Correcting Weather and Climate Models by Machine Learning Nudged Historical Simulations

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

Abstract Due to limited resolution and inaccurate physical parameterizations, weather and climate models consistently develop biases compared to the observed atmosphere. Using the FV3GFS model at coarse resolution, we propose a method of machine learning corrective tendencies from a hindcast simulation nudged toward observational analysis. We show that a random forest can predict the nudging tendencies from this hindcast simulation with moderate skill 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 500 hPa height, surface pressure and near-surface temperature. Furthermore, the root mean square error of the annual-mean precipitation is reduced by about 20%. Biases of other variables remain similar or in some cases, like upper-atmospheric temperature, increase in the year-long simulations.

Plain Language Summary After initialization from a realistic snapshot of the atmosphere, weather and climate models inevitably develop prediction 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, equally skillful predictions of important midtropospheric and surface variables are possible half-a-day to a day further into the future. 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 issues like limited resolution, inaccurate physical parameterizations, and imperfect initial conditions. Typically, postprocessing 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 (Arribas et al., 2011; Stockdale et al., 1988). 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 atmospheric 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 (DelSole & Hou, 1999; Leith, 1978; Saha, 1992) 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 nonlinear 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 (Brenowitz & Bretherton, 2018; Krasnopolsky et al., 2013; Rasp et al., 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. There is evidence that ML can do effective online bias correction in more idealized contexts (e.g. Watson, 2019). 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 general circulation models (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 the FV3GFS model (Zhou et al., 2019). FV3GFS is based on the FV3 nonhydrostatic dynamical core on a cubed-sphere grid (Putman & Lin, 2007) coupled to physics parameterizations implemented by NOAA’s Environmental Modeling Center. It is a component of NOAA’s Unified Forecast System (UFS Community, 2020) used for operational weather forecasting in the United States. 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 or 127 vertical levels (NOAA, 2018). We use a coarse C48 (~200 km) horizontal resolution with 79 vertical levels (same levels used for global storm-resolving FV3 simulations [Stevens et al., 2019]) and a physics timestep of 15 min. Sea surface temperatures and sea ice extent are specified.

2.2. Nudging Approach

In order to estimate the atmospheric model biases across seasons and the diurnal cycle, we perform a 2-year hindcast simulation in which the prognostic state is continuously nudged toward an observational analysis (Figure 1). Specifically, a linear relaxation term is added to the prognostic equations of certain variables:

\[
\frac{\partial a}{\partial t} = -\nabla \cdot a + Q_a^p - \frac{a - a_{\text{obs}}}{\Delta t},
\]

where \(a\) is a prognostic variable, \(-\nabla \cdot a\) 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), \(a_{\text{obs}}\) is an estimate of the observed value of \(a\) and \(\Delta t\) is a nudging timescale. For the horizontal winds, Equation 1 would have additional terms for Coriolis acceleration and the pressure gradient force, which we omit for simplicity. The nudging tendencies \(\Delta Q_a\) 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 model error throughout the simulation.

The nudging is active for temperature (nudging tendency labeled $\Delta Q_a$), specific humidity ($\Delta Q_s$), horizontal winds ($\Delta uQ$ and $\Delta vQ$), and surface pressure. Surface pressure is not a prognostic variable in the non-hydrostatic FV3GFS model. Therefore, 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. A 6-hour timescale $\tau$ is used for all variables. The observational dataset is the Global Forecasting System (GFS) analysis (NCEI, 2020) on a 1.4° latitude-longitude grid. The analysis was produced using the operational GFS at the simulated time (2015–2016), which used a spectral dynamical core and somewhat different physical parameterizations than our version of FV3GFS. The analysis is available every 6 h and is linearly interpolated to obtain a state in between these times. During the simulation, the analysis is interpolated vertically to the model’s pressure surfaces and horizontally to FV3GFS’s cubed-sphere grid. No nudging is applied to any variable in the top-most model level and no nudging is applied for specific humidity above 100 hPa.

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. When the moistening due to nudging is larger than the physics precipitation, we set the total precipitation rate to zero:

$$P = \max \left(0, P_{\text{physics}} - \langle \Delta Q_s \rangle\right).$$

where $P_{\text{physics}}$ is the surface precipitation rate produced by the physics parameterizations and

$$\langle \Delta Q_s \rangle = \frac{1}{g} \int_{0}^{\infty} \Delta Q_s dp.$$  \hspace{1cm} (3)

The clipping at zero in Equation 2 acts a moisture source for the coupled land-atmosphere system with consequences described in the discussion section. Pressure thickness in FV3 includes the mass of dry air and all water species. Therefore, when specific humidity is adjusted by nudging, a corresponding change is made to the pressure thickness to conserve dry air mass.

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.

Over the ocean, sea-surface temperature from the GFS analysis dataset is prescribed. 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 depend on the vertical level and the others are scalars. Although surface pressure is nudged (via the pressure thickness, see Section 2.2) in the training run, the ML model does not make predictions of these tendencies.

The random forest is trained by minimizing a mean squared error loss function. Outputs are normalized by subtracting the mean across samples and then dividing by the standard deviation. A separate mean and standard deviation is computed for each vertical level to account for differing variances as a function of model level. There is no preprocessing of inputs. 16 individual decision trees form the random forest; each tree has a maximum depth of 13 and is trained using nonoverlapping batches, each one sixteenth of the training data. All other hyperparameters use the defaults defined in scikit-learn v0.22.1.

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 2 months. The effects of the column moisture tendency from the ML on surface precipitation is handled in the same way as the nudging case (Equation 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. The random forest trained for this study requires about 360 MB of memory, which is a substantial addition to the ~600 MB required on each processor to run a baseline version of FV3GFS at C48 resolution, assuming one processor per cubed-sphere tile.

2.5. Experiment Configuration and Validation

A procedure is designed to (a) generate training data from across the seasonal cycle and (b) test the online and offline model skill on a time period independent from the training data. We first perform a 2-year long simulation that is initialized from GFS analysis on January 1, 2015 and continuously nudged toward the GFS analysis, as described in Section 2.2. The nudging tendencies and diagnostic state are saved every 5 h to ensure sampling around the diurnal cycle.

The random forest is trained on output from the first year of the 2-year simulation. Columns from 160 time steps which uniformly span 2015 are used for training, resulting in about 2.2 M samples (all 6·48·48 = 13824 columns are used for each time). Using a greater number of samples for training resulted in a random forest that required more memory to store with limited improvements in performance (not shown). 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 2-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 12 10-day forecasts each starting from the first of the month for every month of 2016 to evaluate the error growth on weather forecasting timescales. Second, we initialize a single year-long run on January 1, 2016 in order to evaluate longer timescale statistics. 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 nudged simulation as truth. For variables which are directly nudged, this is a reasonable representation of the true state of the atmosphere. However, precipitation and 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 $\degree 1 \times 1$ latitude-longitude grid and so for this comparison the model output is regridded from the cubed-sphere using the fregrid tool (https://github.com/NOAA-GFDL/FRE-NCtools).

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 observational analysis dataset is equal to the negative of the time-mean nudging tendency multiplied by the nudging timescale (Equation 1). Therefore, Figures 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 intertropical convergence zone, and in midlatitude fronts (see also Supplemental Movie S1), likely correcting a bias of the convective and microphysics parameterizations to generate insufficient rainfall in realistic conditions for this grid resolution. The imprinting of the cubed-sphere grid in Figure 2a is due to the nudging tendency correcting artifacts introduced by the dynamical core at the coarse C48 resolution.

When evaluated offline on samples from the test data, the random forest successfully predicts the time-mean pattern of heating and moistening (Figures 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. We suspect some of the poorly predicted tendencies result from nudging tendencies which are due to dynamical core errors. For example, upstream and downstream of poorly resolved sharp topography

![Figure 2](image-url)
(e.g., see tip of South America in Figure 2), there are strong heating and cooling nudging tendencies which the ML model, given only a column profile of the atmospheric state as input, does not predict.

To evaluate the skill of the random forest in making instantaneous predictions of the nudging tendency, Figures S1a–S1b shows the zonal mean $R^2$ skill for the heating and moistening as a function of latitude and pressure. The heating ($\Delta Q_1$) predictions are substantially more skillful than the moistening ($\Delta Q_2$) 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, the Northern Hemisphere mid-latitudes and the polar regions. The humidity nudging predictions are less skillful, with a maximum 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. Furthermore, the similar performance of $R^2$ and bias for the 2015 training data and 2016 test data (Table S1) indicates that the random forest is not overfitting on the training data or 2015 sea surface temperatures.

### 3.2. Online Performance: Weather Skill

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. 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 lower error than the baseline model for all three of these variables at lead times ranging from 1 to 10 days depending on the variable. 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. 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 Figure 3? We trained a random forest to only predict $\Delta Q_1$ and $\Delta Q_2$ and not predict the momentum tendencies (blue lines in Figure S2). We find 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 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, a not insignificant challenge (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 and ideally will have a climate state that is less biased than the baseline model.
We perform a year-long simulation initialized on January 1, 2016. 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. The limiter checks in each grid cell if the specific humidity after adding the increment from the ML would be below zero, and if so reduces the ML tendency to end up with zero specific humidity. 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 (Figure 4). The spatial RMSE of the 2016-mean precipitation substantially decreases by about 23%, from 2.14 mm/day to 1.65 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 toward the GFS analysis is 1.39 mm/day (Figure S3). This is a lower bound on the precipitation RMSE we might expect from the ML-corrected run, suggesting it has realized two-thirds of the greatest possible precipitation bias improvement we might hope to achieve. The reduction of precipitation rate errors mostly arises from the corrective tendencies of temperature and moisture (compare bottom panels of Figure S3).

The baseline model strongly overpredicts precipitation over the Himalaya, Southeast Asia, and the Andes (Figure 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 (Figure S3). The intensity distribution of the precipitation rate (Figure S4) shows that in the ML-corrected run there is too little rainfall at rates less than 1 mm/day and too much at rates between 1 mm/day and 4 mm/day. The ML-corrected run also predicts less precipitation with intensity greater than 10 mm/day rainfall than the baseline model. Above 50 mm/day, this is in better agreement with the observational dataset.

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 (Figure 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 of temperature in the stratospheric and tropospheric polar regions (Figure S6) and related biases in zonal mean zonal wind (not shown).

4. Discussion

The nudging tendencies incorporate 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 (Figure 2 and Supplemental Movie S1) 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 (Figure S7) 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 $\tau$ (Equation 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 di-
urnal 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-h nudging timescale we used provided a bal-
ance between these competing issues and was also a natural choice given the 6-hourly availability of the
GFS analysis.

Ideally, one would apply the ML corrections to 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 toward the
observed state of the atmosphere instead of, for example, the difference between the boundary layer param-
eterizations 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 (e.g., Bonavita &
Laloyaux, 2020; Farchi et al., 2020) and this would ensure consistency between the models.

The coupling between the ML tendencies of the atmosphere and the land surface is important because the
nudging of specific humidity accounts for about a third of the global mean drying of the atmosphere. With-
out adding the column integrated nudging or ML tendency of humidity to the surface precipitation (Equa-
tion 2), there is a strong drying of the land surface globally (not shown). 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 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.

Although conservation of moisture and energy is not strictly imposed by this method, we expect the ML-as-
sisted model to learn tendencies to make it behave similarly to the reference dataset. Indeed, we find that
over the year-long simulation, there is no secular drift in the total atmospheric moist static energy (MSE;
Figure S5d) and furthermore the ML-corrected model more accurately captures the seasonal variations of
MSE than the baseline model.

**5. Conclusions**

We propose a method to perform online bias correction of a GCM using machine learning of nudging ten-
dencies from a hindcast simulation. A random forest is able to make reasonably skillful predictions of the
nudging tendencies using only the atmospheric model state as input. When coupled back to the atmospher-
ic model, the ML-corrected GCM increases its lead-time forecasts for 500 hPa geopotential height and sur-
face 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. The ML correction does not improve all aspects of simulated climate.
It improves the intensity distribution of heavy daily surface precipitation greater than 50 mm/day but gen-
erates excessive light precipitation rates between 1 and 4 mm/day. It also induces significant temperature
biases in the polar lower stratosphere after a number of weeks.

One area for future work is investigating how much this method improves higher-resolution (e.g., opera-
tional 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., 2021).

Due to the use of historical analysis data, the training dataset is restricted to the climate of the last few de-
cades and the proposed method may have limitations for use in climate-change scenarios due to out-of-sam-
ple 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. Using a high-resolution simulation as a target dataset also allows a more careful budget analysis and isolation of physical parameterization errors. We will report on this approach in a forthcoming paper.

Data Availability Statement

The source code for the FV3GFS model is available at https://github.com/ufs-community/ufs-weather-model (doi: https://doi.org/10.5281/zenodo.4460292). The version of FV3GFS used for this work and the code used to do model training and analysis is available at https://github.com/VulcanClimateModeling/nudge-to-obs-manuscript-workflow (doi: https://doi.org/10.5281/zenodo.5116510). GPCPv1.3 data were obtained from NOAA-NCEI (https://doi.org/10.7289/V5RX998Z). GFS analysis data are available at https://www.ncdc.noaa.gov/pmb/products/gfs.

Acknowledgments

The authors thank Vulcan, Inc. for supporting this work. They thank Lucas M. Harris, Larry W. Horowitz, and Jake Huff at GFDL for providing assistance in using the nudging capability of FV3GFS and supplying post-processed GFS analysis data in a form compatible for nudging. They acknowledge NOAA-EMC, NOAA-GFDL and the UFS Community for publicly hosting source code for the FV3GFS model. They thank two anonymous reviewers whose comments greatly improved the manuscript.

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 (pp. 173–265). Elsevier. https://doi.org/10.1016/B978-0-12-460017-7.50009-4

Arrabas, A., Glover, M., Maidens, A., Peterson, K., Gordon, M., MacLachlan, C., & Casuca, S. (2011). The GloSea4 ensemble prediction system for seasonal forecasting. Monthly Weather Review, 139(6), 1891–1910. https://doi.org/10.1175/2010MWR3615.1

Bennah, M., & Lalooyaux, P. (2020). Machine learning for model error inference and correction. Journal of Advances in Modeling Earth Systems, 12(12), e2020MS002232. https://doi.org/10.1029/2020MS002232

Brenowitz, N. D., Beucler, T., Fritchard, M., & Bretherton, C. S. (2020). Interpreting and stabilizing machine-learning parameterizations of convection. Journal of the Atmospheric Sciences, 77(12), 4357–4375. https://doi.org/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. https://doi.org/10.1029/2018GL078510

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

Brenowitz, N. D., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., & Bretherton, C. S. (2020b). Machine learning climate model dynamics: Offline versus online performance. Retrieved from https://arxiv.org/abs/2011.03081

Danforth, C. M., Kalnay, E., & Miyoshi, T. (2007). Estimating and correcting global weather model error. Monthly Weather Review, 135(2), 261–299. https://doi.org/10.1175/mwr3289.1

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

De Sole, T., Zhao, M., Dirmeyer, P. A., & Kirtman, B. P. (2008). Empirical correction of a dynamical land-atmosphere model. Monthly Weather Review, 136(11), 4063–4076. https://doi.org/10.1175/2008mwr2344.1

Farchi, A., Lalooyaux, P., Bonavita, M., & Bocquet, M. (2020). Using machine learning to correct model error in data assimilation and forecast applications. Retrieved from https://arxiv.org/abs/2010.12605

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. https://doi.org/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. https://doi.org/10.1175/WAF-D-10-05038.1

Han, J., Wittke, 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. https://doi.org/10.1175/WAF-D-15-0053.1

Harris, L., Zhou, L., Lin, S.-J., Chen, I.-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. https://doi.org/10.1029/2020MS002223

Huffman, G. J., Adler, R. F., Morrissey, M. M., Bolvin, D. T., Curtis, S., Joyce, R., et al. (2001). Global precipitation at one-degree daily resolution from multisatellite observations. Journal of Hydrometeorology, 2(1), 36–50. https://doi.org/10.1175/1525-7541(2001)002<0036:GPAODD>2.0.CO;2

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

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, 2013, 485913. https://doi.org/10.5194/gmd-14-4401-2021

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. https://doi.org/10.1175/MWR2923.1

Leith, C. E. (1978). Objective methods for weather prediction. Annual Review of Fluid Mechanics, 10, 107–128. https://doi.org/10.1146/annurev.fl.10.050178.000543

McGibbon, J., Brenowitz, N. D., Cheeseman, M., Clark, S. K., Dahm, J. P. S., Davis, E. C., et al. (2021). fv3gfs-wrapper: A python wrapper of the FV3GFS atmospheric model. Geoscientific Model Development, 14(7), 4401–4409. https://doi.org/10.5194/gmd-14-4401-2021

McGibbon, J., & Bretherton, C. S. (2019). Single-column emulation of reanalysis of the Northeast Pacific marine boundary layer. Geophysical Research Letters, 46(16), 10063–10069. https://doi.org/10.1029/2019GL083646

NCEI. (2020). Global forecast system. Retrieved from https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/global-forecast-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. 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. 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. https://doi.org/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. https://doi.org/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. https://doi.org/10.1002/qj.23

Saha, S. (1992). Response of the NMC MRF model to systematic error correction within integration. Monthly Weather Review, 120, 345–360. https://doi.org/10.1175/1520-0493(1992)120<0345:rotnmm>2.0.co;2

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. https://doi.org/10.1175/2010BAMS3001.1

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

Stockdale, T., Anderson, D., Alves, J., Balmaseda, M. A. (1988). Global seasonal rainfall forecasts using a coupled ocean-atmosphere model. Nature, 392, 370–373. 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. https://doi.org/10.1002/2017RG000574

Watson, P. A. G. (2019). Applying machine learning to improve simulations of a chaotic dynamical system using empirical error correction. Journal of Advances in Modeling Earth Systems, 11(5), 1402–1417. https://doi.org/10.1029/2018MS001597

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

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

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

Yuval, J., O’Gorman, P. A., & Hill, C. N. (2021). Use of neural networks for stable, accurate and physically consistent parameterization of subgrid atmospheric processes with good performance at reduced precision. Geophysical Research Letters, 48(6), e2020GL091363. https://doi.org/10.1029/2020GL091363

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