State-Dependent Bias Correction Within Convection-Allowing Ensemble Systems using Random Forests

Jacob Peace[1,4], Aaron Johnson[2,3], Xuguang Wang[2,3]

[1]University of Georgia; Athens, Georgia
[2]Multiscale data Assimilation and Predictability (MAP) Lab, University of Oklahoma
[3]Consortium for Advanced Data Assimilation Research and Education (CADRE)
[4]National Weather Center Research Experiences for Undergraduates Program; Norman, Oklahoma

ABSTRACT

High resolution convection-allowing ensemble systems do well to cover a wide range of possible outcomes, as opposed to deterministic model runs, to provide the best forecast. These ensemble systems, however, still struggle to produce accurate forecasts of 2-meter temperature and 2-meter dew point that are free of error, and much of this error can be due to systematic bias rather than random error. Using a method of bias correction, this systematic error can be corrected by simply removing consistent domain average errors from forecasts, which makes modest improvements. More advanced methods are explored in this study, employing the use of machine learning, specifically random forests (RF), to bias correct beyond what has already been done. RF models were trained on 20 historical ensemble forecast cases and validated on 6 independent events using a specific set of predictors, mainly focusing on land-atmosphere interactions. Multiple hyperparameter configurations were tested to identify the most effective model. RF bias correction showed substantial improvement over 2 forecast lead times examined, 6 and 21 hours. Partial dependence plots were used to interpret predictor influence, with soil moisture pressure-related variables, and latitude and longitude emerging as key contributors. These findings suggest that RF-based corrections are skillful in short-term ensemble forecasting in near-surface temperature fields.

1. INTRODUCTION

The advent of convection-allowing models (CAMs) in Numerical Weather Prediction (NWP) has improved short-term forecasting by leaps and bounds, as they are able to finely resolve smaller-scale features within the model output. CAMs, with higher resolution and smaller grid-spacing, have been found to produce better forecasts in the short-term than models with coarser resolution (Schwartz et al. 2022; 2017). Convection Allowing Ensemble (CAE) forecast systems improve upon this by running multiple simulations to produce an output that is more all-encompassing of future weather. CAEs account for forecast uncertainty by sampling sources of initial condition and/or model uncertainty, which enables probabilistic forecasts and can reduce forecast error and biases compared to deterministic forecasts (Berner et al. 2017). Schwartz et al. 2017 also found ensemble forecasts performed better than individual model runs which had a higher resolution, which highlights the need for ensemble-based short-range forecasting.

More specifically, forecasts of 2-meter temperature and 2-meter dew point have major implications, as the accuracy of such has impacts on public and private sector companies, agriculture, aviation, and operational meteorology. CAEs, and NWP models in general, have struggled to produce forecasts that are bias-free (Boallegue et al. 2023), and therefore it is difficult for meteorologists to accurately predict these values consistently. The bias typically stems from two sources: errors with initial conditions and detriments within the model itself, such as background physics schemes (Duda et al. 2017). Model uncertainty is often more challenging to accurately sample in the CAE design, leading to systematic errors (i.e. bias) in all ensemble members. Since biases common to all members contribute to forecast error but not ensemble spread, either the ensemble spread should be artificially inflated or the biases must be identified and removed to achieve well-calibrated ensemble forecasts and optimally performing ensemble-based data assimilation. This study will therefore focus specifically on biases in 2-m temperature and dew

[1] *Jacob Peace, University of Georgia, jcp72795@uga.edu*

point forecasts, which are expected to be closely tied to model physics errors associated with the land-atmosphere coupling process, with main efforts focusing on discovering the sources of this error.

Before causes of bias can be well understood, the first step is to be able to identify and remove the bias within the model, which is systematic in nature. Systematic error, as defined by Bouallegue et al. 2023, is the difference between forecast and observations that can be corrected for by post-processing through bias correction. Taking it a step further from traditional bias correction methods that are often based on simply averaging all past forecast errors, we can employ the use of machine learning. The use of machine learning (ML) in meteorological research has proven to be extremely useful in turbulence (McGovern et al. 2013; Williams 2013), hail (Gagne et al. 2017), climate (Chapman and Berner 2025), and severe weather (Hill et al. 2020) forecasting. ML post-processing in general has shown to be effective in reducing overall model error (Agrawal et al. 2023; Chapman et al. 2019). ML is unique in its ability to identify patterns in a dataset and then discover a connection between such patterns and a result that optimizes an error metric (Gagne et al. 2017). Hamill 2021 tested multiple ML methods in bias reduction and found that each ML model used reduced bias relative to raw model guidance. The same approach can be applied to forecasts of 2-m temperature and dew point, which models generally struggle with, due to challenges modeling land-atmosphere coupling and its uncertainty.

This study will attempt to confirm the recent successes ML—and more specifically, Random Forest (RF) ML—has in identifying systematic errors, or bias, in the context of CAE 2m temperature and dew point forecasts, as well as its ability to relate regional and flow-dependent factors to forecast model bias. Secondly, we seek to improve forecasts of 2-m temperature and 2-m dewpoint by removing those biases. Finally, we will investigate the RF models themselves to determine which regional or flow-dependent factors are contributing to the forecast bias. Doing so will not only seek to improve the accuracy of model forecasts but also help to guide future model—and model physics—improvements.

2. METHODS AND DATA

The FV3-LAM short-range CAE was selected for the purposes of this study. The model, which runs at a 3-kilometer (3-km) resolution over a domain encompassing the Continental United States (CONUS) was initialized within the Oklahoma University Multiscale Data Assimilation and Predictability (MAP) Lab as a part of the National Oceanic and Atmospheric Association's (NOAA) Hazardous Weather Testbed (HWT) experiment in spring of 2022. For 26 days during this time period, taking place over the months of May and early June, the model was run almost every day at 00z. The ensemble-based system consisted of 10 members that produced forecasts of 2-meter (2-m) temperature and 2-m dew point, among other variables. The underlying physics parameterizations rely on Thompson Microphysics (Thompson and Eidhammer 2014), the MYNN-EDMF boundary and surface layer parameterization (Nakanishi and Niino 2009; Olson et al. 2019), and the RUC-LSM scheme (Benjamin et al. 2004). A more detailed description of the inner workings of the model are detailed in the figure below, which is Table 1 from Gasperoni et al. 2023.

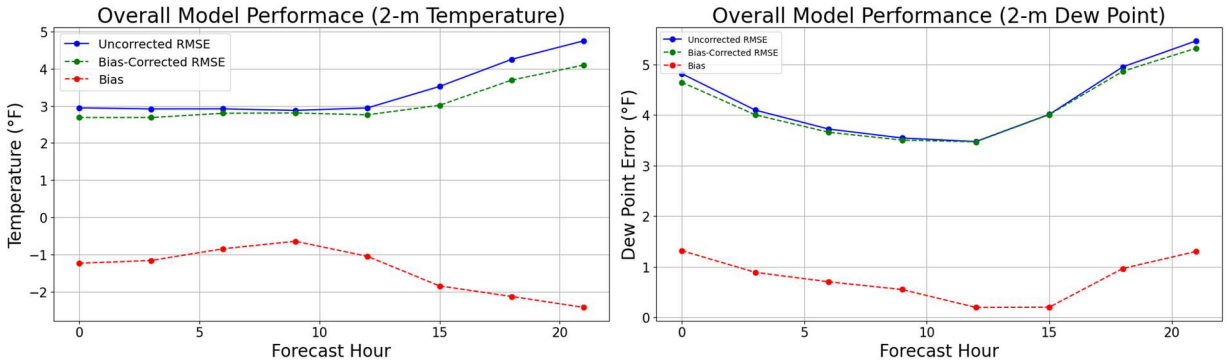
Physics type	Scheme	References
Microphysics	Thompson	Thompson and Eidhammer (2014)
Boundary and surface layer	MYNN-EDMF	Nakanishi and Niino (2009); Olson et al. (2019)
Shortwave/longwave radiation	RRTMG	Mlawer et al. (1997); Iacono et al. (2008)
Land surface model	RUC LSM	Benjamin et al. (2004)
Orographic drag	HRRR gravity wave drag	Dowell et al. (2022)

Table 1. Background physics parameterizations categorizing the FV3-LAM model from Gasperoni et al. 2023.

The forecasts of 2-m temperature and dew point will be verified against Real-Time Mesoscale Analysis (RTMA), an analysis tool which combines short-term model forecasts with recorded observations using data assimilation methods (De Pondeca et al. 2011; Pondeca et al. 2015), that give an estimation of atmospheric conditions at any given time. Recent forecasts from the 3-km

North American Model (NAM), 1-hour forecasts from the 3-km High-Resolution Rapid Refresh Model (HRRR), and 1-hour forecasts from the 13-km Rapid Refresh Model (RAP) are included in the generation of this analysis.

It is important to first identify the overall model performance prior to applying ML post-processing. From this, a consistent bias within the model was determined and removed through domain-average bias correction. Bias correction is simply the process of removing the systematic error from the model outputs. One way of determining the bias is to average error over all cases across the domain (Figure 1. red line). The outputs of 2-m temperature and dew point from the 10 ensemble members from the FV3-LAM were averaged to generate an ensemble mean forecast for a 21-hour lead time at 3-hour increments for a particular initialization of the model, in order to focus on the error common to all ensemble members. This was then averaged over 26 all forecast dates. A root mean squared error (RMSE) analysis was performed on these data to quantify the difference between the forecasted values and the RTMA analyzed values—i.e. the accuracy of the model—and plotted for each lead time. The average bias was also calculated and plotted for each lead time. Both values are averaged for the entire dataset and domain. Finally, an improvement in RMSE was realized through bias correction. By simply removing the domain average bias from each forecast value, we can make an improvement to the accuracy of the forecast. Since bias is a consistent error within the model, we can shift the output of the ensemble mean to account for this error, yielding a more accurate forecast before verifying it against any analysis or observations. Performing yet another RMSE analysis on these data will provide a value of bias-corrected RMSE that is smaller than the uncorrected RMSE value. This is showcased in figure 1.a and 1.b, where all three of these values are plotted for each lead time. Formulas for each value are also defined below Figure 1.a and 1.b.



Figures 1.a and 1.b. Domain average uncorrected RMSE (blue), bias-corrected RMSE (green) and bias (red) for increasing lead times out to 21 hours in 3-hour increments averaged over all 26 forecast dates. A noticeable increase in model accuracy is noted in both cases by bias correction (as noted by the downward shift of the plot). Negative values of bias indicate that the model typically underpredicts the observations (forecast is cooler than analysis), and positive values demonstrate an overprediction of the forecast compared to analysis.

$$RMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{k}(forecast_k - observation_k)^2}$$

Formula 1. Calculation of uncorrected model root-mean squared error

$$BIAS = \frac{1}{K}\sum_{k=1}^{k}(forecast_k - observation_k)$$

Formula 2. Calculation of model bias

$$BCRMSE = \sqrt{\frac{1}{K}\sum_{k=1}^{k}(fcst_k - BIAS - obs_k)^2}$$

Formula 3. Calculation of bias-corrected root-mean squared error

In contrast to a domain-average estimate of bias, ML methods are an attractive option to account for the potentially nonlinear interactions among geographic regions and meteorological variables in determining systematic error. In

particular, the Random Forest (RF) machine learning (ML) model (Breiman 2001) is a collection of decision trees built on random data and features, which it then combines to improve prediction accuracy. At each branch of the tree-growing process, a different subset of random variables is evaluated for inclusion into the tree, which ensures the independence of each tree (Gagne et al. 2017). Each tree learns to make predictions based on different sets of data, and then the forest aggregates them. RF is a powerful tool for model error reduction, because of its ability to pick out nonlinear relationships that would go unnoticed by human eyes. Also, there is a strong need for regional, and flow-dependent model bias correction (Reynolds et al. 2022), and RF ML has proven successful in doing so (Pham et al. 2021). Finally, RF has great skill in a state-dependent approach to bias correction, which is the process of relating background flow patterns (such as cloud cover, wind direction, or soil temperature) (Chapman and Berner 2025).

For the purposes of this project, we will employ the use of the Scikit-Learn Python Library, and the RandomForestRegressor (Pedregosa et al. 2011). To do this, the RF was trained on a selection of so-called predictors, or meteorological variables within the model that potentially cause error. Previous work has shown the importance of multiple-predictor ML methods in calibrating models (Gagne et al. 2014) such as the one included in this research. A brief qualitative investigation of 3-4 forecast cases was done, which involved a comparison of forecast error to the background synoptic regime for a particular day. From this, we determined a set of predictors that may potentially be related to the systematic forecast error. We especially focus on land-atmosphere interactions during the selection of predictors because these generally have the greatest impact on 2-m temperature and dew point observations, but we include other interactions that may play a role in the determination of these values. Table 2 below provides the set of predictors used to train the RF model. Out of the 26 days in the 2022 HWT Spring Experiment, 20 were used to train the RF based upon the predictors, with 6 cases (randomly spread across the experiment period) being left aside to apply the trained model, and a verification process took place to evaluate the accuracy of the RF as compared to the observed error from these 6 validation cases.

List of predictors used to train the RF

- 2-meter Temperature
- 2-meter Dew Point
- U-wind (East/West) at 10 meters
- V-wind (North/South) at 10 meters
- Surface Radar Reflectivity
- Latitude
- Longitude
- Soil Water
- Total Precipitation Accumulation
- 1-Hour Precipitation Accumulation
- Cloud Cover
- Cloud Top Temperature
- Sea Level Pressure
- Soil Temperature

Table 2. List of predictors used to train the RF

RF models are typically trained using a specific set of hyperparameters, which includes a set number of decision trees within the forest, the maximum depth each tree is allowed to grow, the minimum number of samples each leaf node must have after being split, and the maximum number of features to take into account to make the best split of data. Tweaking the hyperparameters will yield differing results when being applied to the same case, so a multitude of RF trainings were completed to deduce which set of hyperparameters were most successful. Only maximum tree depth and minimum samples per leaf were altered for testing; forest size was kept constant at 100 decision trees for each training case. Table 3 below lists out the different combinations of hyperparameters for which the RF was trained on. The RF was trained on these 9 possibilities for forecast lead times of 6 and 21 hours for both 2-m temperature and 2-m dew point, creating 36 trained RF models. This was done to compare the model's performance in a daytime versus a nighttime environment, as well as to visualize how the model performed as lead times increased. The chosen range of hyperparameters from Table 3 was guided by preliminary experiments at early stages of the project which indicated error consistently increased for larger values of minimum samples and smaller values of maximum depth.

Hyperparameters used for RF training	
Maximum Depth	Minimum Samples per Leaf
• 10	• 10
• 10	• 30
• 10	• 100
• 15	• 10
• 15	• 30
• 15	• 100
• 20	• 10
• 20	• 30
• 20	• 100

Table 3. List of combinations of maximum depth and minimum samples per leaf hyperparameters used to train the RF models during the hyperparameter tuning process

Finally, diagnostic statistics were run to examine the trained RF model to determine which predictors were most influential in the training and application of the ML model, as well as the regional and state dependence of the forecast error. This was done on the trained model that performed best compared to the others at each forecast lead time. A test of variable impurity importance (McGovern et al. 2019) was done to determine how much each input feature contributed to the training of the RF and therefore how much each input feature contributes to the RF-based bias. After training, an importance score was calculated for each model, the formula for which is below in figure 3. $I\ (p)$ is the importance score of a given predictor, T is the total number of trees in the ensemble, $s \in S_{t,p}$ is all the splits in tree t where p is used, $\Delta i\ (s)$ is the decrease in impurity due to split s, N is the total number of training samples, and N_s is the number of samples that reached node s. Features with higher values of $I\ (p)$ contribute most to reducing the impurity across the data because they either split the data more effectively or closer to the root of the tree.

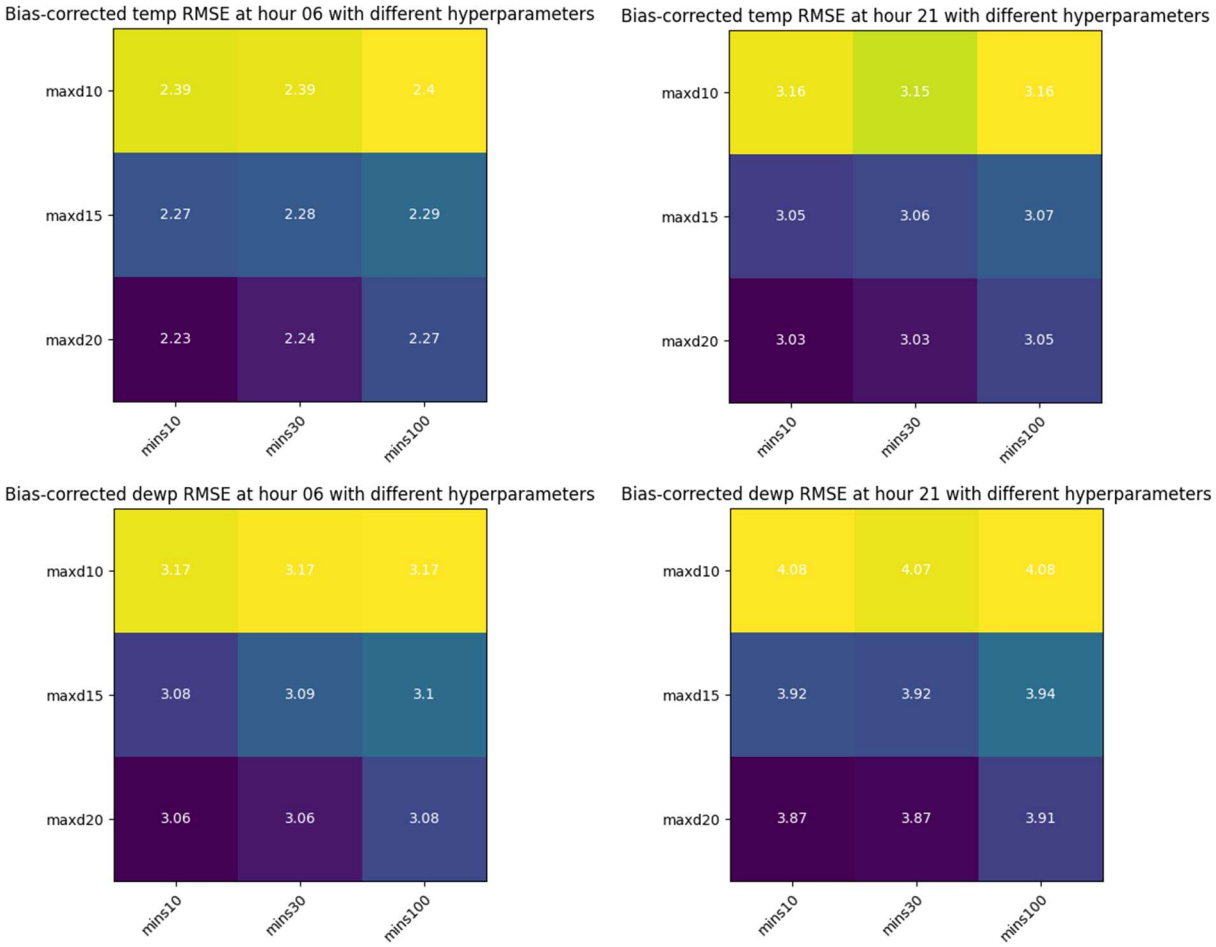
$$I\ (p) = \frac{1}{T}\sum_{t=1}^{T}\sum_{s\in S_{t,p}}\left(\frac{N_s}{N}\right)\ \Delta i\ (s)$$

Formula 4. Importance score formula

3. RESULTS

3.1 Hyperparameter Tuning Results

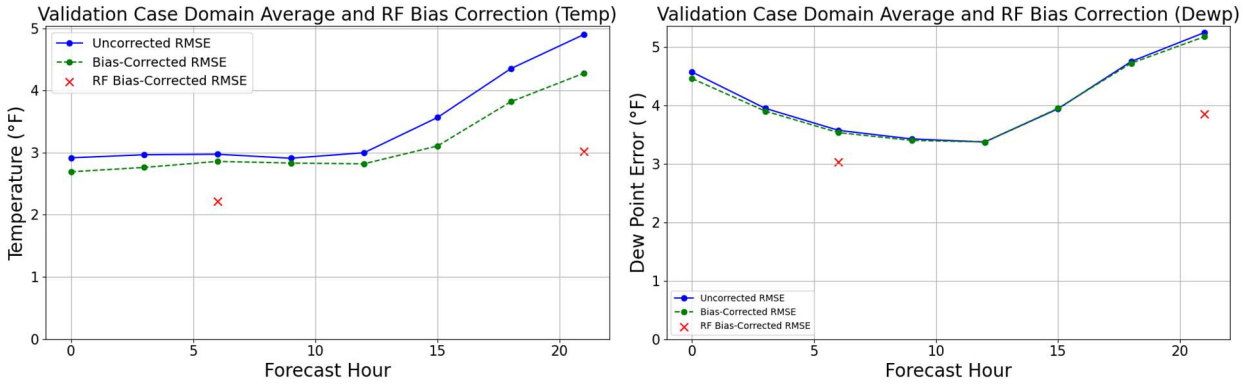
Upon the application of the trained RF models, each model was tested to determine which model had the best performance relative to the others for 2-m temperature and 2-m dewpoint at 6-hour and 21-hour forecast lead times. Each individual model included an output of its RF-based estimate of bias, and a bias corrected output of the forecast error. Using a similar formula to bias corrected RMSE from Formula 3, only changing the domain-average estimate of bias to the trained-RF estimate of bias, an analysis was performed on these to determine which model was best at reducing error through bias correction—i.e. which model generated the lowest bias-corrected RMSE value. Four heatmaps were created to help visualize which set of hyperparameters used to train the RF was most successful in reducing RMSE, one for both 2-m temperature and 2-m dew point at both 6 and 21 hour forecast times. Figures 2.a-d below includes these heatmaps, along with RF bias corrected RMSE values, with cooler colors corresponding to lower error and warmer colors being higher error. For all cases, it appears that training the model on a higher value of maximum depth increases the skill of the RF, but increasing the minimum number of samples per node had either a slightly negative impact, or no impact at all. Overall, the best performing RF model for all four cases used a max depth value of 20 and a minimum samples per leaf value of 10, which produced the lowest RMSE values across the range of hyperparameters considered.



Figures 2.a-d. Heatmaps showcasing the combination of hyperparameters with the best performance for temperature and dew point at forecast hour 6 and 21. Notably, in all four test cases, the RF performed best when the max depth was set equal to 20, and the minimum samples per leaf was set equal to 10.

3.2 Overall RF Performance

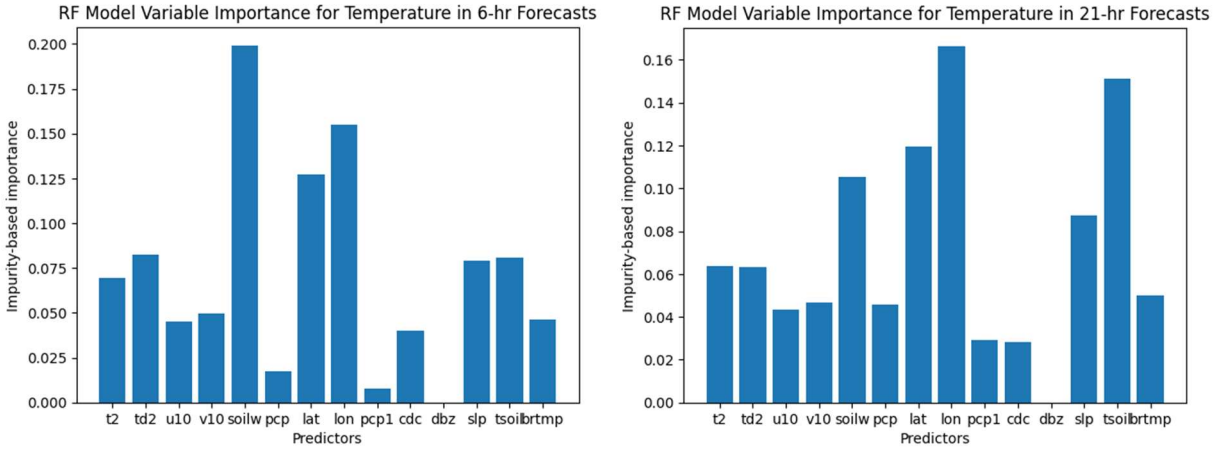
For the sake of comparing apples to apples, model performance was reevaluated to analyze uncorrected RMSE for just the 6 validation cases, rather than the entire dataset, which was then bias corrected using the 20 cases outside the validation cases. This was done to compare RF-based bias correction to domain average bias correction, because the RF models were trained to relate predictors to forecast error for the 20 training cases and then validated against the 6 validation cases without having been exposed to them prior. Figures 3.a and 3.b showcase these two trends with increasing lead time, as well as a comparison to the RF postprocessed RMSE, represented by the two red x's. For both 2-meter temperature and dew point, RF postprocessing outperformed general domain average bias reduction substantially at both 6-hour and 21-hour lead times. For temperature forecasts, domain average bias correction reduced RMSE modestly, creating a 3.9% reduction at hour 6 and a 12.7% reduction at forecast hour 21. RF performed significantly better, reducing RMSE by 25.4% at the 6-hour lead time and 38.4% in hour 21 forecasts. Similarly for dew point forecasts, domain average bias correction provided marginal improvements; a mere 1.1% at hour 6 and 1.3% at hour 21, but RF was far more successful in reducing forecast model error. RF reduced RMSE by 15.2% in 6-hour forecasts and 26.6% at the 21-hour lead time.

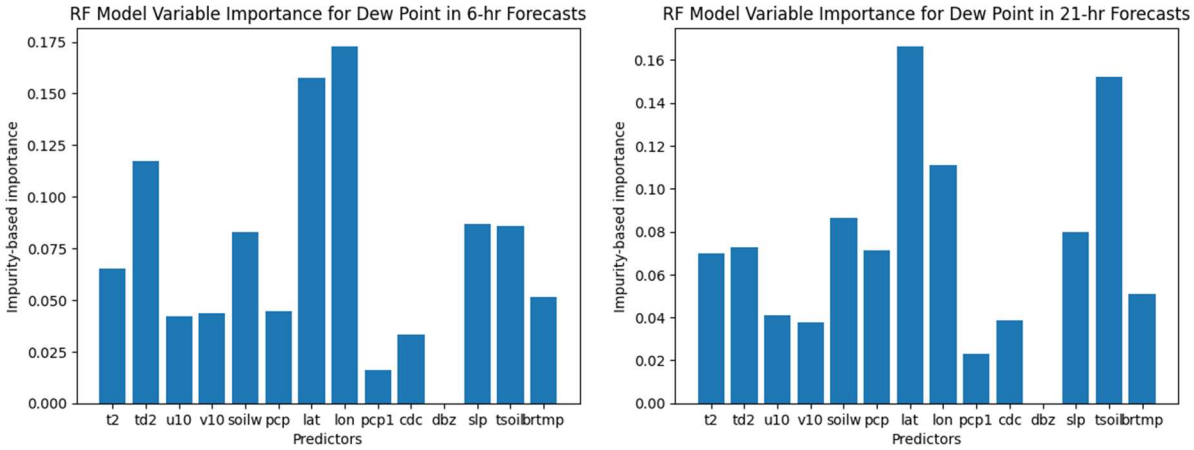


Figures 3.a and 3.b. Plots displaying RMSE of 6 validation cases (blue), which is domain average bias corrected by 20 training cases (green), similarly to how the RF was trained. Red x's represent the RF postprocessed bias corrected RMSE.

3.3 RF Model Impurity Importance

Upon training the RF models, an impurity-based importance score was calculated for each predictor to determine which variables had the greatest impact on the output of the trained RF models. Figures 4.a-d includes the results of this analysis, with higher bars indicating higher feature importance. It is clear to see that latitude and longitude have a major impact, implying that geographic location is a strong predictor for both 2-m temperature and 2-m dew point error at all lead times. This also solidifies the fact that there are regional dependencies in temperature and dew point forecasts, or they are geographically patterned, which the RF was able to pick up on. Soil water appears to be a common indicator of model bias across all cases, but especially for 2-m temperature at forecast hour 6. Subsurface thermal conditions also appear to be correlated with forecast accuracy, especially at the 21-hour lead time. 2-m temperature, 2-m dew point, and sea level pressure are moderately important, and these variables contribute somewhat to explaining where bias occurs, suggesting the bias in the forecast model systematically depends on the local meteorological conditions in addition to geographic location. Other features, such as precipitation accumulations (total and 1-hour), cloud cover, and wind direction appear to be less indicative of forecast bias, perhaps because the forecast model handles these variables well, or their effect on temperature and dewpoint forecasts is negligible, random (i.e. not systematic and predictable), or dependent on other variables not included in our model.



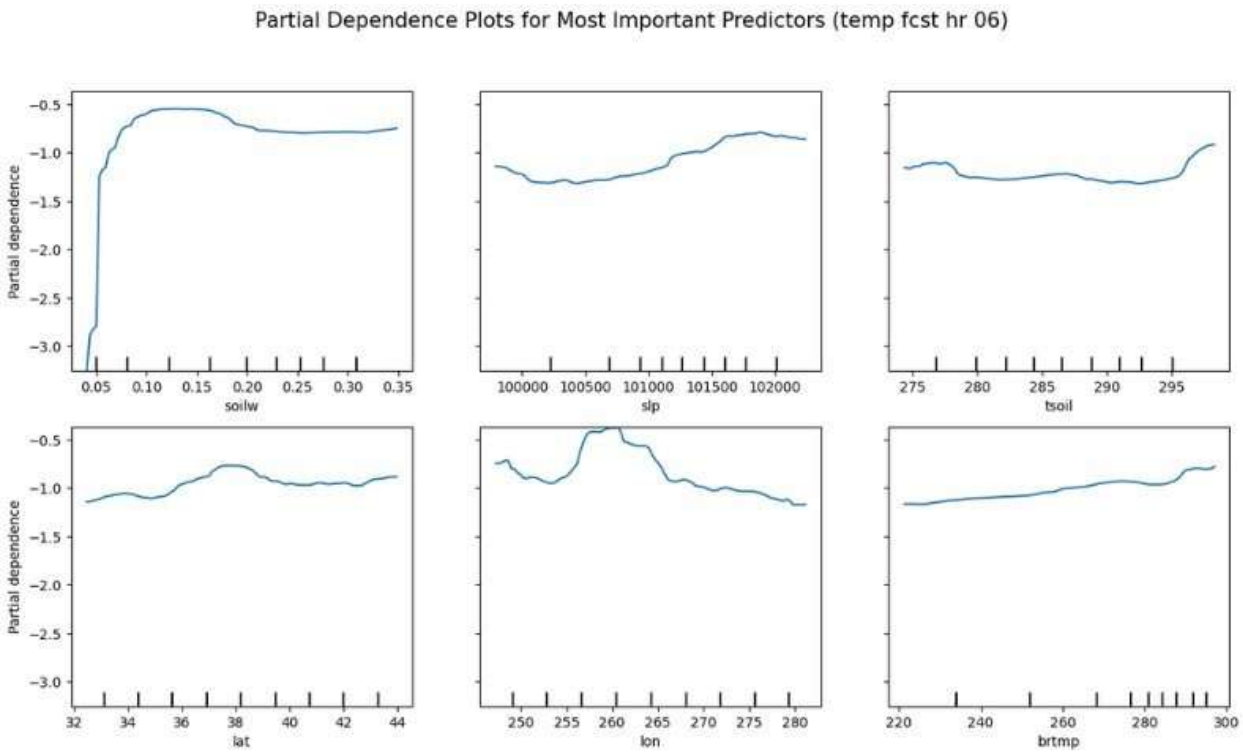


Figures 4a-d. Bar graphs displaying impurity-based variable importance from each RF. Higher bars indicate predictors that were more effective in splitting the data and contributed more to reducing impurity across the trees.

Following sections of this paper will discuss results regarding both 1-dimensional (1-D) and 2-dimensional (2-D) partial dependence plots. 1-D plots show how RF's prediction of forecast bias changes (y-axis) as a certain meteorological variable is altered (x-axis), while holding all other predictors in consideration constant. Similarly, the 2-D plots compare two meteorological variables whose values vary along the x and y-axis, depending on the feature, but the RF predicted bias is plotted spatially, rather than along an axis.

3.4 Partial Dependence Results for Temperature Forecast Hour 6

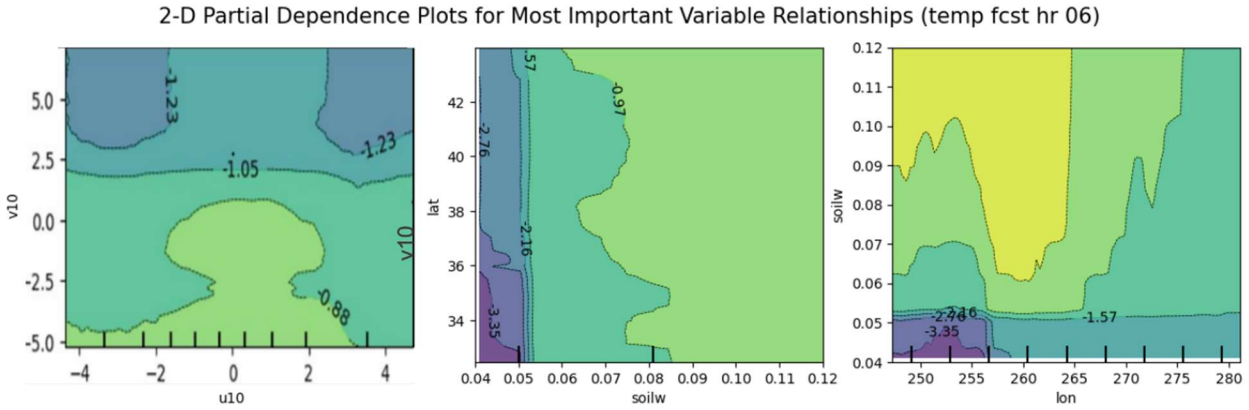
Figures 5a-f below includes 1-D partial dependence plots of the top six predictors from impurity importance calculations for the 6-hour forecast of temperature. The forecast model performs very poorly when soil water contents are very low, or dry, as the RF predicts very high values of bias in these cases, with lower values of bias as water content is increased. Figure 5.b implies that lower sea level pressures are correlated with slightly higher RF predicted bias, indicating that synoptic low-pressure regimes could be more difficult to model. The consistent positive slope of partial dependence for cloud top brightness temperature (brtmp) in Figure 5.f indicates that cooler cloud tops are linked with greater error as compared to higher temperatures, which more represent clear-sky conditions. This implies that the forecast model may struggle to forecast temperature at this lead time when convective cloud cover is present. The overall flatness of Figure 5.c, plotting the soil temperature variable, illustrates its lack of impact on error, and that the model does not strongly depend on this variable in isolation. Finally, looking at latitude and longitude as predictors (Figures 5.d and 5.e), we can better detect the regional dependence of forecast bias within our forecast model.



Figures 5a-f. 1-Dimensional Partial Dependence Plots for the six most important predictors, based upon results from impurity importance tests, for 2-m temperature at 6-hour lead time. Figures a-f, going in order, are plots 1-D partial dependence plots of soil water, sea level pressure, soil temperature, latitude, longitude, and cloud top brightness temperature.

As mentioned before, RF is beneficial because of its ability to pick up on nonlinear relationships between seemingly unrelated meteorological variables. Figures 6.a-c shows a group of those relationships, with the first between u- and v-wind speeds in Figure 6.a, showing that forecasted RF bias for 2-meter temperature appears slightly higher (more negative, in this case) in southerly flow events, with the inclusion of increasingly strong (positive and negative) u-winds generating even more forecasted bias. This implies that the forecast model generally struggles in predicting temperature in strong southeasterly and southwesterly flow patterns across the entire domain. These two flow regimes generally correspond to warm-air advection events, so perhaps the forecast model struggles with modeling these scenarios as compared to cold-air advection events at this forecast lead time. The forecast model appears to perform best when the u-component of wind is close to zero, or a due north/south flow pattern is present, since RF predicted forecast bias tends to increase in the positive and negative directions as the velocity of the u-wind component increases (Note the symmetry along the u10=0 axis).

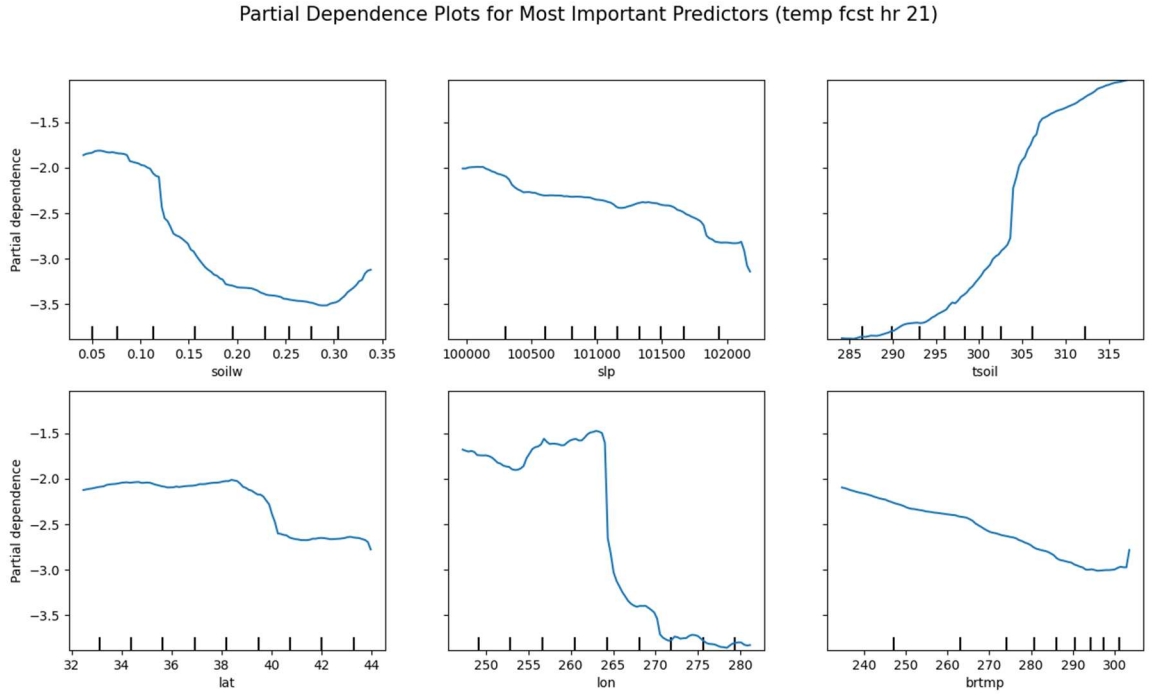
The most important predictor at this forecast hour, soil water, is also found to be dependent on geographic location. Figures 6.b and 6.c detail how dry soil moisture induced bias varies spatially. The highest RF predicted biases tend to be in both low longitudes (west) and low latitudes (south). Though the model struggles with dry soil water content at this forecast hour across the entire domain, this problem is exacerbated in these regions specifically, highlighting the desert southwest as a trouble spot using this predictor.



Figures 6.a-c 2-dimensional partial dependence plots for u- and v-wind (7.a), soil water and latitude (7.b) and soil water and longitude (7.c)

3.5 Partial Dependence Results for Temperature Forecast Hour 21

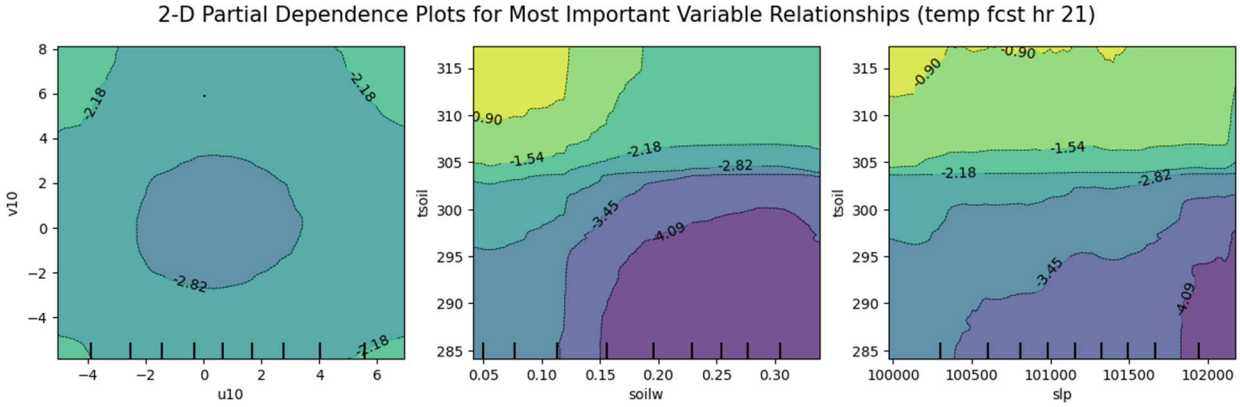
Soil water (Figure 7.a) is still an important predictor of RF forecast temperature bias at the 21-hour lead time, as the forecast model struggles with forecasting temperature when soil water content is higher (> 0.15), dissimilar from earlier forecast lead times, which struggle with drier soil conditions. Most glaringly, there is a very strong dependence on soil temperature at this lead time, higher RF predicted bias values as soil temperatures decrease below approximately 293K (68F), indicating that the forecast model significantly underpredicts 2-m temperature below this threshold. Another strong dependency lies upon east/west extent of the domain, with a sharp drop-off in forecast model accuracy—and a sharp increase in systematic error—east of longitude 264 (as seen in Figure 7.e). Finally, slight dependencies reside in sea level pressure (Figure 7.b) and cloud top brightness temperature (Figure 7.f) at this lead time. They are rather modest but could imply that the model struggles to model 2-m temperature with higher values of each (high-pressure regimes, and little to no cloud cover, respectively), typically producing cold biases.



Figures 7.a-f. 1-Dimensional Partial Dependence Plots for the six most important predictors, based upon results from impurity importance tests, for 2-m temperature at 21-hour lead time. Figures a-f, going in

order, are plots of soil water, sea level pressure, soil temperature, latitude, longitude, and cloud top brightness temperature.

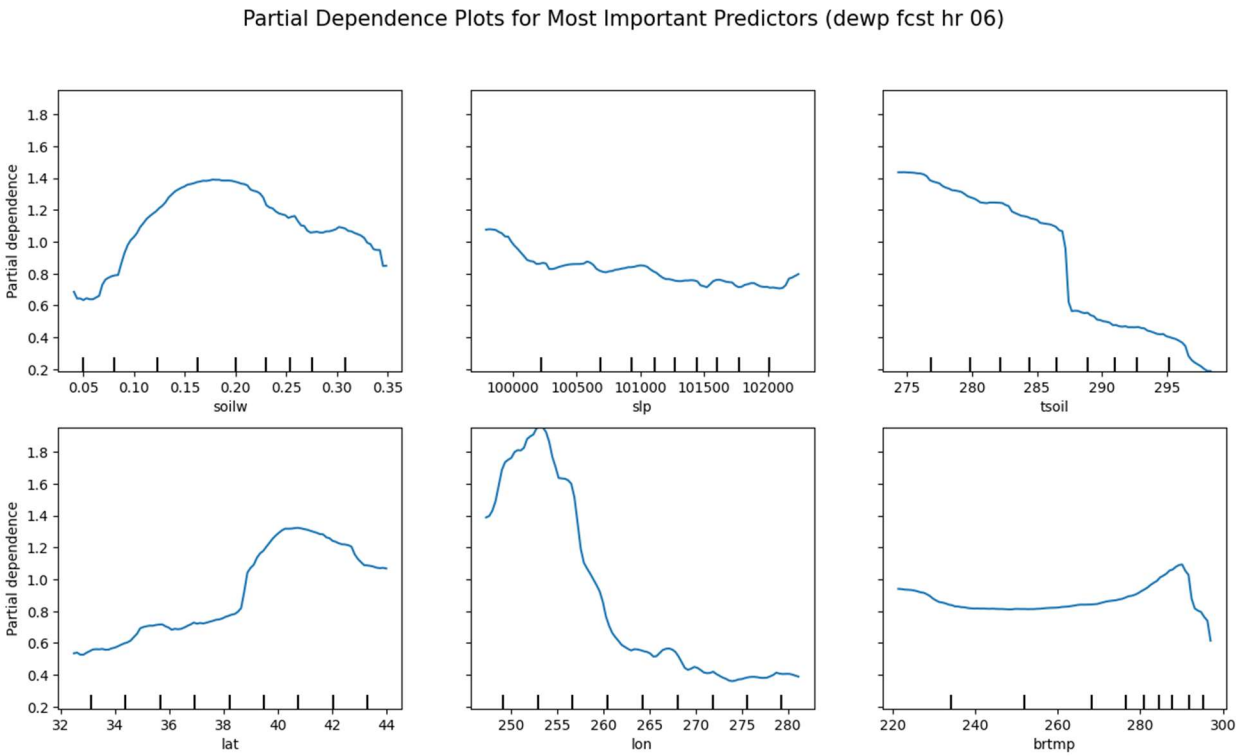
Figure 8.a below illustrates another relationship between u and v-wind components, but notes that the highest forecasted bias by the RF is when wind speeds are calm, or below around 3 knots. Predicted bias begins to decrease slightly by increasing the wind speeds in any given direction. Given that the forecast model initializes at 00z, the 21-hour forecast falls within late afternoon hours where peak heating typically occurs. A prevailing hypothesis for this bias is that the forecast model may struggle with the extent of afternoon surface heating when wind conditions are calm, as the forecast model often underpredicts 2-m temperature at this time. Figure 8.b highlights another relationship between soil temperature and soil water content, noting that there is a strong cold bias within our model when soil moisture is higher and soil temperature is lower. Similarly, another strong cold bias appears in high pressure regimes that are co-located with cool soil temperatures (Figure 8.c).



Figures 8a-c. 2-D partial dependence plots of u and v-wind components (9.a), soil water and soil temperature (9.b) and sea level pressure and soil temperature (9.c).

3.6 Partial Dependence Results for Dew Point Forecast Hour 6

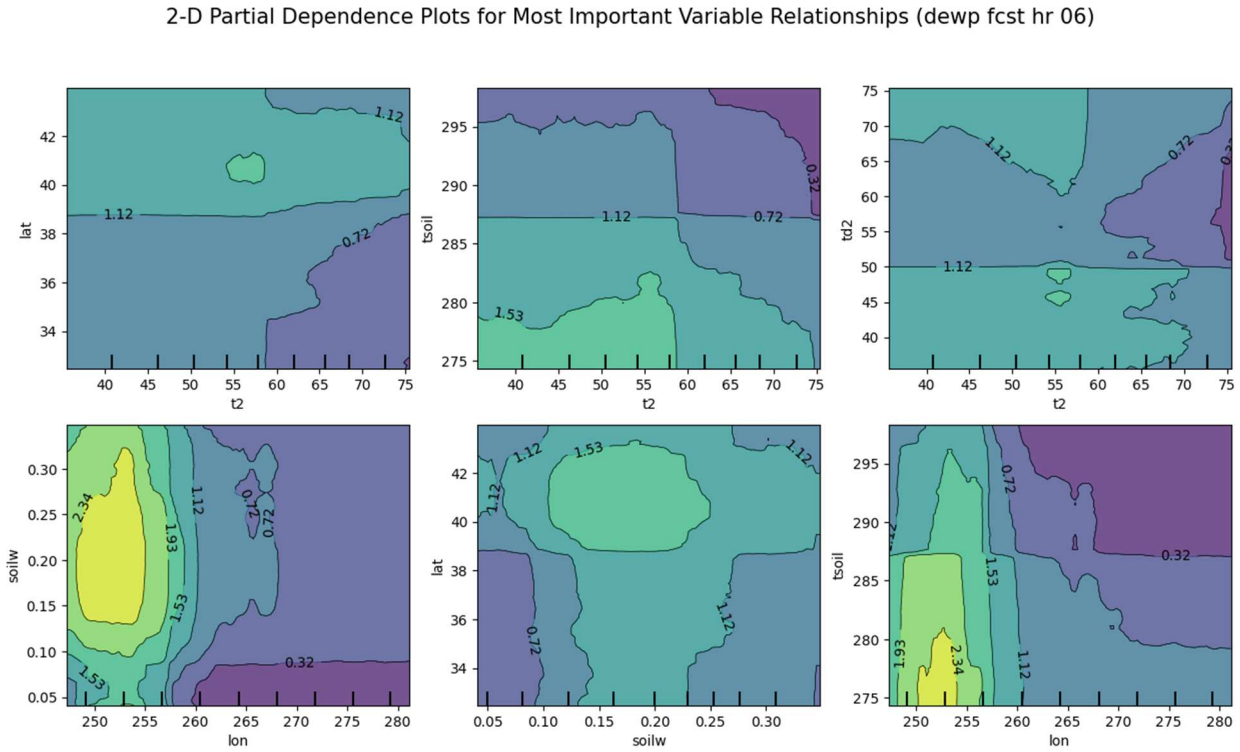
Looking now at the same predictors for 2-m dew point at the hour 6 lead time in Figures 9.a-f, the forecast model continues to struggle with these forecasts depending on soil moisture content (Figure 9.a) and soil temperature (Figure 9.c), as well as geographic location (Figures 9.d and 9.e). Focusing on the first, RF predicted bias becomes more prevalent as soil water moves beyond 0.10, and predicted bias is highest in the middle range of soil moisture (0.10-0.25). This could be because the forecast model incorrectly predicts the extent to which water is evaporated from the soil, resulting in higher dew point forecasts than observations. Forecast model accuracy decreases significantly with decreasing soil temperatures, especially below 287K, where the forecast model increasingly overpredicts the dew point temperature at this forecast hour. Forecast model dew point forecast accuracy is also highly dependent on geographic location, with forecasts being significantly worse at lower longitudes (west) and higher latitudes (north). Finally, varying values of sea level pressure (Figure 9.b) and cloud top brightness temperature (Figure 9.f) yields no significant changes in RF forecasted bias, as they are relatively flat, so the forecast model does not depend on these variables in isolation, though a slight relationship to forecast bias could be present at lower sea level pressures or at the peak in RF forecasted bias at around 290K cloud top brightness temperature.



Figures 9.a-f. 1-Dimensional Partial Dependence Plots for the six most important predictors, based upon results from impurity importance tests, for 2-m dew point at 6-hour lead time. Figures a-f, going in order, are plots of soil water, sea level pressure, soil temperature, latitude, longitude, and cloud top brightness temperature.

Switching gears to analyze relationships between the features, we note a maximum in RF predicted bias across most forecasted temperatures, but latitudes 39-43N typically tend to have larger systematic errors in this relationship (Figure 10.a). Higher forecasted temperatures and lower latitudes tend to be a strength of the forecast model. Another dependence exists between 2-m temperature and soil temperature features (Figure 10.b), where RF predicted bias tends to increase as air temperature and soil temperature decrease, hinting that the forecast model overpredicts surface moisture flux in conditions with limited evaporation. At lower forecasted temperatures and dew points at 2 meters, where the forecast model tends to overpredict the dew point in these forecasts (Figure 10.c). For this same relationship, the forecast model tends to perform best in a window where temperatures are at the upper extreme of the range at this hour (75F), and surface moisture conditions are higher.

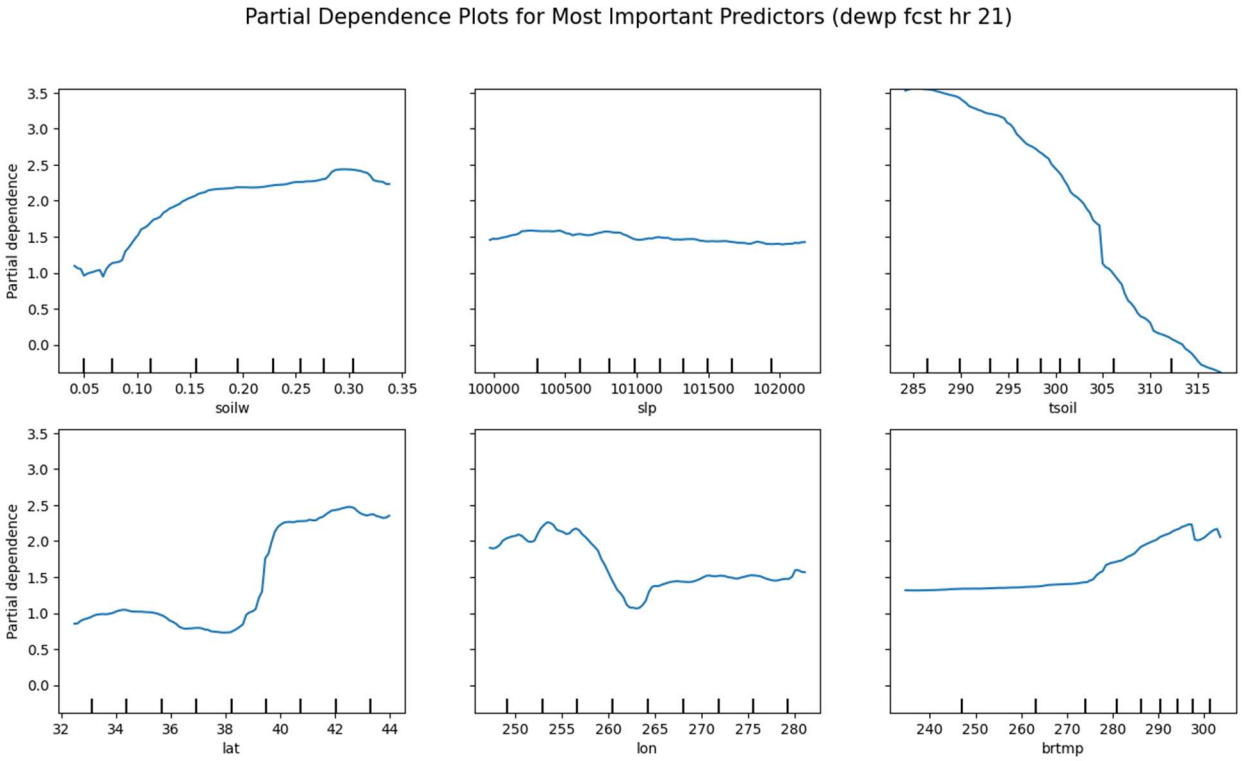
As we compare the regional differences for soil moisture content for 6-hour forecasts of dew point to 6-hour forecasts of temperature, we can see that the highest RF predicted bias is highest at very similar longitudes (Figure 10.d) despite having differing levels of accuracy under different amounts of soil water. However, a difference lies in the latitudinal dependence, as the forecast model overpredicts dew points the most at latitudes north of 40N, whereas temperature forecasts are too cool at lower latitudes (36N and south) (Figure 10.e). Finally, dew point forecasts are too warm in western portions of the domain, specifically when soil temperatures are below 280K (Figure 10.f).



Figures 10.a-f. 2-Dimensional partial dependence plots of 2-m temperature and latitude (11.a), 2-m temperature and soil temperature (11.b), 2-m temperature and 2-m dew point (11.c), soil water and longitude (11.d), soil water and latitude (11.e), and soil temperature and longitude (11.f).

3.7 Partial Dependence Results for Dew Point Forecast Hour 21

Consistent with Figure 9.a, the forecast model continues to struggle with the mid-range (0.10-0.25) for soil water at forecast hour 21, but in increasing the lead time, so has the RF predicted bias on the high end (0.25-0.35), whereas in Figure 9.a there was a slight increase in RF predicted bias. Trends from sea level pressure (Figure 11.b) and cloud top brightness temperature (Figure 11.f) have remained constant with bias only increasing due to increased lead time. Similarly to Figure 7.c, forecast model bias is highly dependent on soil temperature at the 21-hour forecast time, with effects amplifying significantly. Continuing the trend from hour 6, cooler soil temperatures are responsible for great overestimations in forecasted dew point temperature. Geographic dependencies seemed to remain consistent between forecast hours, with differences coming from increased biases with longer lead times.



Figures 11.a-f. 1-Dimensional Partial Dependence Plots for the six most important predictors, based upon results from impurity importance tests, for 2-m dew point at 21-hour lead time. Figures a-f, going in order, are plots of soil water, sea level pressure, soil temperature, latitude, longitude, and cloud top brightness temperature.

RF predicted bias appears to increase at the 21-hour lead time as soil water increases and soil temperature decreases, finding a consistent hot bias under these conditions (Figure 12.a). The RF also found a significant relationship between soil temperature and cloud top brightness temperature (Figure 12.b) for dew point forecasts at this lead time, displaying a hot bias under cooler soil temperatures with high cloud top brightness temperatures (which correlates with clear air/weak and shallow cloud cover). The model overpredicts dew point in these conditions likely because it overestimates near-surface moisture under shallow clouds and cold soils that suppress evaporation and vertical mixing. Finally, a trend we saw from hour 6 dewpoint forecasts, 21-hour forecasts continue to show low soil temperature regimes struggle in western portions of the domain (Figure 12.c), which indicates a consistent increase of this bias with increasing lead times.

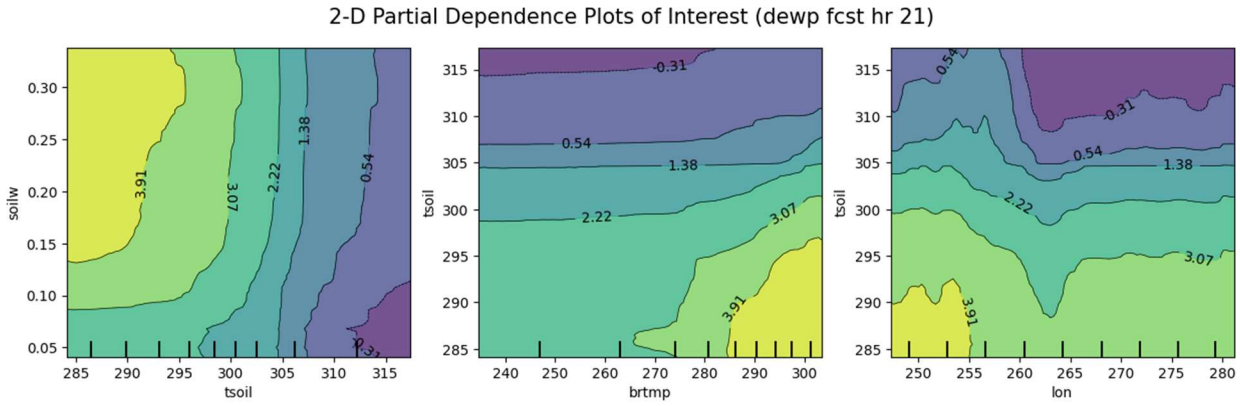


Figure 12.a-c. 2-Dimensional partial dependence plots of soil water and soil temperature (13.a), cloud top brightness temperature and soil temperature (13.b), and longitude and soil temperature (13.c)

3.8 Latitude and Longitude as Predictors

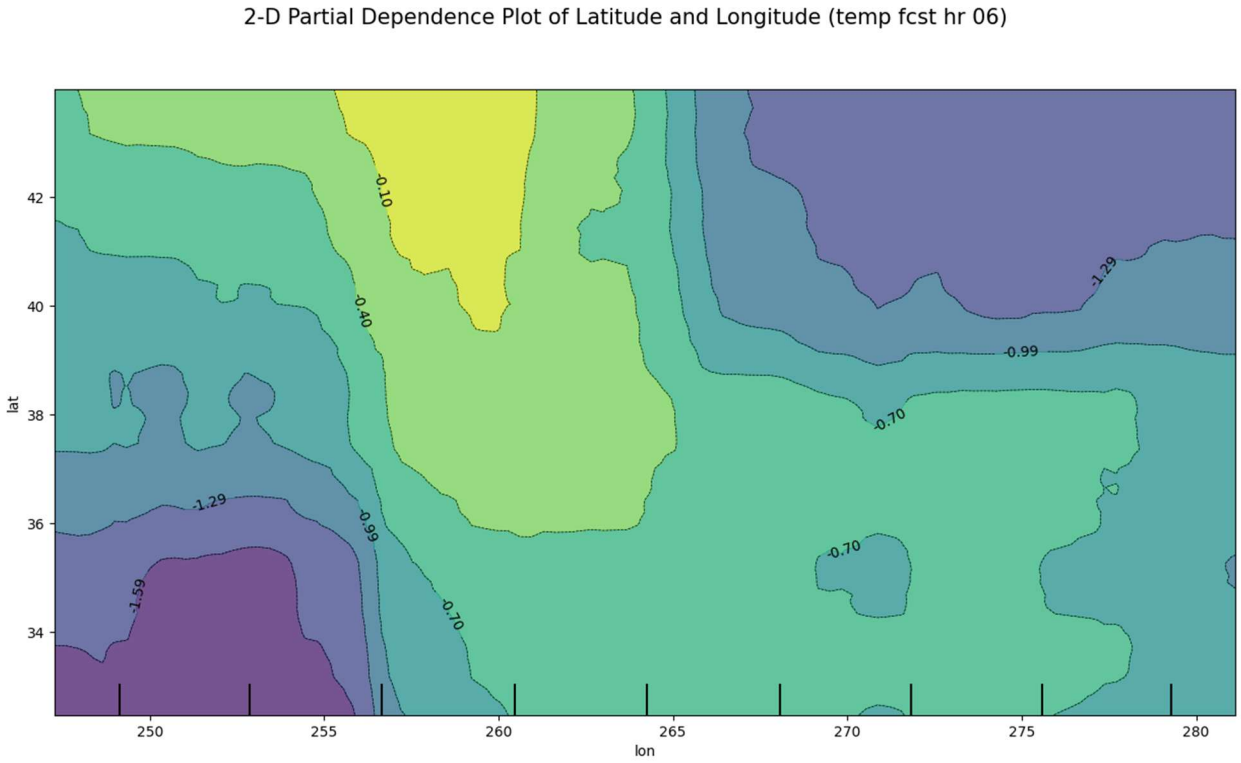
Plotting them individually, latitude and longitude variables show intriguing patterns, but plotting them together on a 2-D partial dependence plot allows us to see how forecast model bias varies spatially, granting the ability to determine the extent to which geographic location plays a role in forecast bias. Higher latitudes represent more northern locations, and higher longitudes describe more eastern locations, with the domain representing the lower 48 Continental United States (CONUS). Figures 13.a-d below show there is a strong regional dependence in RF's forecasted bias for all lead times and forecast cases, with contours being labelled with corresponding values of RF forecasted bias. Beginning with Figure 13.a, which displays RF forecasted bias for temperature at forecast hour 6 shows the RF is expecting the greatest forecast bias in the southwest (approx 247-256 lon, 32 – 36 N lat) and northeast (approx 267-281 lon, 40-44 N lat) showing spatially that the RF is sensitive to these geographic locations. Warmer colors correspond to areas where forecasts are typically more accurate, finding that systematic errors are generally smaller there, whereas the cooler colors represent greater negative values of RF forecasted bias, showing that the forecast model consistently underpredicts temperature in these locations at the 6-hour lead time.

As we increase the lead time from hour 6 to hour 21 for temperature, a shift in RF predicted bias (Figure 13.b) is noted. Firstly, the bias originally noted in the southwestern portion of the domain appears to have levelled out, while the northeastern area of interest has increased in RF predicted bias, indicating that the forecast model continues to struggle with forecasts of 2-m temperature in this area with increasing lead times. As seen from Figure 7.e, there is a very strong dependency of RF predicted forecast bias on east-west extent, with a sharp drop off in model accuracy as you pass east of longitude 264, which corresponds to far eastern portions of Oklahoma, Texas, and Kansas, and far western portions of Iowa and Minnesota. RF also predicts higher forecast bias north of latitude 40 N, which places a bullseye for the greatest RF predicted bias in the northeastern portion of the domain, where RF expects the largest systematic errors. West of longitude 264, RF seems to perform consistently, finding fewer systematic errors here.

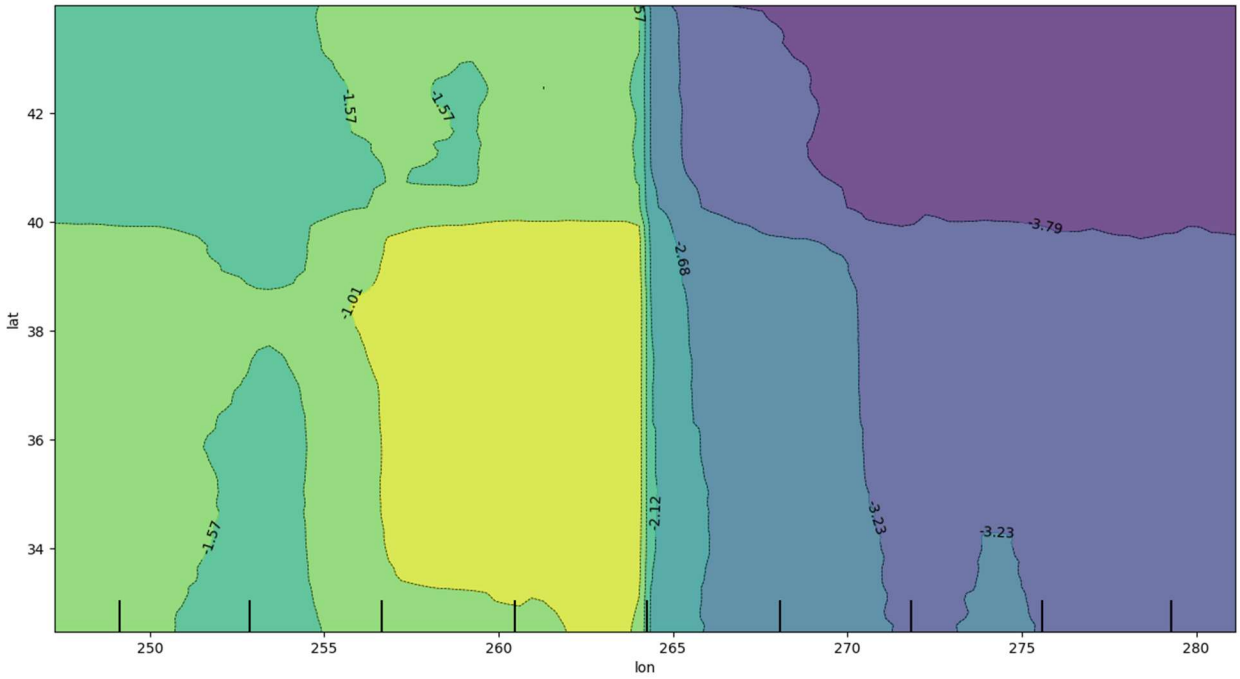
RF predicted bias for dew point at a 6-hour lead time (shown in Figure 13.c appears to be dominated most in part by longitude, with forecast accuracy decreasing with westward extent especially west of longitude 260. RF predicted bias also slightly increases north of latitude 39 N, placing the greatest forecasted dew point bias in the northwest CONUS, where the dewpoint is consistently overpredicted by the forecast model. Western portions of the domain represent mountainous and semi-arid terrain, where larger dew point biases could occur due to complex surface and atmospheric dynamics. Contrasted to the east, which is generally flatter and more humid, where model biases are more predictable or better captured.

Moving forward in lead time, we see in Figure 13.d that the dependencies of RF predicted forecast bias on latitude and longitude increased, especially the former (north of latitude 39 N). Increases in RF predicted bias west of longitude 260 were modest. Based off this interpretation, it

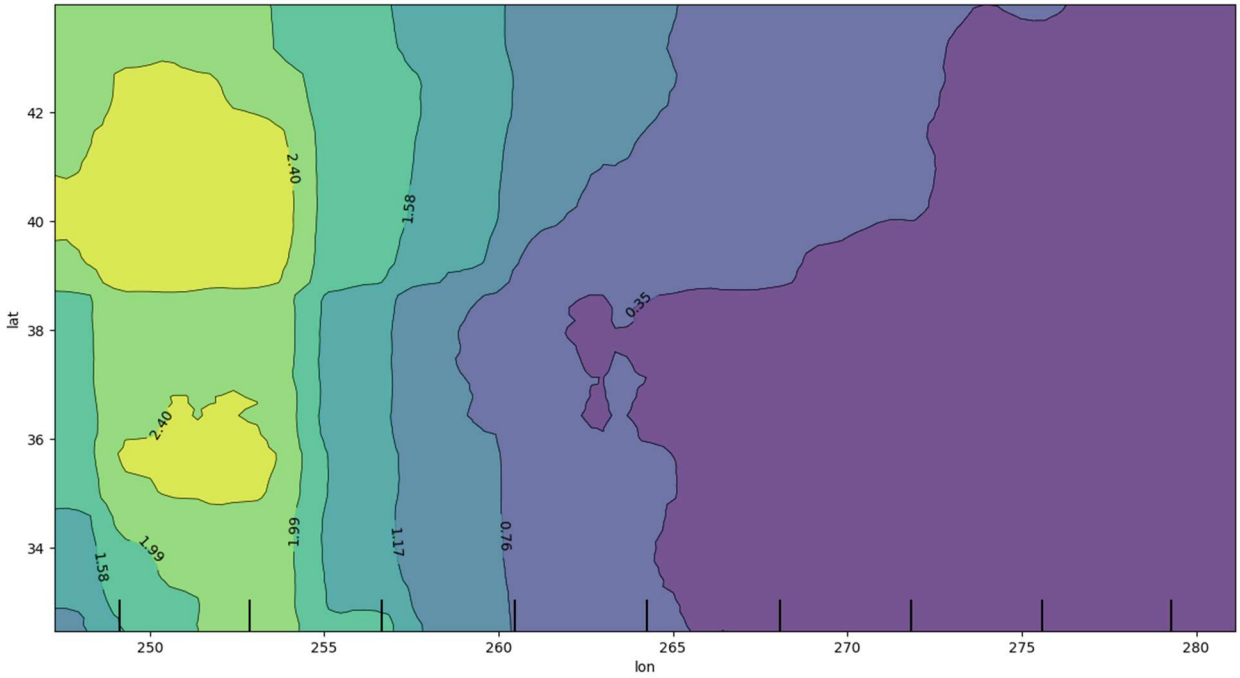
can be concluded that the forecast model is rather skillful in predicting dew point temperatures for southeastern regions of the domain (includes Louisiana, Arkansas, Missouri, and areas directly east), with systematic biases being lower in those regions at both lead times. RF predicted biases become more prevalent with northwestward extent, with a bullseye located in the far northwestern portion of the domain; predicted dew point bias is highest in this region, and the forecast model typically overpredicts the dew point temperature in its forecast. This result is consistent with dew point forecast hour 6, which also noted the northwestern portion of the domain as a trouble spot.

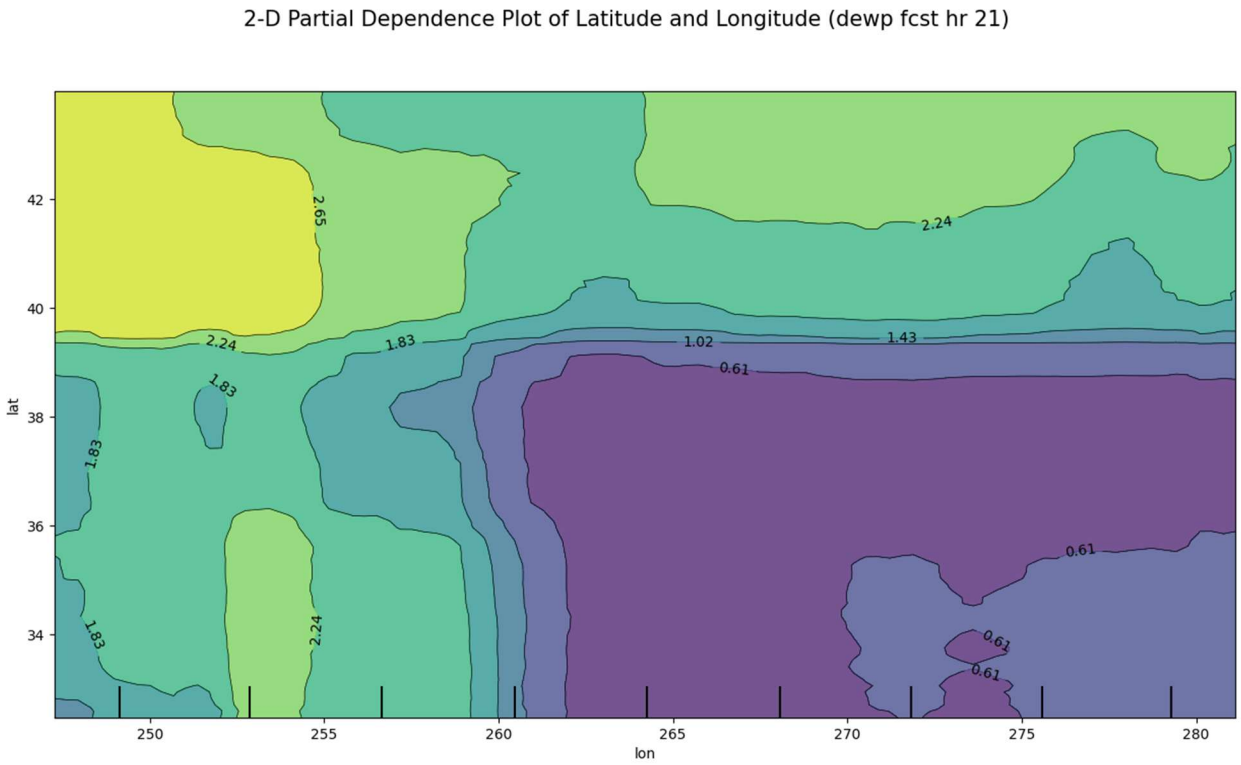


2-D Partial Dependence Plot of Latitude and Longitude (temp fcst hr 21)



2-D Partial Dependence Plot of Latitude and Longitude (dewp fcst hr 06)





Figures 13.a-d. 2-D partial dependence plots displaying longitude on the x-axis, latitude on the y-axis, and how bias is forecast by the RF is displayed spatially on the plot. These plots give a great indication of any geographic dependence of forecast bias.

4. Conclusions

In this study, we applied machine learning (ML) postprocessing methods to convection-allowing ensemble (CAE) forecasts of 2-meter temperature and 2-meter dew point using random forests (RF). These ensembles, though better than deterministic outputs, often still struggle with forecasts of these variables. Domain-average bias correction was performed, but the purpose of this study was to use a more advanced method of bias correction and compare its success to the former. A secondary goal was to find regional or flow-dependent biases that may correlate to specific predictors for the purpose of future model improvements.

Our RF model was trained on a wide range of predictors that may potentially play a role in model error, with an emphasis on land-atmosphere interactions in the selection of predictors. An extensive test of RF hyperparameters was done to determine which set was correlated with the best RF model performance, which was found to be when max depth was 20 and minimum samples per leaf node was 10 for all four forecast cases (hour 6 and 21 forecasts for temperature and dew point). The four trained models with this set of hyperparameters were then tested for impurity importance and analyzed to determine which predictors were most impactful in their predictions of forecast error, with variables such as soil water, soil temperature, latitude, sea level pressure, and cloud top brightness temperature being the strongest predictors for model bias. Analysis of 1-dimensional and 2-dimensional partial dependence plots of RF forecast error found that the forecast model struggles immensely with dry soil conditions in short term temperature forecasts (with a local maximum in the southwest CONUS) and wetter conditions in longer lead times. Soil temperature is a strong predictor of error in 21-hour temperature forecasts for both temperature and dew point, with forecasts often underpredicting temperature and overpredicting dew points when soil temperatures are cooler. Forecasts at this hour are also dependent on sea level pressure and soil moisture. Other biases were found dependent on wind speed and direction in temperature forecasts, with 6-hour forecasts struggling with southerly wind regimes

and 21-hour temperature forecasts underpredicting temperatures in calm conditions. Finally, latitude and longitude were found to be strong predictors in every forecast case, with the RF successfully picking up on systematic model errors that depend solely on geographic location or an interaction between geographic location and meteorological variables.

Overall, RF based postprocessing was found to be very successful in reducing model RMSE (i.e. increasing model accuracy), with substantial quantified improvements from uncorrected RMSE to RF bias corrected RMSE listed below for each case; a finding that solidifies RF postprocessing as a proven method for bias correction.

- 6-hour Temperature Forecasts: 25.4%
- 21-hour Temperature Forecasts: 38.4%
- 6-hour Dew Point Forecasts: 15.2%
- 21-hour Dew Point Forecasts: 26.6%

A major limitation of this study was such that it was to be completed in an accelerated timeframe, leaving little room for a more in-depth analysis of RF postprocessing. Future work in RF bias correction of 2-m temperature and dew point could be done to provide further analysis of RF's skill as a bias reduction method and expand upon the results of this study. Firstly, using a larger dataset could give more tuned interpretations of any flow dependent errors. It should be noted once again that the dataset used during this study was generated from the OU MAP Lab during the spring 2022 HWT experiment, which included forecasts for much of the month of May and early June. Perhaps conducting a similar study during a different time of the year, such as a collection of dates in the wintertime, could have differing outcomes, flow-dependencies, regional biases, or strong predictors of error. Using a wider array of predictors could also be done to give the RF more to work with to improve RF predictive skill. This would then enable the connection of more potential error relationships such as ones discussed in this paper. More connections to latitude and longitude could also be done to determine if certain variables struggle based upon geographic location. In this study, RF was only verified at 6- and 21-hour lead times, but future work could include more forecast times in this analysis, even including forecasts beyond 21 hours. Finally, more extensive hyperparameter tuning could be conducted to determine which set provides the best overall performance. This could also include playing with the total number of trees in the forest as well, which was not done in this study.

5. ACKNOWLEDGEMENTS

A very special thanks to the National Science Foundation, which funded this project under Grant No. AGS-2050267. I would also like to thank the University of Oklahoma, the National Weather Center REU program, and its directors, Alex Marmo and Daphne LaDue, for the opportunity to participate in this spectacular experience, as well as their excellent leadership. Next, I would like to thank my mentors, Dr. Aaron Johnson and Dr. Xuguang Wang, for their tremendous guidance throughout this project; my work would not have been possible without their help. Finally, I would like to thank my professors and peers at the University of Georgia for encouraging me to pursue this research opportunity and supporting me throughout.

5. REFERENCES

Agrawal, S., and Coauthors, 2023: A Machine Learning Outlook: Post-processing of Global Medium-range Forecasts, https://doi.org/10.48550/ARXIV.2303.16301.

Benjamin, S. G., G. A. Grell, J. M. Brown, T. G. Smirnova, and R. Bleck, 2004: Mesoscale Weather Prediction with the RUC Hybrid Isentropic–Terrain-Following Coordinate Model. *Mon. Wea. Rev.*, 132, 473–494, https://doi.org/10.1175/1520-0493(2004)132<0473:MWPWTR>2.0.CO;2.

Berner, J., and Coauthors, 2017: Stochastic Parameterization: Toward a New View of Weather and Climate Models. *Bulletin of the American Meteorological Society*, 98, 565–588, https://doi.org/10.1175/BAMS-D-15-00268.1.

Breiman, L., 2001: Random Forests. *Machine Learning*, 45, 5–32, https://doi.org/10.1023/A:1010933404324.

Bouallègue, Z. B., F. Cooper, M. Chantry, P. Düben, P. Bechtold, and I. Sandu, 2023: Statistical Modeling of 2-m Temperature

and 10-m Wind Speed Forecast Errors. *Monthly Weather Review*, 151, 897–911, https://doi.org/10.1175/MWR-D-22-0107.1.

Chapman, W. E., and J. Berner, 2025: Improving climate bias and variability via CNN-based state-dependent model-error corrections. *Geophys. Res. Lett.*, 52, e2024GL114106. https://doi.org/10.1029/2024GL114106

Chapman, W. E., A. C. Subramanian, L. Delle Monache, S. P. Xie, and F. M. Ralph, 2019: Improving Atmospheric River Forecasts With Machine Learning. *Geophysical Research Letters*, 46, 10627–10635, https://doi.org/10.1029/2019GL083662.

De Pondeca, M. S. F. V., and Coauthors, 2011: The Real-Time Mesoscale Analysis at NOAA's National Centers for Environmental Prediction: Current Status and Development. *Weather and Forecasting*, 26, 593–612, https://doi.org/10.1175/WAF-D-10-05037.1.

Duda, J. D., X. Wang, and M. Xue, 2017: Sensitivity of Convection-Allowing Forecasts to Land Surface Model Perturbations and Implications for Ensemble Design. *Monthly Weather Review*, 145, 2001–2025, https://doi.org/10.1175/MWR-D-16-0349.1.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, null (2/1/2011), 2825–2830.

Gagne, D. J., A. McGovern, and M. Xue, 2014: Machine Learning Enhancement of Storm-Scale Ensemble Probabilistic Quantitative Precipitation Forecasts. *Weather and Forecasting*, 29, 1024–1043, https://doi.org/10.1175/WAF-D-13-00108.1.

Gagne, D. J., A. McGovern S. E. Haupt, R. A. Sobash, J. K. Williams, and M. Xue, 2017: Storm-Based Probabilistic Hail Forecasting with Machine Learning Applied to Convection-Allowing Ensembles. *Weather and Forecasting*, 32, 1819–1840, https://doi.org/10.1175/WAF-D-17-0010.1.

Gasperoni, N. A., X. Wang, Y. Wang, and T.-H. Li, 2023: Evaluating the Multiscale Implementation of Valid Time Shifting within a Real-Time EnVar Data Assimilation and Forecast System for the 2022 HWT Spring Forecasting Experiment. *Weather and Forecasting*, 38, 2343–2362, https://doi.org/10.1175/WAF-D-23-0096.1.

Hamill, T. M., 2021: Comparing and Combining Deterministic Surface Temperature Postprocessing Methods over the United States. *Monthly Weather Review*, 149, 3289–3298, https://doi.org/10.1175/MWR-D-21-0027.1.

Hill, A. J., G. R. Herman, and R. S. Schumacher, 2020: Forecasting Severe Weather with Random Forests. *Monthly Weather Review*, 148, 2135–2161, https://doi.org/10.1175/MWR-D-19-0344.1.

McGovern, A., D. J. Gagne, J. K. Williams, R. A. Brown, and J. B. Basara, 2014: Enhancing understanding and improving prediction of severe weather through spatiotemporal relational learning. *Mach Learn*, 95, 27–50, https://doi.org/10.1007/s10994-013-5343-x.

McGovern, A, R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the Black Box More Transparent: Understanding the Physical Implications of Machine Learning. *Bulletin of the American Meteorological Society*, 100, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

Nakanishi, M., and H. Niino, 2009: Development of an Improved Turbulence Closure Model for the Atmospheric Boundary Layer. *Journal of the Meteorological Society of*

Japan, 87, 895–912, https://doi.org/10.2151/jmsj.87.895.

Olson, J. B., J. S. Kenyon, Wayne. A. Angevine, J. M. Brown, M. Pagowski, and K. Sušelj, 2019: A Description of the MYNN-EDMF Scheme and the Coupling to Other Components in WRF–ARW, https://doi.org/10.25923/N9WM-BE49.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, 2011: *Scikit-learn: Machine learning in Python*. J. Mach. Learn. Res., 12, 2825–2830. https://doi.org/10.5555/1953048.2078195

Pham, L. T., L. Luo, and A. Finley, 2021: Evaluation of random forests for short-term daily streamflow forecasting in rainfall- and snowmelt-driven watersheds. *Hydrol. Earth Syst. Sci.*, 25, 2997–3015, https://doi.org/10.5194/hess-25-2997-2021.

Pondeca, M., Levine, S., Carley, J., Lin, Y., Zhu, Y., McQueen, J., Manikin, G., Purser, R., Whiting, J., & Yang, R. (2015). *Ongoing improvements to the NCEP Real-Time Mesoscale Analysis (RTMA) and UnRestricted Mesoscale Analysis (URMA)*. [Technical report / American Meteorological Society Blue-Book].

Reynolds, C. A., and Coauthors, 2022: Analysis of Integrated Vapor Transport Biases. *Monthly Weather Review*, 150, 1097–1113, https://doi.org/10.1175/MWR-D-21-0198.1.

Schwartz, C. S., G. S. Romine, K. R. Fossell, R. A. Sobash, and M. L. Weisman, 2017: Toward 1-km Ensemble Forecasts over Large Domains. *Monthly Weather Review*, 145, 2943–2969, https://doi.org/10.1175/MWR-D-16-0410.1.

Schwartz, C. S., J. Poterjoy, G. S. Romine, D. C. Dowell, J. R. Carley, and J. Bresch, 2022a: Short-Term Convection-Allowing Ensemble Precipitation Forecast Sensitivity to Resolution of Initial Condition Perturbations and Central Initial States. *Weather and Forecasting*, 37, 1259–1286, https://doi.org/10.1175/WAF-D-21-0165.1.

Thompson, G., and T. Eidhammer, 2014: A Study of Aerosol Impacts on Clouds and Precipitation Development in a Large Winter Cyclone. *Journal of the Atmospheric Sciences*, 71, 3636–3658, https://doi.org/10.1175/JAS-D-13-0305.1.

Williams, J. K., 2014: Using random forests to diagnose aviation turbulence. *Mach Learn*, 95, 51–70, https://doi.org/10.1007/s10994-013-5346-7.