Correcting weather and climate models by machine learning nudged historical simulations

Oliver Watt-Meyer[1], Noah Brenowitz[1], Spencer K. Clark[1,2], Brian Henn[1], Anna Kwa[1], Jeremy McGibbon[1], W. Andre Perkins[1], Chris Bretherton[1,3]

[1]Vulcan Inc., Seattle, WA
[2]Geophysical Fluid Dynamics Laboratory, NOAA, Princeton, NJ
[3]Department of Atmospheric Sciences, University of Washington, Seattle, WA

Key Points:

- Nudging an atmospheric model towards observations is a good way to estimate state-dependent biases
- Machine learning of state-dependent biases improves hindcast skill of a coarse-resolution GCM
- Bias-corrected year-long simulations are stable and reduce mean precipitation errors

Corresponding author: Oliver Watt-Meyer, `oliwm@vulcan.com`

Abstract

Due to limited resolution and inaccurate physical parameterizations, weather and climate models consistently develop biases compared to the observed atmosphere. These biases are problematic for forecasting on timescales from medium-range weather to centennial-scale climate. Using the FV3GFS model at coarse resolution, we propose a method of machine learning corrective tendencies from a hindcast simulation nudged towards an observational analysis. We show that a random forest can predict the nudging tendencies from this hindcast simulation using only the model state as input. This random forest is then coupled to FV3GFS, adding corrective tendencies of temperature, specific humidity and horizontal winds at each timestep. The coupled model shows no signs of instability in year-long simulations and has significant reductions in short-term forecast error for 500hPa height, surface pressure and near-surface temperature. Furthermore, the root mean square error of the annual-mean precipitation is reduced by about 20%.

Plain Language Summary

After initialization from a realistic snapshot of the atmosphere, weather and climate models inevitably develop predictable errors compared to the real world. This decreases the usefulness of forecasts. These errors arise from the coarse resolution of the numerical models and from the uncertain treatment of small-scale processes. We propose a method to reduce these errors by training a machine learning model to correct for them as the atmospheric model proceeds. We show that a random forest can make reasonably skillful predictions of the required correction using a snapshot of the model state as input. When we make a forecast with the machine-learning corrected model, the lead-time for the prediction of important mid-tropospheric and surface variables is increased by half a day to a day. The pattern of precipitation predicted by the machine learning corrected model is also more realistic, with a decrease in excessive rainfall over high mountains. On the other hand, the corrected model develops larger errors in temperature in the high latitudes, particularly in the lower stratosphere.

1 Introduction

Despite steady improvements in the skill of numerical weather and climate models over the last decades, a longstanding issue is the development of biases after initialization. These biases (systematic forecast errors) cause degradation of performance for both medium range weather forecasting and subseasonal to decadal climate predictions. They arise from limited resolution and inaccurate physical parameterizations. Typically, post-processing steps are developed to handle these biases such as model output statistics for weather forecasting (Glahn & Lowry, 1972) or ensemble bias correction for seasonal prediction (Stockdale et al., 1988; Arribas et al., 2011). In this study, we propose an online bias correction method using machine learning (ML). We apply a corrective tendency to the prognostic state of the atmospheric model at each time step in order to reduce model error growth. The necessary corrective tendencies are estimated from a hindcast simulation which is linearly nudged towards an observational analysis. An ML model is trained to predict the nudging tendencies using only the state of the model as inputs. This ML model can then be used in a forecast to keep the model evolution on a more realistic manifold.

Online bias correction has been previously proposed (Leith, 1978; Saha, 1992; DelSole & Hou, 1999) and implemented in a prototype manner (Danforth et al., 2007; DelSole et al., 2008; Yang et al., 2008). In these studies, a corrective tendency is typically estimated from the error growth within the first day of a forecast and the applied tendencies are time-mean or seasonal-mean values. It was found that applying such a correction can lead to the reduction of error growth of corrected variables. State-dependent corrections, typically linearly dependent on the atmospheric state (e.g. DelSole et al., 2008), have been attempted but with little benefit over time-mean tendencies. The distinguishing features of this work are the use of a non-linear function estimator (specifically a random forest) to estimate the corrective tendencies, and the consideration of the effects of correcting specific humidity onto the surface precipitation.

The use of ML for atmospheric model parameterization has seen significant recent effort (Krasnopolsky et al., 2013; Rasp et al., 2018; Brenowitz & Bretherton, 2018). The typical goal has been whole-scale replacement of physical parameterizations either by emulating the behavior of an existing scheme (Krasnopolsky et al., 2005; O'Gorman & Dwyer, 2018) or by learning from high-resolution simulations (Brenowitz & Bretherton, 2018, 2019; Yuval & O'Gorman, 2020) or reanalysis (McGibbon & Bretherton, 2019). In this work, we leverage the significant effort that has already been put into developing skillful physics routines and use ML to provide a correction on top of a full suite of parameterizations. This empirical strategy also reveals physical processes in the target model which are behaving unrealistically (Rodwell & Palmer, 2007). Thus, it provides information that can be used to tune existing physical parameterizations and an automated way to correct for remaining biases after tuning. The proposed method uses existing observational analysis data and does not require costly high-resolution simulations to generate training data. This makes it amenable for groups who wish to explore improving their GCMs with ML but do not have capability for global storm-resolving simulations (Stevens et al., 2019; Harris et al., 2020).

2 Methods

2.1 Atmospheric model

To test our proposed method we use NOAA's global weather forecast system FV3GFS (Zhou et al., 2019). FV3GFS is based on the FV3 non-hydrostatic dynamical core on a cubed-sphere grid (Putman & Lin, 2007) coupled to physics parameterizations implemented by NOAA's Environmental Modeling Center. Briefly, we use the hybrid eddy-diffusivity mass flux turbulence scheme (Han et al., 2016), GFDL microphysics (Zhou et al., 2019), scale-aware mass flux convection scheme (Han & Pan, 2011), RRTMG radiation (Iacono et al., 2008), and the mountain blocking and orographic gravity wave drag parameterization. The operational version uses C768 (13 km) grid resolution and 64 vertical levels (NOAA, 2018). We use a coarse C48 (approximately 200km) horizontal resolution with 79 vertical levels and a physics timestep of 15 minutes.

2.2 Nudging approach

In order to estimate the atmospheric model biases across seasons and the diurnal cycle, we perform a two-year hindcast simulation in which the prognostic state is continuously nudged towards an observational analysis (Fig. 1). Specifically, a linear relaxation term is added to the prognostic equations of certain variables:

$$\frac{\partial a}{\partial t} = -\mathbf{v} \cdot \nabla a + Q_a^p \underbrace{-\frac{a - a_{obs}}{\tau}}_{\Delta Q_a}, \tag{1}$$

where a is a prognostic variable, $-\mathbf{v}\cdot\nabla a$ is advection by the dynamical core, Q_a^p is the tendency of a due to all physical parameterizations (e.g. Yanai et al., 1973), a_{obs} is an estimate of the observed value of a at the given time and position, and τ is a nudging timescale. The nudging tendencies ΔQ_a are saved as a diagnostic and are the target for the ML described in Section 2.3. The nudging keeps the model simulation tracking close to the observed evolution of the atmosphere and the nudging tendencies are an estimate of the (negative) model error throughout the simulation.

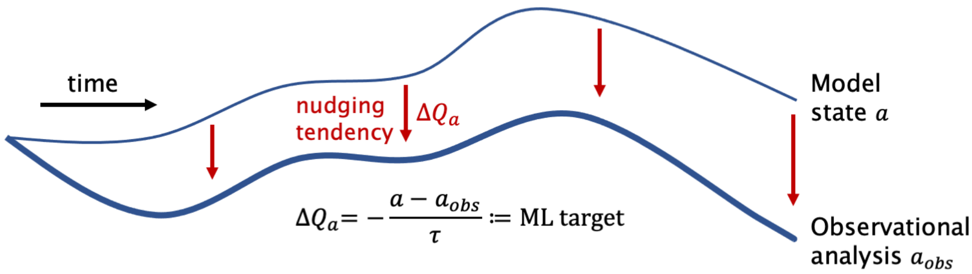

Figure 1. Schematic of procedure to generate the nudging tendencies which serve as a target for the ML model.

The nudging is active for temperature (nudging tendency labeled ΔQ_1), specific humidity (ΔQ_2), horizontal winds (ΔQ_u and ΔQ_v) and surface pressure.[1] A 6-hour timescale τ is used for all variables. The reference dataset is the GFS analysis (NCEI, 2020) on a 1.4° latitude-longitude grid. The analysis is available every 6 hours, and is linearly interpolated to obtain a state in between these times. At each timestep during the simulation, the analysis is interpolated vertically to the model's pressure surfaces as well as horizontally to FV3GFS's cubed-sphere grid. No nudging is applied to any variable in the top-most model level to avoid the sponge layer, and no nudging is applied for specific humidity above 100hPa due to low confidence in the analysis dataset at these levels.

Nudging specific humidity impacts the hydrological cycle. For example, if the column-integrated humidity nudging is non-zero, then the nudging is a source or sink of moisture for the atmospheric column. As will be shown in Section 3.1, the humidity nudging dries the vast majority of columns so can typically be interpreted as additional precipitation. Therefore, we subtract the column-integrated moistening due to nudging from the surface precipitation rate generated by the physics parameterizations. For the cases when the moistening due to nudging is larger than the physics precipitation we set the total precipitation rate to zero:

$$P = \max\left(0, P_{physics} - \langle \Delta Q_2 \rangle\right), \tag{2}$$

where $P_{physics}$ is the surface precipitation rate produced by the physics parameterizations (the shallow convection, deep convection and microphysics schemes) and

$$\langle \Delta Q_2 \rangle = \frac{1}{g} \int_0^{p_s} \Delta Q_2 \, dp. \tag{3}$$

The clipping at zero in Eq. 2 effectively acts a moisture source for the coupled land-atmosphere system with consequences described in the discussion section.

The FV3 dynamical core uses D-grid staggering (Arakawa & Lamb, 1977) and the horizontal winds point in grid-relative directions instead of east and north. To nudge the winds, they are interpolated to the grid center and rotated to latitude-longitude coordinates before the nudging tendencies are computed and then transformed back to the D-grid. This is analogous to how the GFS physical parameterizations interact with the dynamical core winds.

[1] Surface pressure is not a prognostic variable in the non-hydrostatic FV3GFS model. The nudging tendency is computed using the diagnosed surface pressure, and then applied to the pressure thickness of each atmospheric layer proportionally to the coefficient of relation between the layer pressure and surface pressure specified by the vertical hybrid-sigma coordinate.

Over the ocean, the surface boundary condition is a prescribed sea-surface temperature from the same GFS analysis dataset used for the nudging. The monthly 1982-2012 climatology of sea ice extent from the NCEP Climate Forecast System Reanalysis (Saha et al., 2010) is used to determine the ice-ocean boundary.

2.3 Machine learning architecture

A random forest is trained using the scikit-learn Python package (Pedregosa et al., 2011) to predict the nudging tendencies for a particular GCM column given the atmospheric profile at this column. The inputs and outputs are taken from the nudged hindcast simulation described above. The random forest predicts the nudging tendencies of temperature, specific humidity, eastward wind and northward wind. Its inputs are temperature, specific humidity, eastward wind, northward wind, the land/sea/sea-ice mask, surface geopotential and the cosine of the solar zenith angle. The first four inputs, which depend on the vertical level, describe the state of the atmosphere. The mask and surface geopotential distinguish between land and ocean and indicate surface topography. The cosine of zenith angle is a proxy for insolation.

The random forest is trained by minimizing a mean squared error loss function in which each scalar output is normalized by its standard deviation. Sixteen individual decision trees form the random forest; each tree has a maximum depth of thirteen. Section 2.5 will describe the sampling of the training and test data in more detail.

2.4 Coupling of machine learning to GCM

We use a Python wrapper of the FV3GFS Fortran model (McGibbon et al., 2021) in order to execute Python code during the model simulation. Briefly, the wrapper allows viewing and modifying the model state from a Python script at certain checkpoints in the main Fortran time loop. We obtain the input variables at the end of each timestep, evaluate the random forest to compute tendencies of temperature, humidity and winds, multiply these by the physics timestep and then apply these increments to the model state. The tendency of specific humidity predicted by the random forest is limited so that the resulting specific humidity is not negative. Without this adjustment, regions of negative humidity arise near the poles and typically lead to model crashes after about two months. The effects of the column moisture tendency from the ML on surface precipitation is handled in the same way as the nudging case (Eq. 2).

The random forest prediction at each timestep takes about one quarter the time as the full suite of physics parameterizations. This is about 10% of the total wall clock time for the simulation, only a slight increase in computational cost. On the other hand, the random forest trained for this study requires about 360 MB of memory, which is a substantial addition to the approximately 600 MB required on each processor to run a baseline version of FV3GFS at C48 resolution, assuming one rank per cubed-sphere tile.

2.5 Experiment configuration and validation

A procedure is designed to 1) generate training data from across the seasonal cycle and 2) test the online and offline model skill on a time period independent from the training data. We first perform a two-year long simulation that is initialized from GFS analysis on 1 January 2015 and continuously nudged towards the GFS analysis as described in Section 2.2. The nudging tendencies and prognostic state are saved every five hours to ensure sampling around the diurnal cycle.

The random forest is trained on output from the first year of the two-year simulation. Columns from 160 time steps which uniformly span 2015 are used for training, resulting in about 2.2M samples (all $6 \cdot 48 \cdot 48 = 13824$ columns are used for each time).

To evaluate the offline skill of the random forest, a test dataset of 90 evenly spaced times is chosen from the second year (2016) of the two-year nudged run. The performance of FV3GFS coupled to the random forest, which we call online skill, is tested in two ways. First, we initialize twelve 10-day forecasts each starting from the first of the month for every month of 2016. These will be used to evaluate the error growth on short- to medium-range weather forecasting timescales. Second, we initialize a single year-long run on 1 January 2016 in order to evaluate longer timescale statistics. Time-mean biases in precipitation and other fields will be diagnosed from this simulation. All forecasts are initialized from GFS analysis. We compare the ML-corrected simulations against identically-configured baseline runs without ML.

To compute errors in the online simulations, we use the second half of the two-year nudged simulation as truth. For variables which are directly nudged (temperature, humidity, surface pressure and horizontal winds) this is a reasonable representation of the true state of the atmosphere for our purposes, assuming that model errors are substantially larger than the errors of the GFS analysis. However, precipitation or other diagnostic quantities in that simulation may differ strongly from observational estimates. Thus, we compare the simulated precipitation patterns against daily data from the Global Precipitation Climatology Project v1.3 (GPCPv1.3; Huffman et al., 2001). The observed product is on a 1° by 1° latitude-longitude grid and the model output is interpolated from the cubed-sphere to this grid for comparison.

3 Results

3.1 Nudging tendencies and offline performance

Before evaluating the performance of the random forest, it is useful to examine the structure of the nudging tendencies. By definition, the time-mean model bias relative to the reference analysis dataset is equal to the negative of the time-mean nudging tendency multiplied by the nudging timescale (Eq. 1). Therefore, Figs. 2a and 2c show that our baseline configuration of the FV3GFS model drifts moister and cooler than the GFS analysis in the column integral since the nudging tends to dry and heat in most regions. The spatial pattern of the nudging indicates that it especially strengthens the drying and heating in convective regions—Indo-Pacific warm pool and inter-tropical convergence zone—and in midlatitude fronts (see also Supplemental Movie). The imprinting of the cubed-sphere grid in Fig. 2a is due to the nudging tendency correcting artifacts introduced by the dynamical core at the coarse C48 resolution, and we expect this signal would be diminished at higher resolutions (Zhao et al., 2018). The pattern of the tendencies suggest that the nudging of temperature and humidity amplifies precipitation and latent heating, likely correcting a bias of the convective parameterization to generate insufficient rainfall in realistic conditions for this grid resolution.

When evaluated offline on samples from the test data, the random forest successfully predicts the time-mean pattern of heating and moistening (Figs. 2b and 2d). The random forest also has only small global-mean column-integrated biases: about 1.3% too much heating and 2.5% too much drying. On the other hand, the ML does not reproduce some finer-scale features of the test data such as the heating/cooling dipole near the tip of South America, regional patterns of heating/cooling over land and the cubed-sphere grid artifacts.

It is important to also evaluate the skill of the random forest in making instantaneous predictions of the nudging tendency. Figure S2a-b shows the zonal mean R^2 skill for the heating and moistening as a function of latitude and pressure. The heating (ΔQ_1) predictions are substantially more skillful than the moistening (ΔQ_2) predictions. In the tropical lower troposphere, the random forest predictions explain 30-50% of the variance of temperature nudging tendencies. There is also notable skill around the tropical tropopause,

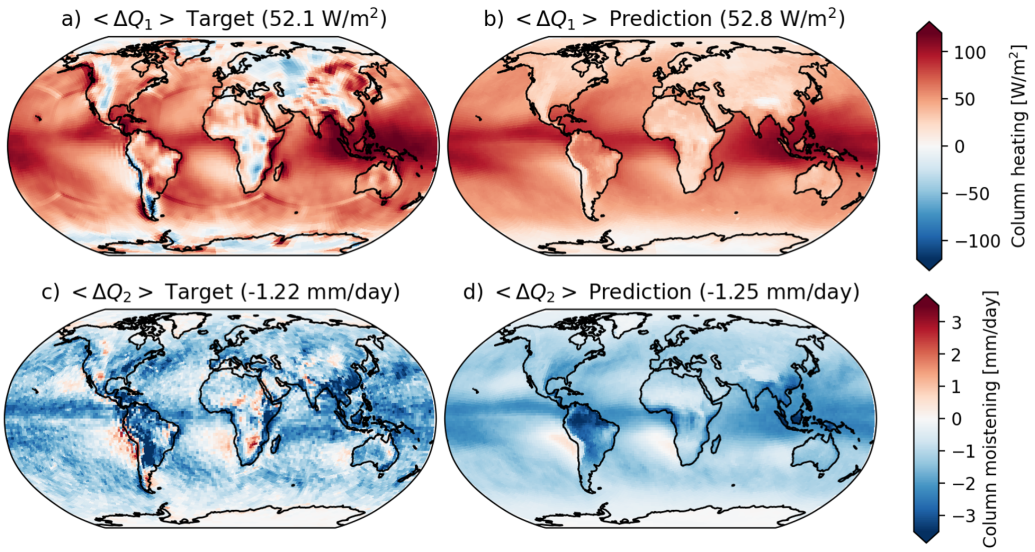

Figure 2. Column integrated heating (top) and moistening (bottom) from nudging, time-averaged across offline test data. Actual tendencies from nudged simulation shown on left and offline random forest predictions on right. Values in titles are the global mean of each panel.

the Northern Hemisphere mid-latitudes and the polar regions. The humidity nudging predictions are less skillful, with a maximimum of 20% of variance explained in the upper troposphere near the equator. Despite the apparent low skill, recall that the random forest accurately predicts the time-mean humidity nudging tendency. Part of the reason for poor R^2 performance is that the nudging tendency of ΔQ_2 is particularly noisy due to fast variability of the model state. For example, note the local speckling in Fig. 2c, which is already an average over 90 timesteps. We do not expect this noisiness to be learnable and indeed it is smoothed by the random forest (Fig. 2d).

3.2 Online performance: weather skill

How does the ML-corrected FV3GFS performs when evaluated with metrics for weather forecast skill and climate drift? A key measure for the skill of a weather model is the speed at which the global root mean squared error (RMSE) of particular variables grows. This indicates how well the model simulates the evolution of the circulation of the atmosphere.

Figure 3 shows global RMSE of 500-hPa geopotential height, surface pressure and lowest model layer temperature (see Section 2.5 for details of the forecast experiments). The ML-corrected FV3GFS has significantly lower error than the baseline model for all three of these variables at lead times ranging from 1-day to 10-days. Depending on the variable and time elapsed, the ML-corrected FV3GFS model is able to make equally skillful forecasts from half to a full day further into the future. This is a substantial improvement given the marginal increase in computational cost associated with evaluating the random forest once per timestep. Furthermore, no variable we have examined has significantly worse skill on the 10-day timescale in the ML-corrected model compared to the baseline.

What drives the improvements in Fig. 3? We trained an random forest to only predict ΔQ_1 and ΔQ_2 and not predict the momentum tendencies (blue lines in Fig. S3). Clearly, the increase in forecast skill for surface pressure arises from predicting the wind tendencies. The baseline model has a biased zonal mean surface pressure pattern, with

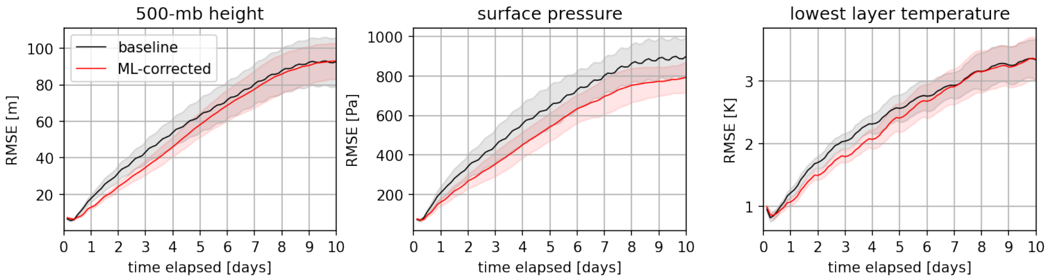

Figure 3. Global root mean squared error for (left) 500hPa geopotential height, (middle) surface pressure and (right) lowest model layer temperature. Averaged across 12 forecasts initialized on the first of every month of 2016. Shading shows one standard deviation. Baseline (black) is standard FV3GFS model and ML-corrected (red) is the FV3GFS coupled to the ML model.

overly high pressure in the polar regions and low pressure in the tropics. The ML correction of winds strongly decreases this bias. On the other hand, the increase in skill for near-surface temperature is similar for the two ML models, indicating that the corrective tendencies of temperature and/or specific humidity are responsible for this improvement.

3.3 Online performance: climate skill

For multi-year climate simulations there are additional requirements for any machine-learning corrected GCM. The model must be able to run indefinitely without numerical instabilities arising. Some previous works using ML for parameterization replacement have struggled with this issue, especially when using neural networks (e.g. Brenowitz & Bretherton, 2019; Brenowitz, Beucler, et al., 2020; Rasp, 2020). Furthermore, the climate of the model must not drift far from a realistic state over the course of a months- to years-long run. Ideally, the machine-learning corrected model will have a climate state that is less biased than the baseline model.

We perform a year-long simulation initialized on 1 January 2016. Since the training data for the random forest is drawn from 2015 only, this is an independent verification time period. The ML-corrected model runs for the full year without any crashes or any special effort to tune its architecture or hyperparameters. It was necessary to add a limiter to the online predictions of the specific humidity tendencies by the random forest to ensure that the specific humidity did not become negative. Without this limiter, which is active in the upper troposphere in about 15% of grid columns on average, regions of negative specific humidity develop and lead to very cold temperatures near the surface that eventually cause model crashes.

The climatological spatial pattern of precipitation in the ML-corrected simulation is notably improved compared to the baseline run (Fig. 4). Using the GPCPv1.3 dataset (Huffman et al., 2001) as a reference, the spatial RMSE of the 2016-mean precipitation substantially decreases by about 24%, from 2.14 mm/day to 1.62 mm/day. While there is a slight increase in the global mean bias of precipitation, this quantity is not well-constrained by the observations (Sun et al., 2018). For comparison, the RMSE of 2016-mean precipitation in the run that is directly nudged towards the GFS analysis is 1.39 mm/day (Fig. S4). This is a lower bound on the precipitation RMSE we might expect from the ML-corrected run, suggesting it has realized over two-thirds of the greatest possible precipitation bias improvement we might hope to achieve. The reduction of precipitation er-

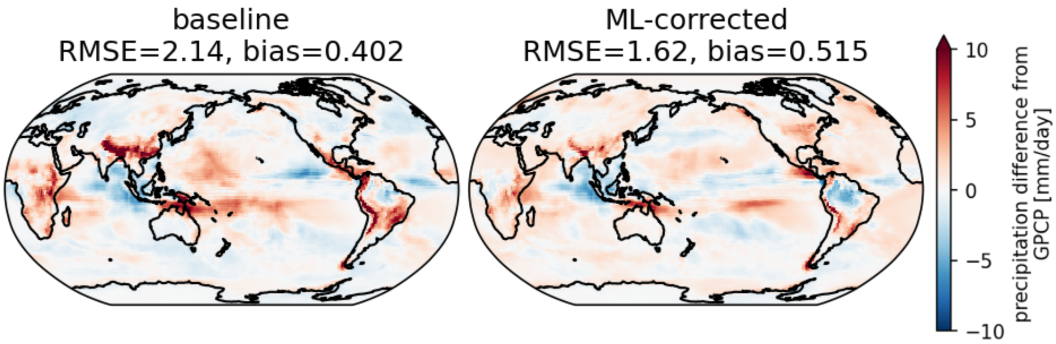

Figure 4. Bias of precipitation, $P_{physics} - \langle \Delta Q_2 \rangle$, averaged over 2016. Bias is computed relative to GPCPv1.3 observational product. Global root mean square error and global mean bias are shown in titles for each run in units of mm/day.

rors mostly arises from the corrective tendencies of temperature and moisture (compare bottom panels of Fig. S4).

The baseline model strongly overpredicts precipitation over the Himalaya, Southeast Asia, and the Andes (Fig. 4). In the ML-corrected FV3GFS, the biases of precipitation over these regions are much smaller in magnitude and cover a smaller area. Over the ocean, the largest biases are mostly decreased in the ML-corrected run (e.g. see Western Pacific). However, the corrected run also has slightly too much precipitation in subtropical regions where there is typically descent. This artifact arises from the nudging method rather than the ML, as the nudged run has a similar bias (Fig. S4).

In the global mean, the year-long ML-corrected runs remain fairly close to the verification data for total water path and lower tropospheric temperature (Fig. S5). However, over the first few weeks of the simulation, prognostic variables such as temperature and zonal wind develop substantial regional biases in the ML-corrected runs. There is a strong annual-mean bias (up to 30K) of temperature in the polar regions at around 100hPa - 250hPa (Fig. S6) and related biases in zonal mean zonal wind (not shown).

4 Discussion

The nudging tendencies can be interpreted as biases of the physical parameterizations of the FV3GFS model at our chosen resolution. For example, the additional heating and moistening done by the nudging in regions of convection (Fig. 2 and Supplemental Movie) indicate that the convective parameterization is generating too little precipitation. Similarly, the nudging tendency of winds show an acceleration over topography in the time-mean (Fig. S1) suggesting that the gravity wave drag parameterization may be overly active in the column mean. It is likely that tuning of the parameterizations could reduce the size of these biases and decrease the corrective nudging tendencies that must be machine-learned.

The nudging timescale τ (Eq. 1) is a free parameter for this method. In principle, a shorter nudging timescale will allow the ML correction to represent faster timescale processes and better represent the diurnal cycle. On the other hand, for physical processes such as boundary layer turbulence which happen on the timescale of hours, there can be a constant tug-of-war between the nudging tendencies and boundary layer tendencies if the nudging timescale is too short. The 6-hour nudging timescale we used provided a balance between these competing issues and was also a natural choice given the 6-hourly availability of the GFS analysis. Using a 6-hour nudging timescale for the temperature and humidity nudging while using a 24-hour timescale for the wind and surface pressure nudging lead to worse offline and online performance of the random forest (not shown).

Although the offline skill of the trained random forest is somewhat modest (Section 3.1), our strategy is to use ML to apply a correction to a complete set of parameterizations. Thus, we neither expect nor require that the ML model we train have exceptional offline skill. Even a time-mean tendency prediction would provide some benefit (DelSole et al., 2008) and any additional ML-derived skill has potential to gain further improvements once coupled back to the atmospheric model.

Ideally, one would apply the ML corrections to the the same model that is used to generate the nudging target (i.e. the analysis). This would ensure that the nudging tendencies represent actual corrections towards the observed state of the atmosphere instead of, for example, the difference between the boundary layer parameterizations of the two models. In an operational weather forecasting context, it would be possible to adapt this method to learn analysis increments from a fully-fledged data assimilation system and this would ensure consistency between the models.

The coupling between the ML tendencies of the atmosphere and the land surface is a key aspect of this method, in particular because the nudging of specific humidity accounts for about a third of the global mean drying of the atmosphere. Without adding the column integrated nudging or ML tendency of humidity to the surface precipitation (Eq. 2) there is a strong drying of the land surface globally. However, due to the requirement of maintaining positive precipitation and not having a simple way to modify the evaporation predicted by the land-surface model, we have introduced a small but significant moisture source to the coupled land-atmosphere. The nudging in turn has to counteract this moisture source with further drying, and this may lead to a biased estimate of the proper nudging tendency of moisture. Ongoing work is exploring whether nudging soil moisture and learning these tendencies (e.g. DelSole et al., 2008) could help address this issue.

5 Conclusions

We propose a method to perform online bias correction of a general circulation model using machine learning of nudging tendencies from a hindcast simulation. A random forest is able to make reasonably skillful predictions of the nudging tendencies using only the model state as input. When coupled back to the atmospheric model, the ML-corrected GCM increases its lead-time forecasts for 500hPa geopotential height and surface pressure by about a day, and for near-surface temperature by about half a day. Furthermore, the RMSE of the time-mean pattern of precipitation is reduced by about 20%. These improvements come with only slight increase in computational cost. However, significant temperature biases develop in the polar lower stratosphere after a number of weeks in the ML-corrected simulations.

One area for future work is investigating how much this method improves higher-resolution (e.g. operational weather forecast) simulations. Second, being able to predict the ML correction with a neural network architecture would also be useful for highly parallel simulations where memory use is a limitation. Neural networks also show better skill than random forests in offline tests, although this is not necessarily a key factor for online skill (Brenowitz, Henn, et al., 2020; Yuval et al., 2020). Generating a less noisy training dataset, for example by smoothing the nudging tendencies in time, could also lead to better offline skill.

Due to the use of historical analysis data, the training dataset is restricted to the climate of the last few decades and the proposed method may have limitations for use in climate-change scenarios due to out-of-sample inputs (e.g. O'Gorman & Dwyer, 2018).

To handle this limitation, one can use a high-resolution model as a target dataset for nudging and run the high-resolution simulations for current and future warmed simulations. We will report on this approach in a forthcoming paper.

Acknowledgments

We thank Vulcan, Inc. for supporting this work. Lucas M. Harris and Larry W. Horowitz at GFDL provided assistance in using the nudging capability of FV3GFS and supplied post-processed GFS analysis data in a form compatible for nudging. We acknowledge NOAA-EMC, NOAA-GFDL and the UFS Community for publicly hosting source code for the FV3GFS model (https://github.com/ufs-community/ufs-weather-model) and NOAA-EMC for providing the necessary forcing data to run FV3GFS. The version of FV3GFS used for this work (https://github.com/VulcanClimateModeling/fv3gfs-fortran) contains minor changes to ease use on cloud computing and as part of a python wrapper (https://github.com/VulcanClimateModeling/fv3gfs-wrapper). GPCPv1.3 data was obtained from NOAA-NCEI (https://doi.org/10.7289/V5RX998Z). GFS analysis data is available at https://www.nco.ncep.noaa.gov/pmb/products/gfs. If and when this manuscript is accepted, we will provide DOI's for the two Vulcan Climate Modeling GitHub repositories listed above.

References

Arakawa, A., & Lamb, V. R. (1977). Computational design of the basic dynamical processes of the UCLA general circulation model. In J. CHANG (Ed.), *General circulation models of the atmosphere* (Vol. 17, p. 173 - 265). Elsevier. doi: https://doi.org/10.1016/B978-0-12-460817-7.50009-4

Arribas, A., Glover, M., Maidens, A., Peterson, K., Gordon, M., MacLachlan, C., ... Cusack, S. (2011). The GloSea4 ensemble prediction system for seasonal forecasting. *Monthly Weather Review*, *139*(6), 1891 - 1910. doi: 10.1175/2010MWR3615.1

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

Brenowitz, N. D., & Bretherton, C. S. (2018). Prognostic validation of a neural network unified physics parameterization. *Geophysical Research Letters*, *45*(12), 6289-6298. doi: https://doi.org/10.1029/2018GL078510

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

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., ... Bretherton, C. S. (2020). *Machine learning climate model dynamics: Offline versus online performance.*

Danforth, C. M., Kalnay, E., & Miyoshi, T. (2007). Estimating and Correcting Global Weather Model Error. *Monthly Weather Review*, *135*(2), 281–299. doi: 10.1175/mwr3289.1

DelSole, T., & Hou, A. Y. (1999). Empirical correction of a dynamical model. Part I: Fundamental issues. *Monthly Weather Review*, *127*(11), 2533 - 2545. doi: 10.1175/1520-0493(1999)127⟨2533:ECOADM⟩2.0.CO;2

DelSole, T., Zhao, M., Dirmeyer, P. A., & Kirtman, B. P. (2008). Empirical Correction of a Coupled Land–Atmosphere Model. *Monthly Weather Review*, *136*(11), 4063–4076. doi: 10.1175/2008mwr2344.1

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology and Climatology*, *11*(8), 1203 - 1211. doi: 10.1175/1520-0450(1972)011⟨1203:TUOMOS⟩2.0.CO;2

Han, J., & Pan, H.-L. (2011). Revision of convection and vertical diffusion schemes in the NCEP Global Forecast System. *Weather and Forecasting*, *26*(4), 520 - 533. 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). 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

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. (e2020MS002223 2020MS002223) doi: https://doi.org/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. *Journal of Hydrometeorology*, *2*(1), 36 - 50. doi: 10.1175/1525-7541(2001)002⟨0036:GPAODD⟩2.0.CO;2

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

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Chalikov, D. V. (2005). New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Monthly Weather Review*, *133*(5), 1370 - 1383. doi: 10.1175/MWR2923.1

Krasnopolsky, V. M., Fox-Rabinovitz, M. S., & Belochitski, A. A. (2013). Using ensemble of neural networks to learn stochastic convection parameterizations for climate and numerical weather prediction models from data simulated by a cloud resolving model. *Advances in Artificial Neural Systems*, 485913. doi: 10.1155/2013/485913

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

McGibbon, J., & Bretherton, C. S. (2019). Single-column emulation of reanalysis of the Northeast Pacific marine boundary layer. *Geophysical Research Letters*, *46*(16), 10053-10060. doi: https://doi.org/10.1029/2019GL083646

McGibbon, J., et al. (2021). fv3gfs-wrapper: a python wrapper of the FV3GFS atmospheric model. *to be submitted to Geoscientific Model Development*.

NCEI, N. (2020). *Global forecast system.* Retrieved 2020-12-16, from `https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forcast-system-gfs`

NOAA. (2018). *Strategic implementation plan for evolution of NGGPS to a national unified modeling system.* Retrieved from `https://www.weather.gov/media/sti/nggps/UFS%20SIP%20FY19-21_20181129.pdf`

O'Gorman, P. A., & Dwyer, J. G. (2018). Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *Journal of Advances in Modeling Earth Systems*, *10*(10), 2548-2563. doi: https://doi.org/10.1029/2018MS001351

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

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

Rasp, S. (2020). Coupled online learning as a way to tackle instabilities and biases in neural network parameterizations: general algorithms and lorenz 96 case study (v1.0). *Geoscientific Model Development*, *13*(5), 2185–2196. doi: 10.5194/gmd-13-2185-2020

Rasp, S., Pritchard, M. S., & Gentine, P. (2018). 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

Rodwell, M. J., & Palmer, T. N. (2007). Using numerical weather prediction to assess climate models. *Quarterly Journal of the Royal Meteorological Society*, *133*(622), 129–146. doi: 10.1002/qj.23

Saha, S. (1992). Response of the NMC MRF model to systematic error correction within integration. *Mon. Wea. Rev.*, *120*, 345 - 360.

Saha, S., Moorthi, S., Pan, H.-L., Wu, X., Wang, J., Nadiga, S., . . . Goldberg, M. (2010). The NCEP climate forecast system reanalysis. *Bulletin of the American Meteorological Society*, *91*(8), 1015 - 1058. doi: 10.1175/2010BAMS3001.1

Stevens, B., Satoh, M., L., A., et al. (2019). DYAMOND: the dynamics of the atmospheric general circulation modeled on non-hydrostatic domains. *Prog Earth Planet Sci*, *6*, 61. doi: https://doi.org/10.1186/s40645-019-0304-z

Stockdale, T., Anderson, D., Alves, J., et al. (1988). Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. *Nature*, *392*, 370-373. doi: https://doi.org/10.1038/32861

Sun, Q., Miao, C., Duan, Q., Ashouri, H., Sorooshian, S., & Hsu, K.-L. (2018). A review of global precipitation data sets: Data sources, estimation, and intercomparisons. *Reviews of Geophysics*, *56*(1), 79-107. doi: https://doi.org/10.1002/2017RG000574

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

Yang, X., DelSole, T., & Pan, H.-L. (2008). Empirical Correction of the NCEP Global Forecast System. *Monthly Weather Review*, *136*(12), 5224–5233. doi: 10.1175/2008mwr2527.1

Yuval, J., O'Gorman, P. A., & Hill, C. N. (2020). *Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision.*

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

Zhao, M., Golaz, J.-C., Held, I. M., Guo, H., Balaji, V., Benson, R., . . . Xiang, B. (2018). The GFDL global atmosphere and land model AM4.0/LM4.0: 2. Model description, sensitivity studies, and tuning strategies. *Journal of Advances in Modeling Earth Systems*, *10*(3), 735-769. doi: https://doi.org/10.1002/2017MS001209

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