els requires a correction to the area-weighted surface elevation, which is almost always
192 negative and depends on the local topographic variability. Over the Andes and Himalayas,
193 it locally exceeds -150 m (Fig. A1). We calculate this topographic correction at the ini-
194 tial time of a coarse simulation and leave it fixed thereafter.

195 3 Methodology for nudged training simulation

196 Our methodology for constructing an ML training data set evolved to overcome
197 early shortcomings. Key advances were the use of nudging (Secs. 3.1-3.3) and the use
198 of fine-model downwelling radiation and precipitation to force the land surface (Sec. 3.4).

199 3.1 Failure of a tendency-difference method as motivation for nudging

200 We first tried a 'tendency difference' method (Brenowitz & Bretherton, 2019). The
201 coarse model was initialized with a p-coarsened state from the fine model and integrated
202 over one 15-minute physics timestep to determine average tendencies of four prognos-
203 tic variables T, q, u, v at each grid point. Fluctuations in these 'memory variables' per-
204 sist over many time steps so they are important to accurately forecast. The coarse-model
205 tendencies were subtracted from the p-coarsened fine-grid tendencies averaged over the
206 same period to get fields of tendency differences which we hoped to use as ML targets.
207 Unfortunately, the coarse model immediately spun up strong vertical velocities around
208 orography that contaminated these tendency differences. The excess vertical velocities
209 took about three hours to damp out. This was a form of initialization shock, a challenge
210 noted and addressed since the early days of numerical weather prediction (Daley, 1981).

211 As in WM21, we have circumvented initialization shock by smoothly nudging the
212 coarse training model with an appropriate time scale τ so it stays close to the evolving
213 coarsened fine model output, but remains near a dynamically adjusted state of the free-
214 running coarse model. We choose $\tau = 3$ hours, the time vertical velocity variance takes
215 to equilibrate. WM21 chose $\tau = 6$ hours for their similar approach of nudging a coarse
216 model to a reanalysis, because that was the frequency of the available GFS analysis data.

We will show that both choices perform similarly well. Our shorter τ has the conceptual advantage of keeping the nudged atmospheric state closer to the coarsened-fine state.

3.2 Mathematical formalism for nudging approach

Our mathematical notation for the nudging approach is summarized in Table 1. We consider an arbitrary scalar field a (e.g. T, q) that has been prognosed by the fine reference model on the fine grid ($a^f(x^f, y^f, p^f, t)$). We p-coarsen a^f to the field $\overline{a}(x^c, y^c, p^c, t)$ given on the coarse grid. We initialize the coarse model with the coarsened-fine output from some time t_0:

$$a^n(x, y, p, t_0) = \overline{a}(x, y, p, t_0).$$

We run the coarse model, nudging its prognosed fields (denoted by a superscript n) toward their reference values:

$$\frac{\partial a^n}{\partial t} = -\nabla \cdot (\mathbf{v}^n a^n) + Q_a^p + \Delta Q_a, \tag{1}$$

Here $Q_a^p(x, y, p, t)$ is the tendency of a due to the physical parameterizations of the coarse model. Analogous equations with additional pressure-gradient and Coriolis terms are used for the eastward and northward velocity components u and v.

The nudging tendency

$$\Delta Q_a = -\frac{a^n - \overline{a}}{\tau}. \tag{2}$$

corrects the coarse training simulation to evolve similarly to the coarsened fine simulation.

Eq. 2 also shows that the nudged coarse atmospheric state differs slightly from the reference coarsened fine-grid state in proportion to the nudging tendency and the nudging timescale. Thus plots of nudging tendency also translate (by relabeling the color bar) into plots of the nudged state difference.

The primary ML targets are the time-dependent nudging tendencies of the memory variables in all coarse-model grid points sampled over a sufficiently long nudged simulation. Since the nudged run is initialized from coarsened fine model output, the first few hours of the nudged simulation suffer from initialization shock and should not be used for ML training. Instantaneous nudging tendency fields look very noisy where the coarse model state in each grid column varies strongly from time step to time step, e.g. due to episodic cumulus convection. If the nudging is applied with a relaxation timescale τ, it

Notation	Description
$a(x, y, p, t)$	Arbitrary scalar field.
$a^f(x^f, y^f, p^f, t)$	a represented on the fine grid
$a^c(x^c, y^c, p^c, t)$	a represented on the coarse grid (x^c, y^c implied for a single grid column)
$\overline{a}(x^c, y^c, p^c, t)$	Area-weighted pressure-level average of $a^f(x^f, y^f, p^c, t)$ across a coarse grid cell
$Q_a^p(x^c, y^c, p^c, t)$	Source of a from coarse-model physics parameterizations
$\overline{Q}_a(x^c, y^c, p^c, t)$	Apparent source of a from fine reference simulation on coarse grid
$\mathbf{V}(x, y, p, t)$	Velocity vector with components $u = Dx/Dt, v = Dy/Dt, \omega = Dp/Dt$

Table 1. Coarsening notation

suffices to use the nudging tendency averaged over the timescale τ (3 hours by default) as our ML target.

We treat the nudging tendency as a correction to the parameterized physics. Physical processes represented in GCMs, like radiative transfer, boundary-layer turbulence, cloud microphysics, or cumulus convection, can be approximated as operating within individual coarse grid columns, as long as the grid spacing is larger than the O(10 km) depth of the troposphere. Thus, GCM parameterizations of these processes are generally formulated column-wise, with tendencies that only depend on the atmospheric state in the corresponding column. In this work, as in WM21, we make the same assumption for machine-learning the nudging tendencies, although we will also note its flaws.

3.3 Nudging, training, and coarse-grid forecast periods

The nudged coarse simulation has a C48 (200 km) horizontal grid spacing. It is started at 01 UTC on 1 August 2016, one hour into the 40-day fine-grid simulation, when it has all necessary coarsened fine-grid initialization data. It extends to the end of the fine-grid simulation.

The first four days of the fine-grid reference and nudged coarse simulations are treated as a spin-up period. We use Days 5–40 of the nudged coarse simulation to generate a 36-day dataset of field values $a^n(x, y, p, t)$ and nudging tendencies $\Delta Q_a(x, y, p, t)$ that is sam-

pled for ML training and testing. Time-means from the fine and nudged coarse datasets are computed over this 36-day period.

Free-running prognostic (forecast) runs are by default initialized from the coarsened fine output at the beginning of Day 5 (0 UTC 5 August 2016) and are run for 36 days. The fine-grid simulation is used as reference 'truth' to measure their forecast skill. Time-means are computed using their last 30 days to allow for forecast spin-up. A set of four shorter ten-day simulations initialized on Days 5, 13, 21 and 29 are used to assess weather forecast error. The skill of ML-corrected model versions is compared against an identically initialized baseline version of the coarse model with no added ML.

3.4 Forcing the land surface in a nudged training run

We specify sea surface temperature (SST) but the land model is interactively coupled to the atmosphere. For this paper, our ML approach is to correct the atmosphere but not the land model or the ocean surface flux algorithm. Our nudged coarse training run supports this approach by forcing the land surface consistently with the fine reference model.

The land model is forced by the atmosphere through downwelling shortwave and longwave radiation, precipitation, and lowest-level wind, humidity and temperature (which affect the turbulent heat and moisture fluxes from land to atmosphere). In the training run, the lowest-level quantities are already nudged toward coarsened fine model predictions. The downwelling fluxes and precipitation diagnosed by the physical parameterizations of the nudged coarse model are typically biased relative to the fine-grid reference model. These biases are large for our case, due to the nudged coarse model generating much less cloud than the fine-grid model. The global-mean fine-grid surface downwelling shortwave radiation flux is 33 W/m^2 less than the coarse-grid model. This is partly compensated by a downwelling longwave flux increase of 11 W/m^2, to give a net of -22 W/m^2. Fig. 2 shows a time-mean map of this quantity, showing the bias is largest in land and ocean regions with high insolation and extensive high cloud.

To minimize land surface drift in the nudged training run, we therefore force the surface with the coarsened downwelling radiation and precipitation from the fine-grid model. In our simulations, this has no impact over the ocean because the surface forcings do not feed back on SST. The right panels of Fig. 3 show that this keeps the time-

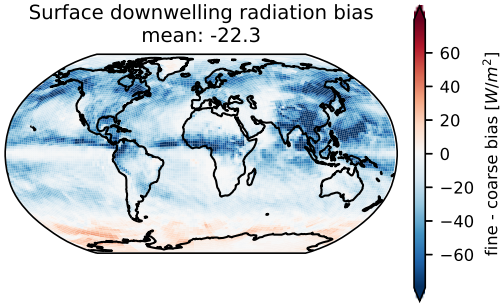

Figure 2. Time-mean difference between the fine-grid and nudged coarse total (shortwave plus longwave) downward radiative flux at the surface.

average land surface latent heat flux (LHF) and sensible heat flux (SHF) desirably close to their reference values. The left panels of Fig. 3 show the large land-surface surface flux biases that develop in the training run if this is not done. As expected, the biases of time-average LHF and SHF over the oceans are insensitive to this change. The LHF over warm oceans is typically somewhat smaller in the nudged training run than in the fine-grid model.

4 Machine-learning methods

Our ML schemes are trained 'offline' (without considering their feedback on other parts of the climate model), because we can take advantage of efficient methods for doing that. They are then applied 'online', for which those feedbacks become important and can lead to climate drifts or model instability. Ultimately, online performance must be the primary metric for evaluating our ML schemes; our hope is this is founded on good offline skill.

We use random forests (RFs) and neural nets (NNs) to learn the three-hour average nudging tendency profiles and the fine-grid surface downward radiative fluxes. Each has its own strengths. RFs do not extrapolate outside their training range, an advantage for prognostic simulations in which climate drifts and extreme events inevitably create out-of-range samples. Prognostic simulations with an RF used for the ML correction may experience significant climate drifts, but generally remain stable until those drifts are already unacceptably large. NNs have many architectural variants that can help op-

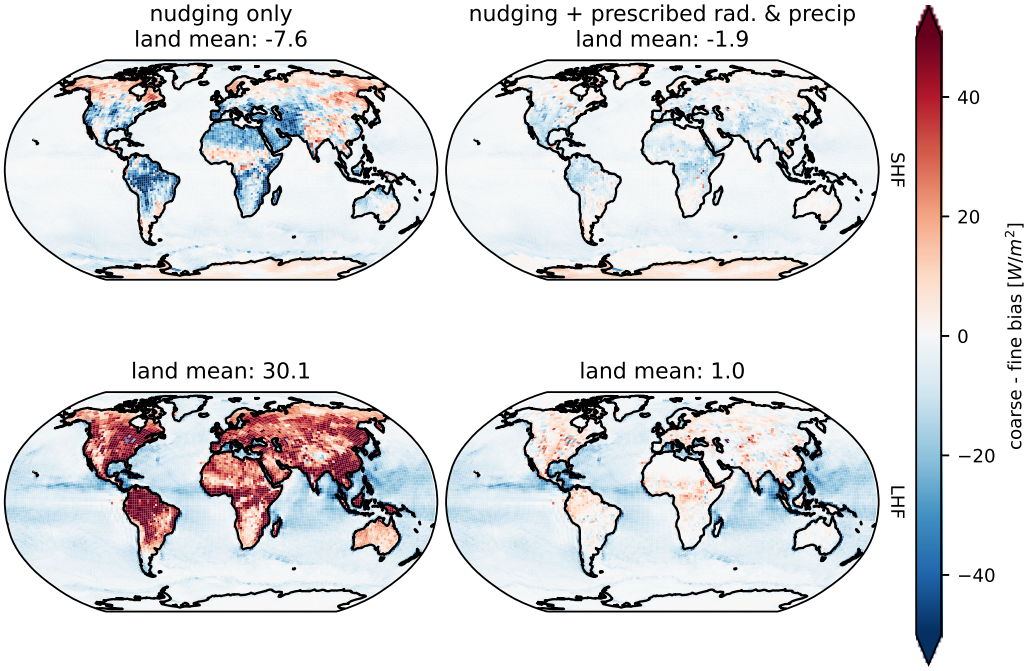

Figure 3. Map of time-mean biases of SHF (top) and LHF (bottom) from the C48 nudged training run relative to the fine-grid reference. Left panels: Surface forced by coarse-model physical parameterizations. Right panels: Surface forced by fine-grid radiation and precipitation.

timize their skill, and they can run more efficiently than RFs on accelerator-based computing architectures.

To ensure moisture conservation, in prognostic runs we use an atmospheric column humidity budget to infer surface precipitation (see Sec. 6.1).

4.1 ML methodology

We store the model state and the averaged nudging tendencies from the nudged coarse run every three hours. Our primary ML scheme predicts vertical profiles of 3-hour averaged nudging tendencies in each GCM grid column. Our training sample comprises the global fields of T, q and (where noted) u, v nudging tendencies at 130 randomly-selected times from Days 5–31 of the nudged coarse simulation. This temporal sub-sampling is needed for efficient training of random forests. For C48 grid resolution, there are 13824 grid columns over the globe and hence ∼1.8M atmospheric columns used for training.

The diagnostic testing is on another sub-sample of 50 times during the last 9 days (Sept. 1–9).

We machine-learn the nudging tendencies $\Delta Q_{T,q}$ and (where noted) $\Delta Q_{u,v}$ as functions of the column state from the nudged coarse run, defined as the 79-level column profiles of T, q and (where noted) u, v, plus the cosine of solar zenith angle (needed for radiation), and the surface geopotential (an indicator of mountains). For RFs we also use the land-sea-ice mask (0=sea, 1=land, 2=sea ice), but this categorical variable is not a suitable feature for NNs. The learned approximations are denoted with a superscript ML, e.g. ΔQ_a^{ML}.

The loss function is a sum of normalized mean squared errors (for RFs) or mean absolute errors (for NNs) in the target nudging tendency profiles. For each nudging tendency, the loss at each vertical level is normalized by dividing the prediction error by the standard deviation of the target nudging tendency at that level, such that all levels are weighted roughly equally in the total loss.

A secondary ML scheme is trained to predict fine-grid surface downwelling shortwave and longwave radiative fluxes and to deduce the surface net shortwave flux, which is also a required input for forcing the FV3GFS land surface model. This is needed to correct the large surface radiation biases of the coarse model that feed back on LHF and SHF over land. This scheme uses the same features as the tendency ML, not including wind profiles, and the same set of test and training samples. RFs are straightforward to train for this purpose; NNs are trained with a positivity constraint to avoid model crashes driven by negative ML-predicted fluxes.

The ML approaches are named using a string of letters describing the learned nudging tendencies (plus downwelling radiation, if learned) followed after a hyphen by the type of ML. For instance, for $TquvR$-RF, a random forest is used to learn the T, q, u, v nudging tendencies and a second random forest is used to learn the downwelling radiation R.

4.2 RF configuration

The RF for nudging tendencies is implemented in `scikit-learn`. It uses 13 trees of maximum depth 13. The Tq option learns the T and q nudging tendencies from the

T and q profiles. The $Tquv$ option learns all four nudging tendencies from all four profiles. A RF for longwave and shortwave downwelling radiation is similarly configured using a mean squared error loss between the prediction and the fine-grid targets.

4.3 NN configuration

Neural nets are implemented in `keras`. We use fully-connected NNs with two hidden layers and learning rate 2×10^{-3}. To achieve 40-day stable prognostic NN-corrected simulations, we made the following changes to the RF training protocol:

- Mean absolute error is used in the loss functions to reduce sensitivity to outliers.
- If wind corrections are used ($Tquv$ option), train separate NNs for $\Delta Q_{u,v}$ and for $\Delta Q_{T,q}$ instead of predicting all four tendencies in a single model. Customizing the input feature set and hyperparameters for the separate models enables better online stability. The $\Delta Q_{T,q}$ model has a width of 128 while the $\Delta Q_{u,v}$ model has a width of 32; these widths were selected for good offline performance.
- Include a rectification layer in the training and output that prevents negative surface downwelling radiative fluxes.
- Regularize the NNs using a L2 coefficient $\gamma = 10^{-4}$ for T, q, and $\gamma = 10^{-2}$ for u, v to achieve online stability and smooth dependence of outputs on input profiles.
- Train four different NNs with different random seeds in stochastic gradient descent, and use the median prediction of the four.

5 ML target characteristics and offline ML skill

Our ML targets are time-dependent nudging tendency profiles and downwelling surface radiative fluxes from around the world. This section discusses some salient characteristics of these targets and the offline skill of RFs and NNs in learning them. Five major points are:

1. Three-hour nudging keeps the nudged coarse model state very similar to the coarsened-fine reference.
2. Instantaneous nudging tendency profiles can have complex vertical structure that varies greatly in space and time and challenges ML skill.

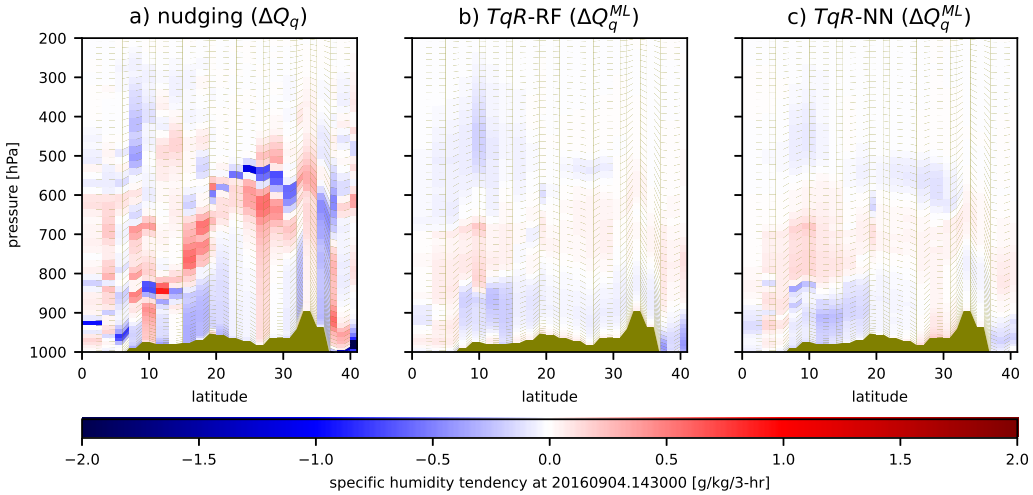

Figure 4. South-north cross-sections along 0°E (a) humidity nudging tendency from the C48 nudged training run, compared to (b) RF and (c) NN.

3. Time-mean nudging tendencies are large and well-learned by ML.
4. Fine-grid surface radiative fluxes are skillfully learned.
5. RFs and NNs have comparable offline skill.

5.1 3-hour mean humidity nudging tendency cross-section

Fig. 4 illustrates challenges in ML of nudging tendencies, Fig. 4a shows the 3-hour mean humidity nudging tendency ΔQ_q along a south-north vertical section through west Africa along 0°E, on the afternoon of Day 36 of the nudged training run, a time in our ML test sample. The legend is given in units of g/kg per 3 hours. Thus, with sign reversed, it corresponds to the difference between the nudged coarse and fine humidity, which is seen to have a typical magnitude less than 1 g/kg.

The humidity nudging tendency has a complex spatial structure, with both sharp and diffuse vertical structures at a range of pressures, presenting a challenging data set for machine learning. Figs. 4b-c show the corresponding RF and NN learned cross-sections at this time in our test sample. They are similar to each other. They both qualitatively resemble Fig. 4a but with much lower amplitude.

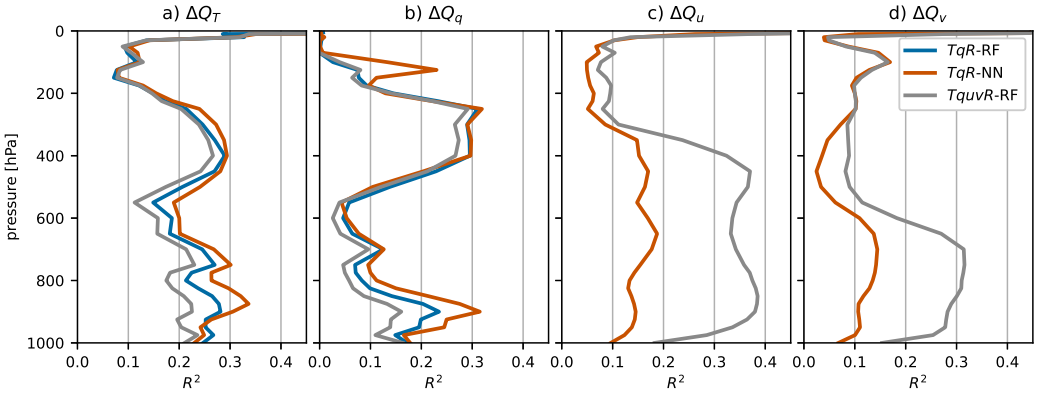

Figure 5. Vertical profiles of R^2 of the nudging tendencies over the test times for offline TqR and $TquvR$ RF and NN configurations. For TqR configurations, only $\Delta Q_{T,q}$ are predicted. The $TquvR$-NN configuration predicts the same $\Delta Q_{T,q}$ (and has the same R^2) as does TqR-NN.

5.2 ML skill for 3-hr nudging tendencies

To evaluate offline skill in learning the vertical profile of each nudging tendency, we use area-weighted fraction of variance (R^2) at each pressure level p, taken over the test data (Appendix A2 gives a mathematical specification). R^2 measures the skill improvement (or degradation if $R^2 < 0$) of a prediction over a trivial default, in this case the global mean of the nudging tendency over the test data at pressure level p.

The cross-section example hints that the RF and NN have low and comparable skill in predicting 3-hr nudging tendencies. Fig. 5 shows the profiles of R^2 vs. pressure for the four nudging tendencies over the test times. For all variables, R^2 is modest, varying between 0.1-0.3 depending on pressure. That is, neither type of ML is able to learn the bulk of the space-time variability of the nudging tendencies based on single-column features. For temperature and humidity, the NN has a somewhat higher R^2 at all pressures. For winds, the NN has much smaller offline skill, due to applying heavy regularization to avoid prognostic instabilities.

The R^2 profiles for ΔQ_T and ΔQ_q are slightly larger for the NN than for the RF, especially in the lower troposphere. For the RF, the results shown are for the TqR case that only T and q are predicted. If all four nudging tendencies are simultaneously predicted, then R^2 is slightly degraded at all levels.

5.3 Time-mean ML skill for humidity nudging tendency

To avoid systematically forcing climate biases, the ML corrective tendencies should be approximately unbiased in time-mean relative to the actual nudging tendencies. This can be checked using column-integrated maps and globally-integrated profiles of the humidity nudging tendency.

The column-integrated humidity nudging tendency $\langle \Delta Q_q \rangle$ measures how much moisture must be supplied to each coarse-model grid column to match the evolution of the fine run. It reflects fine-coarse differences in precipitation, surface evaporation and horizontal moisture convergence. Figure 6a is a map of the column humidity nudging tendency $\langle \Delta Q_q \rangle$ averaged over the 50 test samples. This map is a specklier version of the full Day 5–40 time-mean shown later in Fig. 14a. It also looks very similar to Fig. 2c of WM21, who nudged to an observational analysis rather than a fine-grid model. (WM21 referred to ΔQ_q as ΔQ_2; we have changed their notation to avoid potential confusion, since Q_2 is traditionally the apparent drying given in energy units (Yanai et al., 1973).)

Almost everywhere, and especially in regions of strong mean precipitation, $\langle \Delta Q_q \rangle < 0$, i.e. the fine-grid reference simulation is drying more (has a larger excess of precipitation over evaporation) than the nudged coarse simulation despite similar thermodynamic states. This indicates a substantial global bias of the FV3GFS parameterized physics toward inhibiting precipitation when applied at C48 grid resolution.

Figs. 6b-c show the offline time-mean column-integrated humidity nudging tendency predicted by the RF and NN. Both ML approaches provide smooth but accurate approximations to the target map in Fig. 6a, with similar spatial pattern RMSEs of 1.0-1.3 mm/d. The TqR-RF approach has negligible global-mean bias, but TqR-NN has a global moistening bias of 0.3 mm/d compared to the target. This is partly an undesirable consequence of regularizing the NN loss function, which particularly affects the humidity nudging tendencies. The NN tends to preserve extrema of the target map better, at the expense of creating spurious features such as a drying maximum over coastal Antarctica south of the Indian Ocean. The ML approximations of other nudging tendencies have qualitatively similar time-mean characteristics, except that the off-line global-mean biases of the NNs are comparable or less than for the RFs.

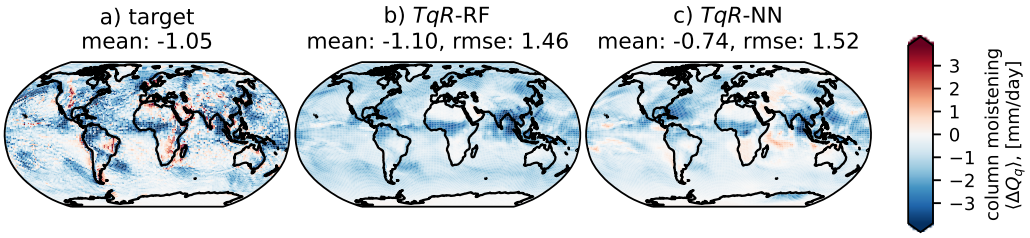

Figure 6. Column humidity (a) nudging tendencies, (b) from offline TqR-RF, and (c) from offline TqR-NN, averaged over the 50 test times from the nudged training run. Spatial pattern RMSEs for the ML methods are with respect to the target nudging tendencies.

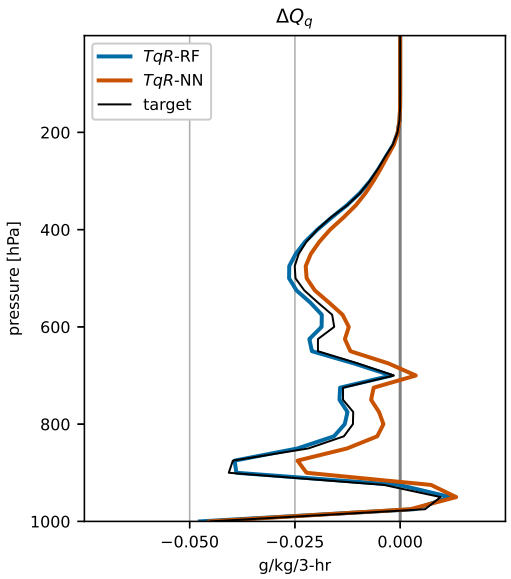

Figure 7. Global mean vertical profile of humidity nudging tendency averaged over the test times, and the offline TqR-RF and TqR-NN approximations to it.

Another useful off-line ML bias measure is the global mean vertical profile of nudging tendency averaged over the test times, shown for humidity in Fig. 7. The RF matches the target profile nearly perfectly; the NN has a small but significant positive bias except near the surface.

5.4 Nudging tendencies of other fields

In this section, we document the time-mean nudging tendencies of the other prognostic memory variables, T, u and v. These are all substantial and well replicated off-

line by both RF and NN, without much bias (not shown). Temperature tendencies are expressed in units of heating rate.

Fig. 8a-c shows time-mean column-integrated nudging tendencies for heat ($\Delta Q_1 = c_v \Delta Q_T$), moist static energy ($\Delta Q_m = \Delta Q_1 + L_v \Delta Q_q$), and u. In these formulas, c_v (the specific heat of air at constant volume) replaces the standard c_p (the specific heat of air at constant pressure) because of a FV3GFS peculiarity of the interfaces between the physical parameterization tendency, the nudging tendency and the FV3 dynamical core. L_v is the latent heat of vaporization.

The column heat nudging tendencies $\langle \Delta Q_1 \rangle$ are dominated by latent heating associated with precipitation, so they look nearly like mirror images of the humidity nudging tendencies. Artifacts at the edges of the cubed-sphere tiles are evident over the Southern Ocean. The column nudging tendency of moist static energy (Fig. 8b) is illuminating because it cancels out effects of latent heating and drying to reveal fine minus nudged coarse differences in atmospheric radiative heating and surface latent plus sensible heat flux. Over most ocean locations, it is 25–50 W/m^2. This is due to the fine model having somewhat stronger latent heat flux and less atmospheric radiative cooling (due to more simulated high cloud) than the nudged coarse model over ocean regions (Fig. 3). Over some land regions such as Eurasia, the column moist static energy nudging tendency is negative. Over land, the surface sensible and latent fluxes are similar in the fine and nudged coarse runs, and the radiative heating correction is smaller because there is less high cloud over land, and hence a lesser opportunity for a fine-coarse atmospheric radiative heating difference induced by high cloud biases.

The u-wind nudging tendencies are strongest around major mountain ranges and windy coastlines. Like for humidity, the maps of column heat and zonal-wind nudging tendencies look similar to those shown by WM21. This is expected, because the temperature and winds of the fine-grid reference runs are lightly nudged to reanalysis. Therefore, nudging these coarse model fields to the fine-grid reference is functionally similar to nudging them to a global analysis.

Fig. 8d shows a latitude-pressure cross section of the zonal-time-average v nudging tendency. It shows low-level meridional convergence and upper-level divergence away from 10°N. That is, the fine-grid reference is nudging the meridional winds in the coarse

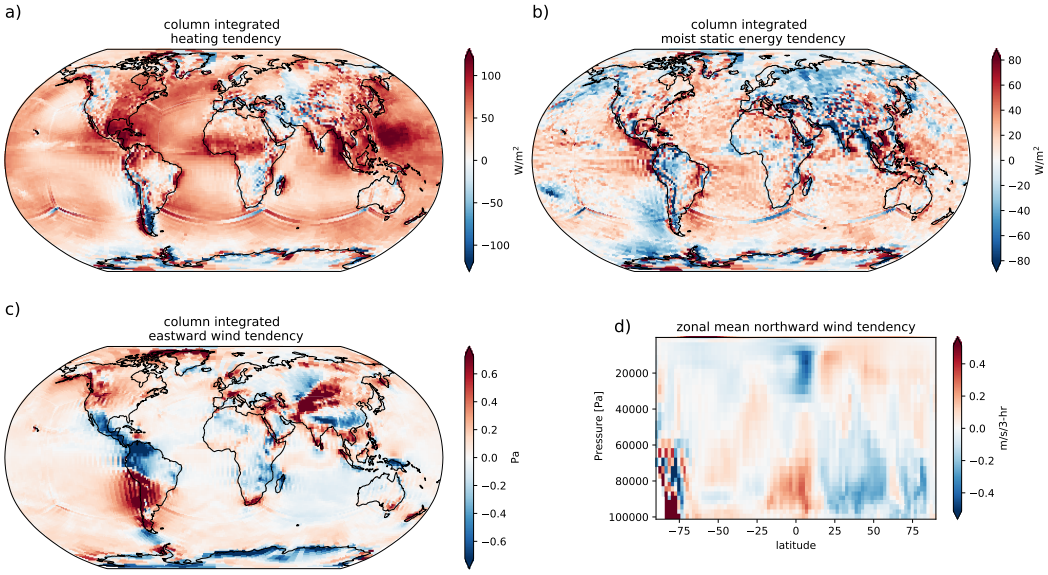

Figure 8. Time-mean nudging tendencies from the C48 nudged training run. (a-c) Column integrals of heating $\langle \Delta Q_1 \rangle$, moist static energy tendency $\langle \Delta Q_m \rangle$, and zonal wind acceleration $\langle \Delta Q_u \rangle$, and (d) zonal-mean latitude-height cross section of meridional wind acceleration ΔQ_v

run (which has less precipitation and latent heating in the Intertropical Convergence Zone) toward a stronger Hadley circulation.

5.5 ML adjustment of surface downwelling radiation

For accurately forcing the land surface in prognostic simulations, we train RFs and NNs to match the fine-grid reference downwelling surface longwave and shortwave radiative fluxes as a function of column thermodynamic state. To add simulation skill, these must match the fine-grid fluxes much more closely than do the coarse-grid parameterized fluxes (whose typical biases are shown in Fig. 2), i.e. within a few W/m^2 or a time-mean relative error of a few percent at each location.

Indeed, both NNs and RFs skillfully predict the time-mean downwelling longwave and shortwave surface radiation. Fig. 9 shows that both methods have comparable small global-time-mean biases of under 2 W/m^2 and low spatial pattern RMSEs in total (longwave plus shortwave) downward radiation over the test data from the nudged training run. Both schemes perform comparably well with a small global-mean bias and similar RMS pattern errors. The NN has a stronger zonal-mean component to the pattern er-

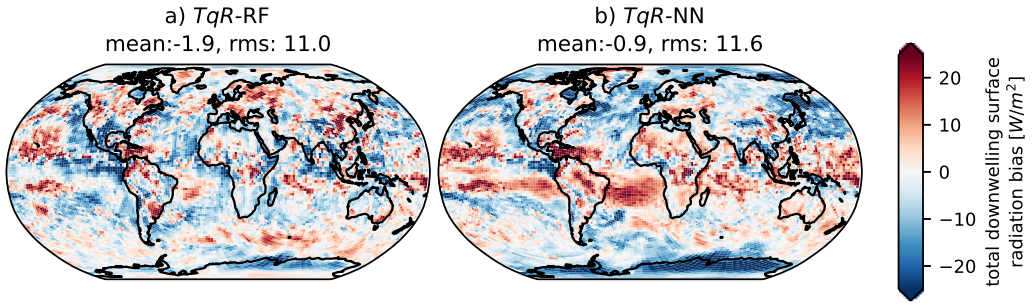

Figure 9. Difference between total surface downward radiative flux from offline ML and fine-grid reference, averaged over 50 test times randomly sampled from the last nine days of nudged training run.

ror, overestimating downwelling shortwave flux over the subtropical South Atlantic and Pacific Oceans and underestimating it over Antarctica.

6 Forecast skill and mean biases

The purpose of our ML is to make prognostic (free-running) weather and climate simulations more closely resemble the reference fine-grid simulation. For prognostic simulations, the nudging tendencies of T and q, and optionally u and v, are replaced by their ML versions (either RFs or NNs), and the ML-predicted downwelling shortwave and longwave radiation are used to force the land surface. In this online application, the corrective ML is fully interactive with the rest of the climate model.

In prognostic simulations, the surface precipitation is calculated from the atmospheric column humidity budget, truncated at a minimum value of zero:

$$\begin{aligned} P &= P^p - \langle \Delta Q_q^{ML} \rangle \\ P^+ &= \max(P, 0) \end{aligned} \tag{3}$$

Here P is the ML-corrected budget-based precipitation, calculated as the physics precipitation plus the column drying from the ML humidity nudging tendency. P may be negative if the ML implies enough column moistening. Enforcing the positivity of surface precipitation creates an artificial source $P^- = P^+ - P = \max(-P, 0)$ of surface precipitation that is not in the atmospheric moisture budget and hence does not have to be balanced by evaporation. In global (and land) mean, this source is small – approximately 0.1 mm/d.

Metric	Units	Base no-ML	Tq RF	TqR RF	$TquvR$ RF	$TquvR$ NN	TqR NN
Z500 RMSE 3–7d fcst	m	64	62	62	**60**	**60**	62
T850 RMSE 3–7d fcst	K	3.1	2.9	2.8	2.9	**2.7**	**2.7**
Prec bias land-time-mean	mm/day	1.1	0.8	0.1	0.4	**0.0**	**0.0**
Prec RMSE time-mean	mm/day	3.7	2.7	**2.5**	2.6	**2.5**	2.6
T200 RMSE time-mean	K	3.4	3.2	**3.1**	5.1	3.9	**3.1**

Table 2. Prognostic weather and climate metrics (details in text) with selected training/ML methodologies. The best results for each metric are bolded. Weather forecast RMSE is based on the average of four initializations; the standard deviation of the mean is about 3 m for 500 hPa height and less than 0.1 K for 850 hPa temperature for all model versions.

Our ideal ML-corrected model would improve weather forecast skill at lead times up to a week or more vs. the baseline, have reduced time-mean biases of key climate metrics such as precipitation patterns and temperature throughout the atmosphere, and run stably for an indefinite period of time from any plausible initial condition given any boundary forcings. However, with our ML approach, we find trade-offs between weather and mean-state skill, especially for upper tropospheric temperature. Prognostic stability also shaped our approach, e.g. in guiding our choice of NN regularization coefficients.

Table 2 shows the sensitivity of some key error metrics to choices of training and ML methodology. It compares the no-ML baseline to RF configurations with just temperature and humidity correction (Tq), added surface radiation correction (TqR), and added wind correction ($TquvR$). It also includes NN versions of the final two configurations.

The first two metrics (500 hPa height and 850 hPa temperature RMSE vs. the fine-grid reference) measure weather forecasting skill. They are based on the average skill over days 3–7 of a set of four 10-day forecasts, initialized from the coarsened fine-grid data on Days 5, 13, 21 and 29 (Fig. 10). The tabulated sample means and the standard deviations of the mean given in the caption for these metrics are estimated from this 4-member set.

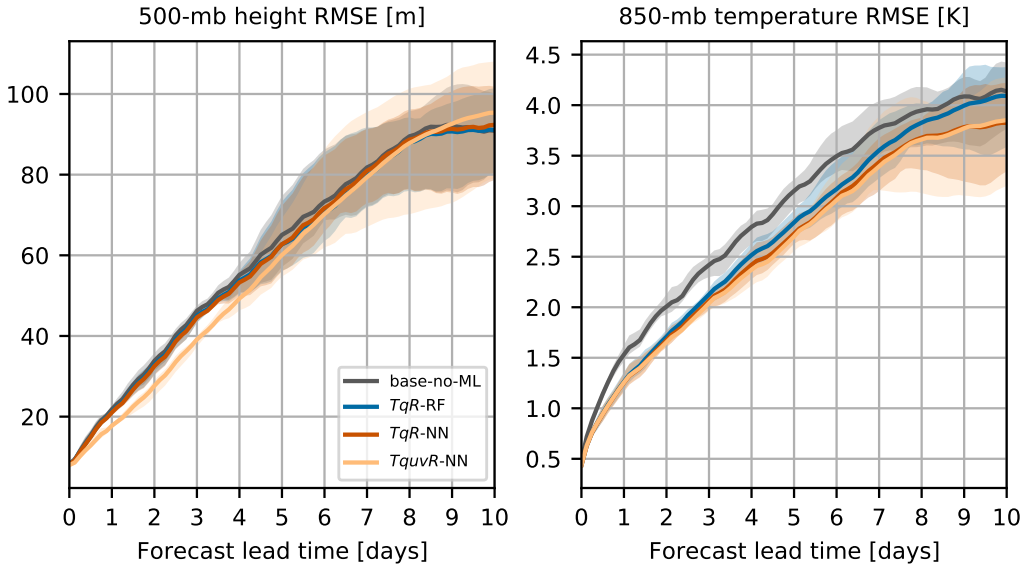

Figure 10. RMSE of 500 hPa height and 850 hPa temperature for baseline and three prognostic ML-corrected configurations in the first 10 days of four forecasts initialized every 8 days. Lines shows the mean, and shading shows the range of results across these forecasts.

For 500 hPa height RMSE, ML correction of temperature and humidity adds marginal skill (3%) over the baseline and wind corrections adds another 3% increment to that. This finding is in the same direction as WM21 found for nudge-to-observations. For 850 hPa temperature RMSE, RF correction of temperature and humidity adds 7% forecast skill, the radiation correction makes a slight additional improvement, while wind correction has little impact. NNs slightly outperform RFs on this metric, though the difference may not be statistically significant.

The remaining three metrics are based on time-mean biases from the last 30 days of 40-day prognostic simulations. The chosen variables are global-mean land-surface precipitation bias, the RMS pattern error of maps of precipitation (see also Fig. 11) and 200 hPa temperature (see also Fig. 13a,d). Two of these metrics focus on precipitation, which was a practical motivation for this work. Note that our ML does not directly target precipitation, so it is not guaranteed to improve these metrics. The third is a measure of upper-tropospheric time-mean skill, which is important to a plausible simulation of the atmospheric general circulation and the movement and intensity of storm systems.

The baseline global-mean land-surface precipitation bias is reduced to zero for both NN configuration and nearly to zero by the TqR-RF configuration. This drastic improvement over the baseline is primarily due to the ML radiation correction, with help from the ML corrections to temperature and humidity tendencies. The RF wind tendency correction slightly worsens this bias. Daily time series of this quantity show the baseline configuration has a large precipitation spin-up as it moistens the atmosphere over the first ten days, while the ML-corrected simulations exhibit much less spin-up. Thereafter, all simulations have daily fluctuations of up to ± 0.2 mm/d with little further drift. We infer that the values of time-mean global-land-mean surface precipitation in Table 2 have less than ± 0.1 mm/d of random uncertainty, so their differences are meaningful.

The RMSE of the time-mean precipitation pattern is reduced nearly 30% from the baseline by inclusion of the ML temperature and humidity tendencies, with an additional 3% improvement from the radiation correction, and no consistent impact from the ML wind tendencies.

The final row in Table 2 shows the time-mean pattern RMSE in 200 hPa temperature. This proved decisive in our choice of optimal ML configuration. ML correction of temperature and humidity tendencies slightly reduced the baseline RMSE. Addition of the ML surface radiation correction decreased this RMSE slightly more. The ML wind tendency correction increased the time-mean 200 hPa temperature errors substantially, mostly at high latitudes (Fig. 13d). Yuval and O'Gorman (2021), using a coarsening approach on an aquaplanet, also found that ML correction of subgrid vertical momentum fluxes had good offline skill but led to time-mean upper-tropospheric zonal wind drifts in prognostic simulations.

Overall, this led us to select the neural net ML architecture TqR-NN as the optimal choice. This configuration learns temperature and humidity nudging tendencies, but not wind tendencies, and includes a learned surface radiation correction. Like its random forest analogue, it improves on the baseline no-ML configuration in all five metrics, and it has smaller errors than the RF in time-mean 200 hPa temperature and land-surface precipitation. Except for the 200 hPa temperature, the NN and RF configurations that also include ML wind tendency correction increase 3–7 day forecast skill and are also competitive for time-mean biases, as found by WM21 for the RF configuration in the related nudge-to-observations application.

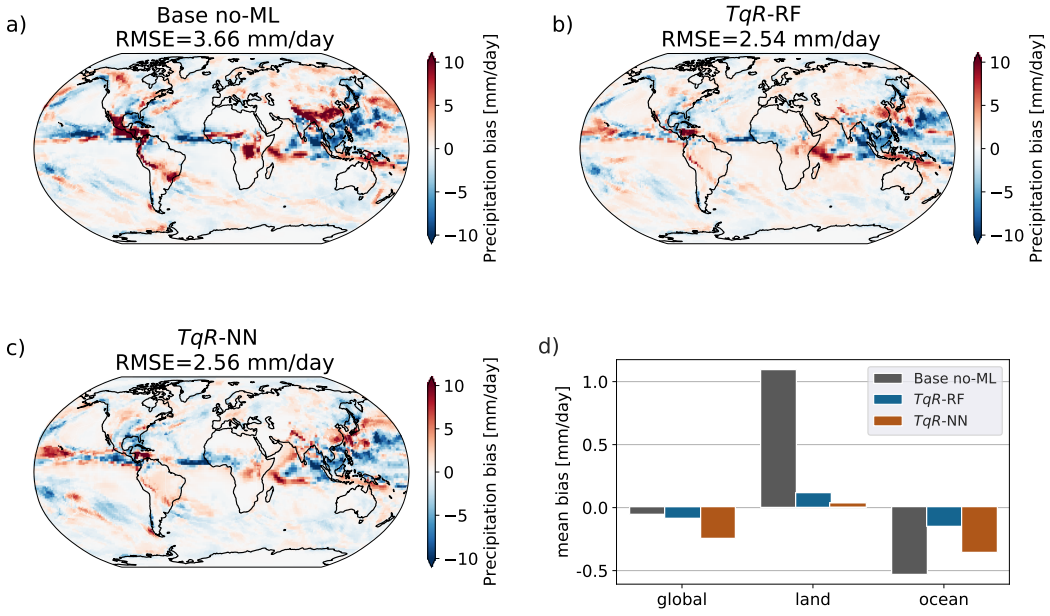

Figure 11. Maps of 30-day time-mean precipitation pattern difference from the fine-grid reference for prognostic simulations: (a) C48 baseline (b) *TqR*-RF, (c) *TqR*-NN; and (d) bar charts of the land-mean, ocean-mean and global-mean precipitation biases for these three configurations.

6.1 ML corrections reduce time-mean precipitation bias

Fig. 11a-c shows maps of the time-mean precipitation differences from the fine-grid reference ('pattern errors') of the C48 baseline without any ML correction, with RF-based corrective tendencies and surface downwelling radiation, and with NN-based corrective tendencies and surface downwelling radiation. RMSE is shown on these plots as an overall global measure of pattern error. Both ML configurations reduce the precipitation RMSE by 30%, an even more substantial reduction than achieved by WM21 using the nudge-to-observations method. As found by WM21, the biggest reductions in precipitation error are over the Himalaya, Andes, and central American mountains, but the precipitation errors are reduced over most land regions. We attribute the additional improvement mainly to our less biased radiative forcing of the land surface, which largely removes small but widespread wet biases over arid subtropical land regions (e.g. the Sahara Desert) found by WM21.

Fig. 11d compares global-time-mean, land-time-mean, and ocean-time-mean precipitation biases (vs. the fine-grid reference) for the three configurations shown in Figs. 11a-c. The fine-grid reference has a 30-day average land surface precipitation of 2.3 mm/d.

The coarse-grid baseline has a 3.4 mm/d average, i.e. a 1.1 mm/d high bias over the reference. Both the RF- and NN-based corrections largely remove this land surface precipitation bias by shifting precipitation from land to ocean.

6.1.1 Diurnal cycle of land surface precipitation

Fig. 12 shows the time-mean diurnal cycle of land precipitation for the TqR-RF and TqR-NN model configurations, based on hourly-mean outputs binned by local solar time. The fine-grid reference has a pronounced diurnal cycle peaked in the late afternoon. This is a long-standing challenge for conventionally-parameterized global climate models (Christopoulos & Schneider, 2021). Indeed, the diurnal cycle of the C48 baseline simulation is rather irregular, with realistic timing but only half as large a land surface precipitation (as measured by its first Fourier harmonic). The NN and RF realistically increase the diurnal cycle amplitude but undesirably shift the timing of maximum precipitation three hours earlier in the day; this result is unaffected by including wind correction (e.g. $TquvR$-NN). This is still an improvement over typical conventionally-parameterized global climate models, which on average have the diurnal rainfall peak over tropical land nearly six hours too early (Christopoulos & Schneider, 2021).

6.2 Other systematic biases of the prognostic runs

Our current version of the nudge-to-fine method does not automatically prevent mean-state drifts of global means or spatial patterns in ML-corrected prognostic runs. Fig. 13a-c compares time series of some global-mean variables in TqR-RF, $TquvR$-NN and $TquvR$-NN prognostic runs with the fine-grid reference and the baseline. This provides a more holistic view of time-mean bias development throughout the troposphere than the metrics discussed so far. Overall, the TqR NN and RF configurations keep mean-state drifts of these variables smaller or comparable to the baseline configuration.

RF-corrected runs are insensitive to different random RF realizations, so just one curve is shown. The NNs are more sensitive to their random seed. The color shadings show the range of results from using the NNs from the four individual random seeds. Ideally, the fine-grid reference would lie within the shaded regions.

Drifts of global-mean 200hPa air temperature (Fig. 13a) vary significantly among the different baseline and ML-corrected runs. TqR-NN best matches the fine-grid ref-

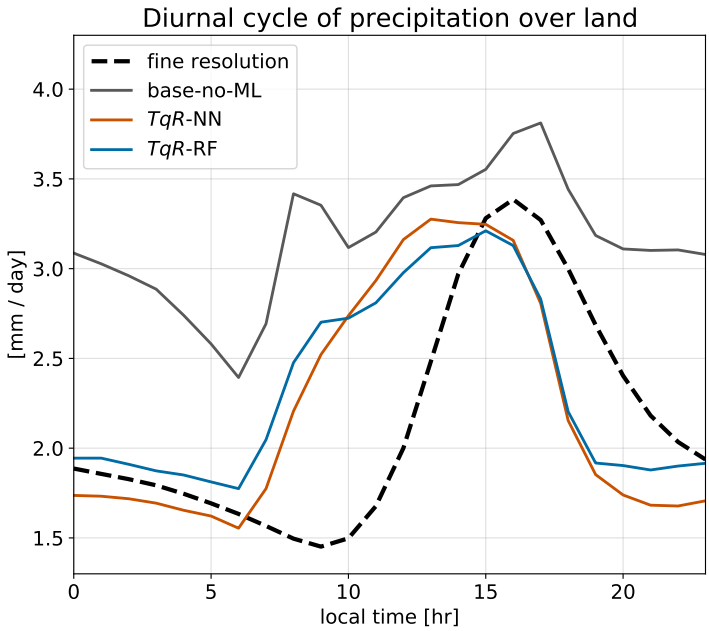

Figure 12. Maps of 40-day land-mean precipitation diurnal cycle from the fine-grid reference for the fine-grid reference, C48 baseline, TqR-RF and TqR-NN prognostic simulations.

erence; TqR-RF drifts slightly warm. $TquvR$-NN drifts cold at a rate comparable to the baseline model. WM21 reported drifts of comparable amplitude during the first month of year-long simulations using the nudge-to-observations method; those drifts stopped growing thereafter.

Fig. 13d shows the 20–40 day zonal-mean 200 hPa temperature, after the global-mean drifts have nearly fully developed. All simulations have little bias in the tropics, but in the north polar region, the wind-corrected run ($TquvR$-NN) has developed a cold bias exceeding 10°K, much larger than for the baseline and other ML configurations.

The global-mean 850 hPa temperature (Fig. 13b) from both the TqR-NN and $TquvR$-NN prognostic runs drifts less from the fine-grid reference than does either the baseline or the RF-corrected run. For the global-mean precipitable water (Fig. 13c), all ML-corrected runs drift less than the baseline (which becomes significantly too moist). The drifts of the two NN simulations are comparable to the RF but of opposite sign.

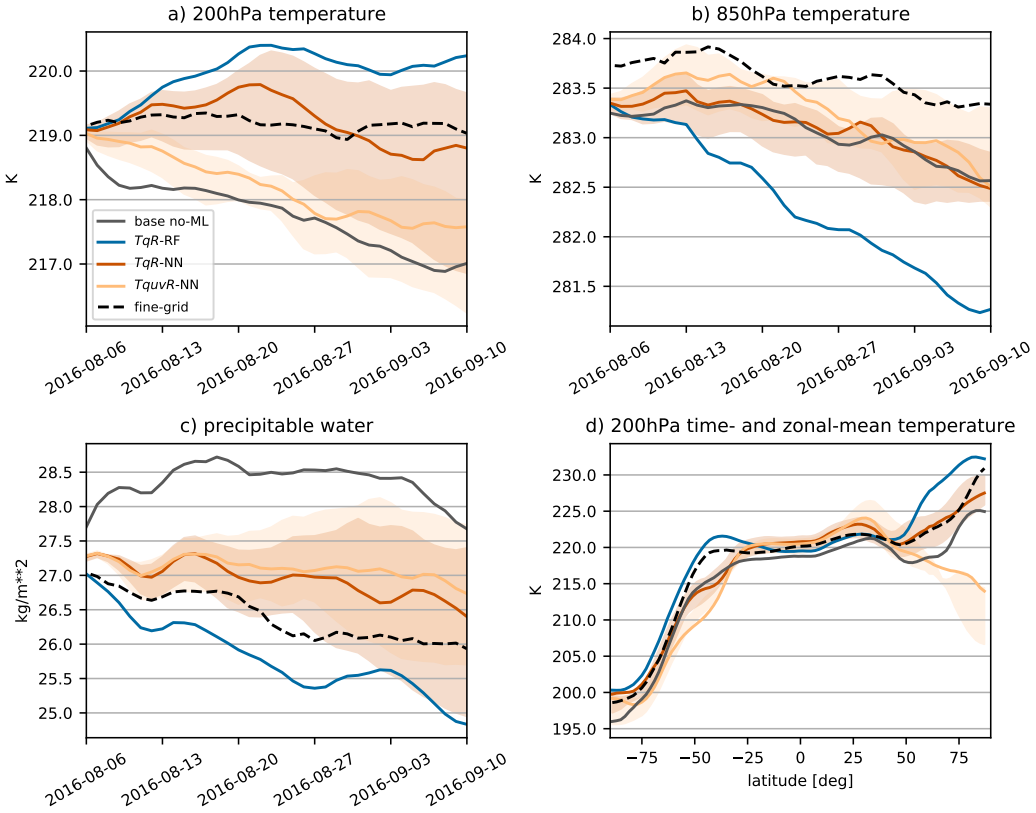

Figure 13. Time series of global daily-mean (a) 200 hPa temperature; (b) 850 hPa temperature, and (c) precipitable water. (d) 20–40 day time-zonal-mean of 200 hPa temperature for three prognostic ML configurations, the baseline coarse simulation and the fine-grid reference. For TqR-NN and $TquvR$-NN, 4 NNs each were trained from different random seeds. The solid line comes from using their ensemble-mean at each time step (as shown in other plots). The shading indicates the range of predictions from prognostic runs using each NN individually.

6.3 Optimality of 3-hour nudging timescale

Appendix A3 discusses the sensitivity of prognostic simulations with and without wind nudging to a range of choices of nudging time scale τ from 1–12 hours. Precipitation biases are not highly sensitive to τ. For the preferred TqR-NN configuration, choosing $\tau = 3$ hours minimizes zonal-time-mean upper-tropospheric temperature biases, narrowly besting $\tau = 6$ hours. If wind nudging is also used, the longest τ, 12 hours, minimizes these biases, but they are still larger than for the TqR-NN configuration with 3 hour nudging.

7 Discussion: Nudging tendencies and physical parameterization correction

Our nudge-to-fine ML approach has treated the nudging tendencies as a correction to the model physical parameterizations, predicted in each grid column using the thermodynamic profiles and horizontal winds within that grid column. We argue that this ML assumption is far from perfect but is good enough to be useful.

Formally, one can decompose the nudging tendency field ΔQ_a of a scalar $a(x, y, p, t)$ into contributions from fine-coarse differences in 'physics' and 'dynamics' (Appendix A4). The decomposition is approximate above orography. Here, the physics component ΔQ_a^p is the fine-coarse difference in the apparent source of a, and the dynamics component ΔQ_a^d is due to the difference of the advection of a between the coarsened fine simulation and the nudged coarse simulation. We can compute ΔQ_a^p directly in each coarse grid column (see Appendix A4) and estimate ΔQ_a^d as a residual $\Delta Q_a - \Delta Q_a^p$.

Since weather and climate respond most strongly to systematic forcing, we compare how similar time-mean nudging tendencies look to their physics components. We use humidity for illustration. Fig. 14a shows the column-integrated time-mean $\langle \Delta Q_q \rangle$. Figs. 14b shows its physics component $\langle \Delta Q_q^p \rangle$. The map of the total nudging tendency looks like a horizontally smoothed version of the physics nudging tendency.

Fig. 14c shows the residual, interpreted as the dynamics nudging tendency $\langle \Delta Q_d^d \rangle$. This has a very small global mean of about -0.01 mm/d, because it is associated with fine-coarse differences in resolved humidity advection, which has no global source or sink. It features sharp structures around maxima of the physics drying tendency, where more humidity is being condensed and removed as precipitation in the fine model than in the

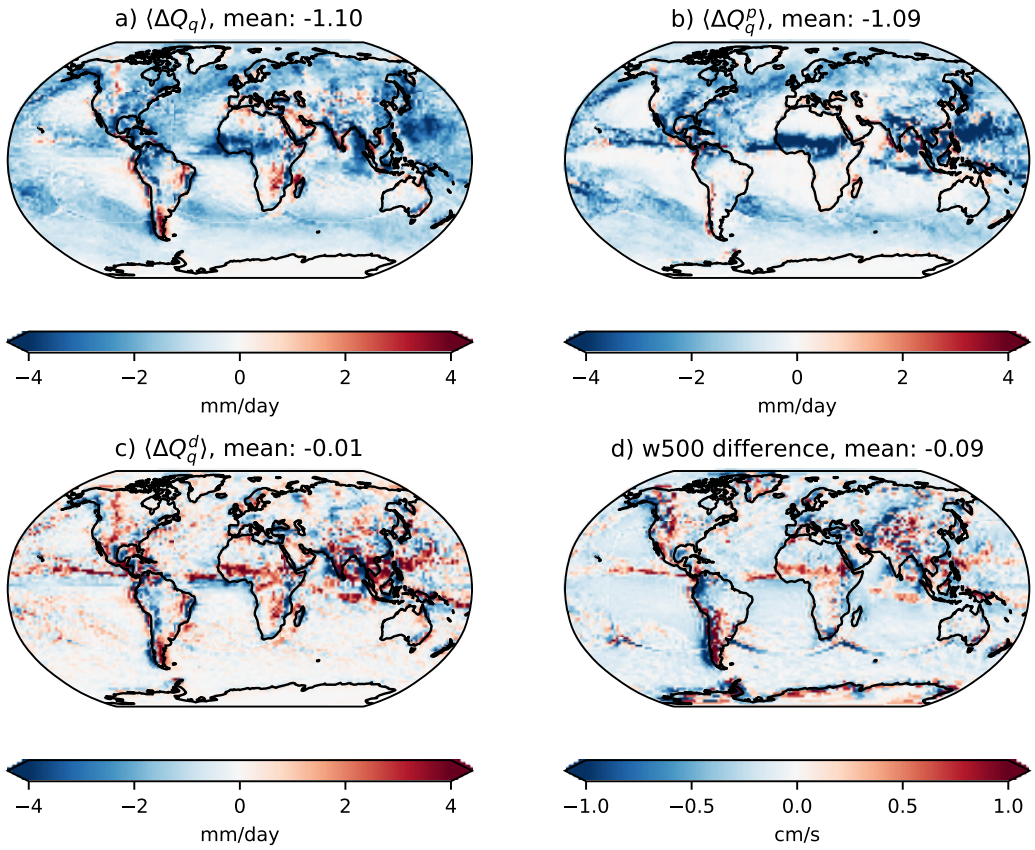

Figure 14. Time-mean (a) total, (b) physics, and (c) dynamics column humidity nudging tendencies from the C48 nudged training run, and (d) corresponding fine-coarse 500 hPa vertical velocity differences.

coarse model. The associated latent heat release buoyantly drives more mid-tropospheric upward motion in the fine model than in the coarse model (Fig. 14d). This forms the upward branch of a Hadley-cell-like dynamical response of the nudged coarse model to the fine-coarse latent heating difference. There are also $\langle \Delta Q_q^p \rangle$ signals where air flows across mountain ranges such as the Andes or Rockies. Comparison of Fig. 14c with Fig. 14d suggests that these signals are also associated with time-mean fine-coarse vertical velocity differences, driven by better channeling of the airflow through the better-resolved topography of the fine model. Near mountains and ITCZs, the dynamical component of the humidity nudging tendency can be as large as the physics component, but over other parts of the globe it is much weaker. These results suggest that column-local prediction of the nudging tendencies may be a useful approximation in most locations.

8 Conclusions

Following the nudge-to-observations corrective ML methodology of WM21, we have developed a nudge-to-fine approach that uses ML to make a coarse-grid global atmospheric model behave more similarly to a reference fine-grid model. Compared to a coarse-grid baseline model, nudge-to-fine ML can improve weather forecasts out to ten days, reduce time-mean precipitation biases by 30%, and reduce global time-mean errors relative to the reference model in other fields such as lower tropospheric temperature and precipitable water.

The ML is trained by nudging a coarse-grid (200 km) version of the FV3GFS model to a 40-day fine-grid (3 km) global simulation made using X-SHiELD, GFDL's modified version of FV3GFS. Both simulations have 79 vertical model levels. The fine-grid output is coarsened in line to allow smaller data volumes. The nudging time scale is 3 hours. The ML is trained to predict the vertical profiles of nudging tendencies of temperature, humidity and (optionally) horizontal wind components on the coarse grid, using the column state for features.

Both the baseline and nudged-coarse simulations simulate too little cloud. During the day, this leads to strong radiative heating of land surfaces, resulting in excess latent and sensible heat fluxes. This bias is successfully corrected in the nudged-coarse simulations by overwriting the coarse-model downward radiative flux with the fine-grid results. ML is used to predict the downward radiative fluxes from these fine-grid results for use in prognostic (forecast) simulations.

The surface precipitation is also overwritten with fine-grid output for the nudged run. For prognostic simulations, the surface precipitation predicted by the physical parameterizations is corrected by subtracting the machine-learned column integrated humidity nudging tendency. As with the nudge-to-observations approach, 40% of global precipitation comes from the humidity nudging. Correcting the surface radiative fluxes, a novel feature of this work, is key to obtaining forecasts with unbiased average land surface precipitation.

We compared off-line and prognostic skill using random forests and neural nets for the ML of nudging tendencies and surface radiative fluxes. The offline skill of instantaneous predictions of all four nudging tendencies (for T, q, u and v) predicted by both RFs

and NNs was modest (explained variance fractions of 0.1-0.4, depending on pressure), but both models accurately captured their time-mean distributions. We used a basket of five metrics of prognostic (online) skill, two measuring weather forecast skill and three measuring mean-state bias relative to the fine-grid reference, to choose an optimal ML configuration. This configuration uses a NN ensemble for temperature and humidity tendency correction, another NN ensemble for surface radiation, but no wind correction. Adding learned wind corrections improves 3–7 day 500 hPa forecast skill but induces substantial 200 hPa temperature biases in the following simulated month. Random forests give results that are almost as good as the optimal NN configuration.

The training and machine learning employ a sophisticated cloud-based workflow that wraps the main components of FV3GFS in Python. While our open-source software for doing this necessarily confronts details of the FV3GFS, its overall structure and the conceptual basis of the nudge-to-fine corrective ML approach can transfer to other global weather and climate models.

The results shown here only scratch the surface of how machine learning using coarsened outputs from fine-grid real-geography global models could improve coarse grid models. Nudge-to-fine corrective ML could be trained and tested using multi-year GSRM simulations across a range of climates in order to improve coarse-grid climate-change simulations. Within the nudge-to-fine framework, we are investigating numerous refinements to the coarsening, training and machine-learning procedures, including better use of energy and momentum conservation constraints and new ML architectures that can improve offline skill while retaining online stability. Groups using more idealized settings such as aquaplanets are also making progress on these issues (Yuval & O'Gorman, 2020; Beucler et al., 2021). Perhaps within a decade, ML will replace complex and often inaccurate assumptions about subgrid variability in physical parameterizations, leading to more reliable global climate models with increased computational efficiency that better use the talents of skilled human model developers.

Appendix A

A1 Surface elevation correction due to pressure-level coarse-graining

In each coarse grid column, p-coarsening will imply some virtual temperature profile $\overline{T_v}(p)$. Assuming hydrostatic balance in the coarse grid column, the heights of the

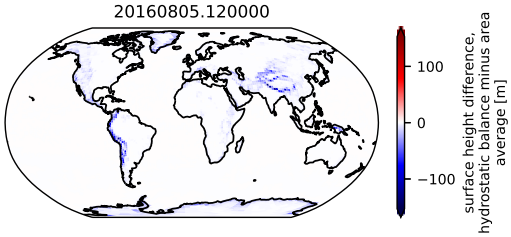

Figure A1. Surface elevation correction needed to maintain hydrostatic balance when p-coarsening the 3 km atmospheric state to a 200 km grid and and conserving column mass.

coarse-grid levels can be found by downward integration of

$$\frac{d(gz)}{d\log p} = -R_d\overline{T_v}(p) \tag{A1}$$

starting with an area-coarsened height of the top model interface at the known top interface pressure of p_T. This implies some surface elevation $z_s^c = \overline{z_s} + \delta z_s^c$ at the known coarsened surface pressure $p_s^c = \overline{p_s}$. Fig. A1 shows a map of the surface elevation correction needed for coarsening the 3 km atmospheric state at one particular time to 200 km resolution, which is strongly negative in coarse grid columns encompassing strong contrasts in fine-grid surface elevation, such as over the Himalayas and Andes.

A2 Definition of explained variance fraction R^2

Let $f(x, y, p, t)$ be the truth and $\tilde{f}(x, y, p, t)$ the prediction. Then R^2 is given by

$$R^2(p) = 1 - \frac{SSE(p)}{SS(p)}. \tag{A2}$$

The sum of squared errors is defined as

$$SSE(p) = \sum_i \Big[f(x_i, y_i, p, t_i) - \tilde{f}(x_i, y_i, p, t_i)\Big]^2 A(x_i, y_i),$$

where A is the grid cell area and the index i ranges over all samples in the test data in which $p < p_s$, the surface pressure. The total sum of squares is given by

$$SS(p) = \sum_i \Big[f(x_i, y_i, p, t_i) - \hat{f}(p)\Big]^2 A(x_i, y_i),$$

where

$$\hat{f}(p) = \frac{\sum_i f(x_i, y_i, p, t_i) A(x_i, y_i)}{\sum_i A(x_i, y_i)}$$

is the pressure-level global average over the test data.

A3 Sensitivity to nudging timescale

We tested the sensitivity of our results to four choices of nudging timescale: $\tau = 1, 3, 6, 12$ hrs. The nudging tendencies are mildly sensitive to τ. For instance, the global-time-mean column drying over the 40-day nudged simulation ranged from 1.15 mm/d for $\tau = 1$ hr to 0.79 mm/d for $\tau = 12$ hr. With a long nudging timescale, the atmosphere moistens slightly (by 1% for $\tau = 12$ hr). The physical parameterizations then make more precipitation and column drying, leaving less for the nudging tendencies to do.

For each τ, we ran two 36-day prognostic simulations using the TqR-NN (no neural net wind correction) and $TquvR$-NN (including neural net wind correction) methodologies. Only a single random seed is trained and shown in each sensitivity test for each timescale. The $\tau = 1$ hr wind-corrected simulation crashed after 13 days. The other simulations all maintained nearly unbiased land-surface precipitation, unlike the baseline simulation (Fig. A2).

Fig. A3 shows zonal-time-mean cross-sections of temperature bias relative to the fine-grid reference. For all nudging timescales, the simulations without wind correction (row (a)) had smaller high latitude upper-tropospheric temperature biases than the wind-corrected simulations (row (b)). The simulations with wind correction were least biased at the longest tested nudging time scale of 12 hr. The simulations without wind correction had minimum temperature biases for $\tau = 3$ hr, closely followed by $\tau = 6$ hr. We obtained similar sensitivities when using random forest learning. These results motivated us to use $\tau = 3$ hrs and no wind nudging as our preferred choice for this paper.

A4 Physics-dynamics decomposition of nudging tendency

To decompose the nudging tendency of an advected scalar a, we start with the advection equations for the nudged coarse model:

$$\frac{\partial a^n}{\partial t} + \nabla \cdot (\mathbf{v}^n a^n) = Q_a^p + \Delta Q_a, \tag{A3}$$

and the coarsened fine model:

$$\frac{\partial \overline{a}}{\partial t} + \nabla \cdot (\overline{\mathbf{v}}\ \overline{a}) = \overline{Q_a}. \tag{A4}$$

By design, the state of the nudged run is forced to evolve similarly to the coarsened fine reference run, so a^n remains close (but not identical) to $\overline{a}$ and similarly for the veloc-

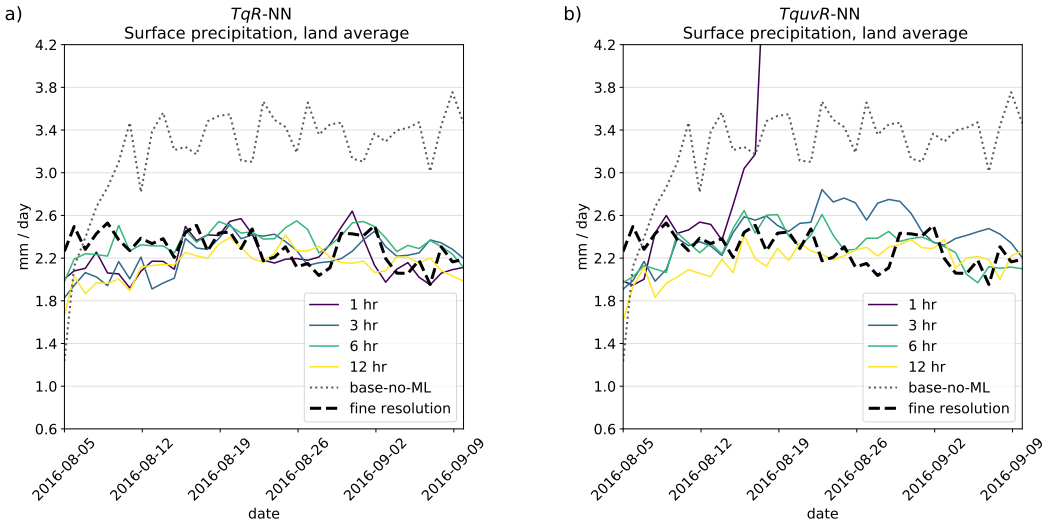

Figure A2. Prognostic run surface precipitation over land for nudging timescales of 1, 3, 6, 12 hrs, compared to the baseline physics and fine-grid model. Panels show (a) TqR-NN and (b) $TquvR$-NN configurations.

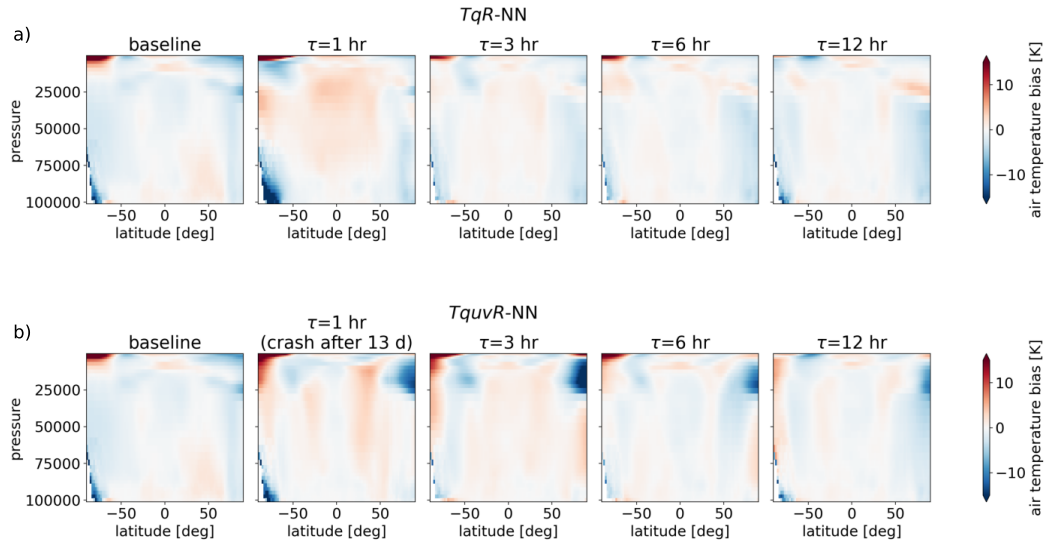

Figure A3. Prognostic run time and zonal mean biases of the air temperature vertical profile for nudging timescales of 1, 3, 6, 12 hrs as well as the baseline model. Note that the time mean of the 1 hr nudging timescale in the $TquvR$-NN case was only taken over the first 13 days of its prognostic run before it crashed; other runs are averaged over the full 36 days.

ity components. Recall that Q_a^p is the source of a due to the coarse-model physical parameterizations. $\overline{Q_a}$ is the apparent source of a for the coarsened fine-grid reference simulation (Yanai et al., 1973). We computed $\overline{Q_a}$ in each coarse grid column every 15 minutes as a sum of contributions from the parameterized physical processes in the fine-grid model plus a vertical eddy flux convergence of a due to fine-grid vertical velocity perturbations from the coarse-grid mean, plus any additional tendency due to nudging of the fine-grid run to an observational analysis. This coarsening is not exact in coarse-model pressure layers that are partly filled by fine-grid topography.

Differencing Eqs. (A3) and (A4) and solving for ΔQ_a, we obtain the decomposition

$$\Delta Q_a = \Delta Q_a^p + \Delta Q_a^d. \tag{A5}$$

Here the physics component is

$$\Delta Q_a^p = \overline{Q_a} - Q_a^p, \tag{A6}$$

The dynamics component is

$$\Delta Q_a^d = \frac{\partial}{\partial t}(\overline{a} - a^n) + \nabla \cdot (\overline{\mathbf{v}}\,\overline{a}) - \nabla \cdot (\mathbf{v}^n a^n), \tag{A7}$$

It has advective and storage terms. The nudging keeps $\overline{a} - a^n$ small. Hence it also keeps the storage term small, especially in time-mean. The advective term is the difference of two flux convergences with zero global mean, and the storage term has near-zero global mean, so the dynamics component of the nudging tendency has a near-zero global mean.

Acknowledgments

We thank Vulcan, Inc. and GFDL for supporting this work. We acknowledge NOAA-EMC, NOAA-GFDL and the UFS Community for publicly hosting source code for the FV3GFS model and NOAA-EMC for providing the necessary forcing data to run FV3GFS. The version of FV3GFS used for this work and the code used to do model training and analysis is available at https://doi.org/10.5281/zenodo.5211066.

References

Beucler, T., Pritchard, M., Rasp, S., Ott, J., Baldi, P., & Gentine, P. (2021). Enforcing analytic constraints in neural networks emulating physical systems. *Phys. Rev. Lett.*, *126*, 098302. doi: 10.1103/PhysRevLett.126.098302

Brenowitz, N. D., & Bretherton, C. S. (2019). Spatially extended tests of a neural network parametrization trained by coarse-graining. *J. Adv. Model. Earth Syst.*, *11*, 2728-2744. doi: 10.1029/2019MS001711

Christopoulos, C., & Schneider, T. (2021). Assessing biases and climate implications of the diurnal precipitation cycle in climate models. *Geophys. Res. Lett.*, *48*, e2021GL093017. doi: 10.1029/2021GL093017

Daley, R. (1981). Normal mode initialization. *Rev. Geophys.*, *19*, 450-468. doi: 10.1029/RG019i003p00450

DelSole, T., Zhao, M., Dirmeyer, P. A., & Kirtman, B. P. (2008). Empirical correction of a coupled land–atmosphere model. *Mon. Wea. Rev.*, *136*(11), 4063–4076. doi: 10.1175/2008mwr2344.1

Han, Y., Zhang, G. J., Huang, X., & Wang, Y. (2020). A moist physics parameterization based on deep learning. *J. Adv. Model. Earth Syst.*, *12*, e2020MS002076. doi: 10.1029/2020MS002076

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. *J. Adv. Model. Earth Syst.*, *12*, e2020MS002223. doi: 10.1029/2020MS002223

Huffman, G. J., Adler, R. F., Morrissey, M. M., Bolvin, D. T., Curtis, S., Joyce, R., ... Susskind, J. (2001). Global precipitation at one-degree daily resolution from multisatellite observations. *J. Hydromet.*, *2*, 36 - 50. doi: 10.1175/1525-7541(2001)002⟨0036:GPAODD⟩2.0.CO;2

Leith, C. E. (1978). Objective methods for weather prediction. *Annu. Rev. Fluid Mech.*, *10*, 107 - 128.

Lyu, G., Köhl, A., Matei, I., & Stammer, D. (2018). Adjoint-based climate model tuning: Application to the planet simulator. *J. Adv. Model. Earth Syst.*, *10*, 207-222. doi: 10.1002/2017MS001194

McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J., Davis, E., ... Fuhrer, O. (2021). fv3gfs-wrapper: a python wrapper of the FV3GFS atmospheric model. *Geosci. Model Dev. Disc.*, *14*, 4401–4409. doi: 10.5194/gmd-14-4401-2021

Mooers, G., Pritchard, M., Beucler, T., Ott, J., Yacalis, G., Baldi, P., & Gentine, P. (2021). Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *J.*

Adv. Model. Earth Syst., *13*, e2020MS002385. doi: 10.1029/2020MS002385

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci.*, *115*, 9684–9689. doi: 10.1073/pnas.1810286115

Stevens, B., Satoh, M., Auger, L., Biercamp, J., Bretherton, C. S., Chen, X., ... Zhou, L. (2019). DYAMOND: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Prog. Earth Planet. Sci.*, *6*, 61. doi: 10.1186/s40645-019-0304-z

Tomita, H., Miura, H., Iga, S., Nasumo, T., & Satoh, M. (2005). A global cloud-resolving simulation: Preliminary results from an aquaplanet experiment. *Geophys. Res. Lett.*, *32*, L08805. doi: 10.1029/2005GL022459

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. *Geophys. Res. Lett.*, *48*, e2021GL092555. doi: 10.1029/2021GL092555

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

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

Yuval, J., & O'Gorman, P. A. (2021). Neural-network parameterization of subgrid momentum transport in the atmosphere. *Geophys. Res. Lett.,* submitted. Preprint:. doi: 10.1002/essoar.10507557.1

Zhou, X., Atlas, R., McCoy, I. L., Bretherton, C. S., Bardeen, C., Gettelman, A., ... Ming, Y. (2021). Evaluation of cloud and precipitation simulations in CAM6 and AM4 using observations over the Southern Ocean. *Earth Space Sci.*, *8*, e2020EA001241. doi: 10.1029/2020EA001241