University of North Carolina Asheville
Journal of Undergraduate Research
Asheville, North Carolina
May 2021

Utilizing an Ensemble Feed Forward Neural Network to Reduce 24 – Hour Weather Forecast Error

Andrew Coleman
Department of Atmospheric Sciences
The University of North Carolina Asheville
One University Heights
Asheville, North Carolina 28804 USA

Faculty Advisor: Dr. Christopher Hennon

Abstract

A novel method of point weather forecasting is presented in which an ensemble of 950 individually trained feed-forward back propagation neural networks were developed to produce 24-hour forecasts for nine different cities across the contiguous United States in varying climate regimes totaling 72 forecasts. The ensemble forecasts consisted of the following parameters: 24-hour maximum temperature, minimum temperature, 2-minute maximum sustained wind speed, and accumulated precipitation. The ensemble neural network was trained on dynamical weather models (comprised of the primitive equations of the atmosphere), statistical weather forecasts (dynamical guidance post-processed using linear statistical methods), and human forecasters as input. More than two years of historical weather observations served as verification/targets for the input. Performance of the ensemble is assessed through a comparative analysis between it and the five input predictors to measure relevant error reduction achieved by the network. Error metrics that were used to assess this are: root mean square error (RMSE) and bias. Results indicate significant error reduction across all forecast parameters between the ensemble network forecasts and input model forecasts. Assessing ensemble forecast performance per city shows ensemble reductions of RMSE from the input models for maximum and minimum temperature, on average, exceeding 2-3 degrees Fahrenheit, wind of 3-5 knots and precipitation of 0.10-0.20 inches.

1. Introduction

A novel ensemble weather forecasting approach has been developed in which 950 feed-forward backpropagation neural networks[1] were individually trained on dynamical and statistical forecast guidance for nine different US cities (Figure 1) to produce 24-Hour forecasts from 06z-06z for maximum and minimum temperature, maximum 2-minute wind, and accumulated precipitation. Forecast model ensembles are developed by way of combining initial conditions (airport observations, weather balloon launches, derived satellite observations, etc.) with the first guess field (the 6-hr forecast from the prior deterministic run of the given model), running a sophisticated algorithm[1] in order to perturb these initial conditions, and finally utilizing the primitive equations of the atmosphere on these varied solutions to develop an ensemble of N number of solutions.

Over the last several decades, these ensemble prediction methods have been increasingly used in operational spaces due to the high level of skill in accounting for a wide range of possible solutions and postprocessing estimation errors of the initial boundary conditions[2]. From the specific model ensemble, several different forecast products can thus be created—one of which has high utility to operational meteorologists - probabilistic forecasts. Probabilistic weather guidance can be developed to quantify and convey overall uncertainty in any given forecast to the general public, whether it be a point forecast or a large geographical area. A study[3] divulging ensemble techniques elaborates by highlighting that the use of ensemble probabilistic forecasts as opposed to a single deterministic run of say, the Global Forecast System (GFS), proves to have systematic higher skill. Moreover, research has shown that when initializing

an ensemble forecast system with both the initial conditions and forecast model uncertainty, that a larger scope of possible true solutions is found and thus uncertainty is quantified on a much more accurate and robust level[4].

Within the focus of a neural network ensemble, a "normal" practice for perturbing possible solutions begins to diverge. One study[5] utilized two separate feed forward backpropagation neural networks with the input training data modified slightly in order to obtain localized precipitation forecasts. For their first iteration they trained on rain gauge data (as verification/targets for forecast model input) without regard for gauge measurement accuracy; for the second iteration all positive rain gauge errors of 0.09" were removed from the training data. The author notes that both individual and consensus neural network forecasts performed significantly better than the input numerical model (the Nested Grid Model (NGM), which has since been discontinued and superseded by the Global Forecast System model (GFS). Another study[6] utilized a unique a method of ensemble creation in order to forecast hail over Northeast Italy. Manzato used a bootstrap ensemble technique[7] in order to continuously retrain and assess for the highest performing feed forward neural network and thus implement them as the ensemble members. Results show that when compared to operational forecasts, this ensemble technique for hail forecasting, using sounding-derived indices, produced an 84% improvement in overall forecast with regards to the studies specific table of performance metrics.

In this study, 72 forecasts for 9 different US cities are produced from an ensemble feed-forward backpropagation neural network (henceforth referred to as the FFE). The FFE was developed to forecast daily 24-hour maximum and minimum temperatures, 2-minute wind speed, and total accumulated precipitation from 06z-06z. In the following section, methodology and datasets will be presented, alongside the selection of tuned hyperparameters, and data sources. Section 3 will outline conclusions**,** results, and future work

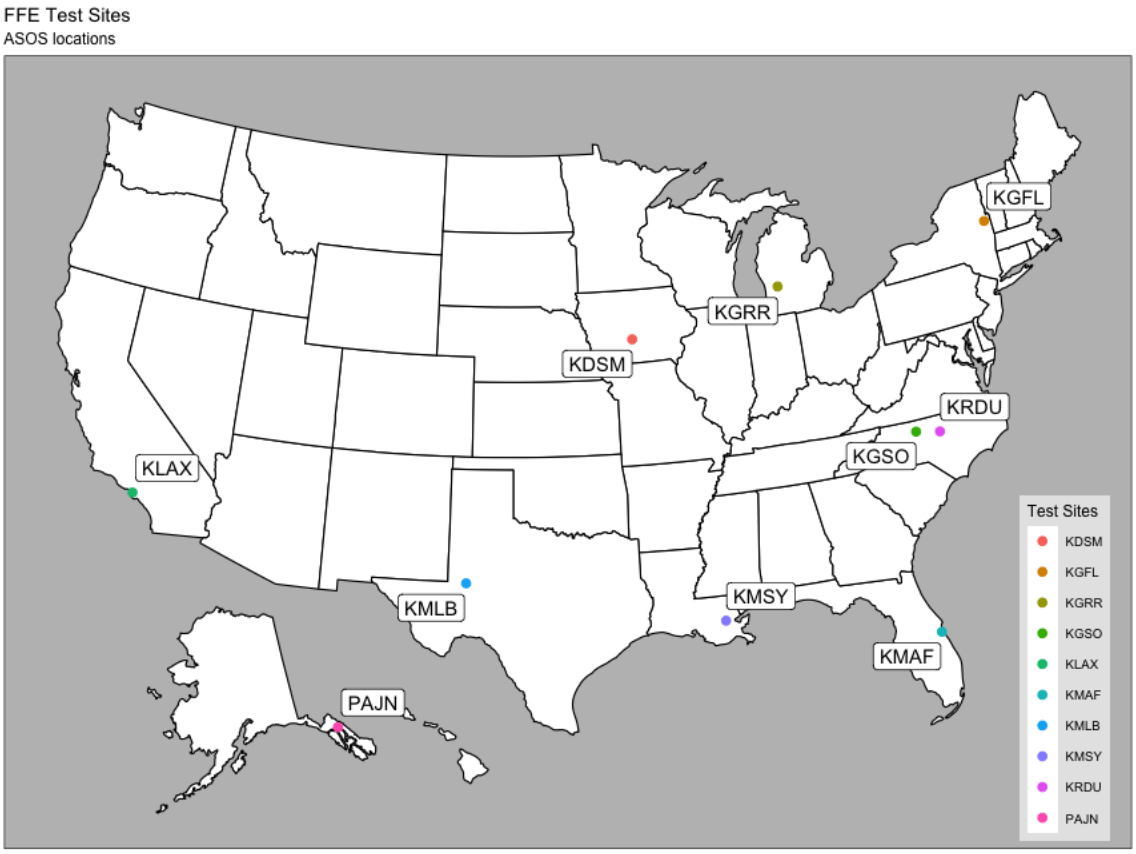

Figure 1. Locations of ASOS test sites.

2. Methodology and Datasets

Within a neural network (NN) framework, there are "tuning" parameters that allow the user to adjust certain aspects of the architecture in order to achieve maximum performance. Some of those parameters include number of neurons, performance statistics utilized in training (MAE, MSE etc.…), number of epochs (iterations through training set),

number of validation checks, learning rate, and much more. Within the FFE, the number of neurons, the training algorithm, validation checks, and epochs were all tweaked to achieve maximum performance. The novelty and power of the FFE is that the ensemble technique utilized is adapting many different training algorithms, training the FFE 50 times per algorithm and pooling each individual training sequence output into a total ensemble. Given such a high volume of non-linear training sequences, this allowed for far less time to be devoted to hyperparameter tuning as no appreciable increases (or decreases) in accuracy were observed in changing either the number of neurons, validation checks, or epochs.

With these small-scale changes of accuracy in mind, the number of epochs initialized within the FFE was 500, validation checks 500, and generally <5 neurons were used to decrease computational wear and tear. Exact specifications of each weather parameter and attendant hyperparameters can be seen in Table 1. Special attention is drawn to the precipitation ensemble hyperparameters as this parameter required the most tweaking due to the highly non-linear and complex nature of quantitative precipitation forecasts. Thus, a more constrained set of algorithms was utilized alongside a higher number of epochs and validation checks per train.

Table 1. FFE Hyperparameter Settings

	Maximum Temperature	Minimum Temperature	Maximum 2-Minute Wind Speed	Quantitative Precipitation Forecast
# Validation checks	500	500	500	500
# Epochs	500	500	300	300
# Neurons	3	3	10	3

Training algorithms implemented into the FFE are as follows: scaled conjugate gradient backpropagation (trainscg), Levenberg-Marquardt backpropagation (trainlm), Bayesian regularization backpropagation (trainbr), Conjugate gradient backpropagation with Powell-Beale restarts (traincgb), gradient descent with momentum backpropagation (traingdm), gradient descent with momentum and adaptive learning rate backpropagation (traingdx), and Gradient descent backpropagation (traingd). Specific training algorithms utilized within the different weather parameter batch trains are included in Table 2. Further, in order to achieve brevity, for in-depth training specifications regarding each algorithm and supplementary methodology the author refers the reader to MATLAB documentation[8] for further reading.

Table 2. Specifications of algorithms utilized within each parameter

	Maximum Temperature	Minimum Temperature	Maximum 2-Minute Wind Speed	Quantitative Precipitation Forecast
Trainscg	X	X	X	X
Trainlm	X	X	X	X
Trainbr	X	X	X	X
Traincgb	X	X	X	X
Traingdm			X	
Traingdx	X	X	X	X
Traingd			X	
Total Ensemble Members (NN Retrains)	N = 250	N = 250	N = 350	N = 250

The FFE was trained on five input predictors and their attendant 24-Hour forecasts from 06z-06z for a 2.5-year period from January 2018—January 2021. Those input predictors are, the North American Mesoscale Model Output Statistics (MOS) product (the NAM), the Global Forecast System MOS product (the GFS), the National Blend of Models MOS product (the NBS), the High-Resolution Rapid Refresh dynamical output (the HRRR)[9], and the human generated Point Forecast Matrix (the PFM) from the National Weather Service. The NAM, GFS, and NBS, all were obtained from the Iowa Environmental Mesonet Model MOS archive[10]. The PFM dataset was also obtained from the

NWS product archive from the Iowa Environmental Mesonet and thus had specific forecasts for the desired time range programmatically mined. The HRRR archive was obtained from the Utah HRRR archive (Blaylock et al. 2017). For verification/targets, the FFE input was validated using the same model input timespan at the designated Automated Surface Observation Station (ASOS). Target data are obtained through the National Climatic Environmental Information Global Historical Climate Network dataset. Finally, the software utilized to produce and design the FFE was MATLAB version R2020a.

Training of the FFE was implemented on the latest runs of each of the five input predictors prior to the 00z hour in order to ingest the most recent and up to date data into the ensemble. Therefore, the NAM 12z, the GFS 18z, the NBS 19z, the HRRR 18z, and generally the PFM 21z output (release times of the PFM from each NWS office differs and would range from 20-22z) were operational runs on all forecasts of the FFE. For each forecast and each parameter, the ensemble mean was implemented as the deterministic forecast of the FFE. Of all the forecast parameters, as can be seen in Table 1, QPF is the only parameter with less than half of the models as its input. The decision to only utilize the HRRR and the NBS was a difficult one but ultimately was a result of data source issues where for the GFS and NAM bulletin's, QPF is given as a code and represents a range of precipitation (e.g., 0.00-0.10") not a deterministic output and the PFM QPF output had inconsistent forecast ranges for the time range (06z-06z) that was settled upon. Thus, the available model output dwindles down to the HRRR and the NBS.

3. Results

Two performance metrics were calculated to evaluate the FFE: root mean squared error (RMSE) and bias: (2).

$$RMSE = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}(d_i - f_i)^2} \tag{1}$$

$$Bias = \overline{\left(\hat{\theta}\right)} = \frac{1}{n}\sum_{i=1}^{n}\left(\left(\hat{\theta}\right) - \theta\right) \tag{2}$$

Where in equation (1), parameter d_i, is the predicted model value and parameter f_i, is the observed/target value. For equation (2), parameter $\left(\hat{\theta}\right)$ is the predicted model value and parameter θ, is the observed target value.

Performance metrics were constrained to the two based on the following criterion: 1) RMSE and its correlation and power within assessing a given models predictive accuracy, and 2) average bias and its simplicity in evaluating the normal tendency of model's predictive behaviors across an entire dataset

3.1 Maximum Temperature

Figures 2-3 compare the average RMSE and bias of each input model over the course of the 9 different ASOS locations against the FFE. Analyzing these results provides insight to the true power of a NN and the significant ability at reducing systematic bias and error within a forecast model output for any given location. Figure 2, the average RMSE, shows promising results with the FFE besting all input models. It's imperative to note the close RMSE score between the FFE and the NAM and then the NBS. Looking at the architecture of these two model inputs, the NAM is a product generated from an already sophisticated model output. With a horizontal grid spacing of 3 km and filtering such high-resolution output through advanced linear statistics (MOS techniques), the NAM thus, on average, is comparable to the FFE with an RMSE of 2.74 and the FFE with 2.13. Similarly, the NBS is a product composed of several different high-resolution model outputs and also filtered through the same MOS techniques as the NAM, making it as well comparable to the NAM and FFE with an RMSE of 3.45.

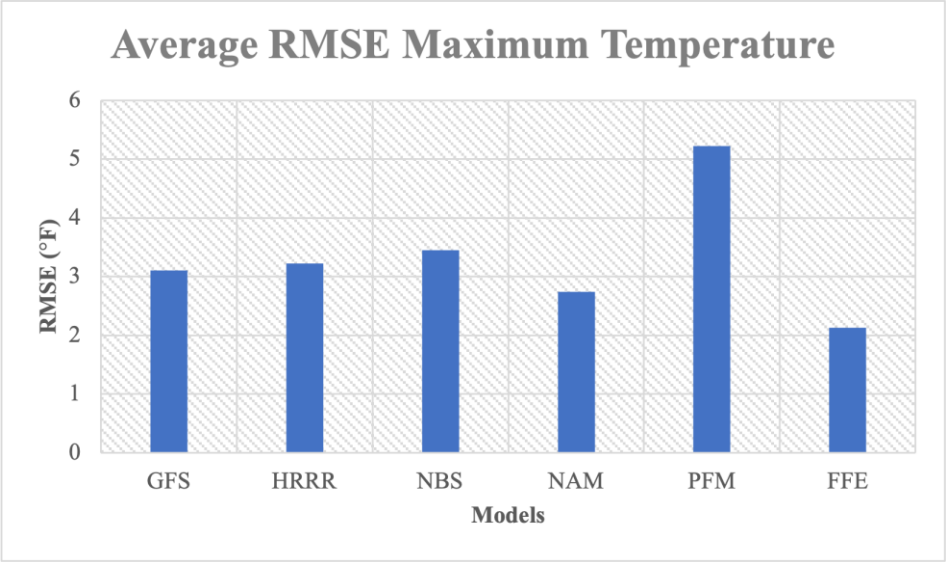

Figure 2. Average RMSE of maximum temperature forecasts.

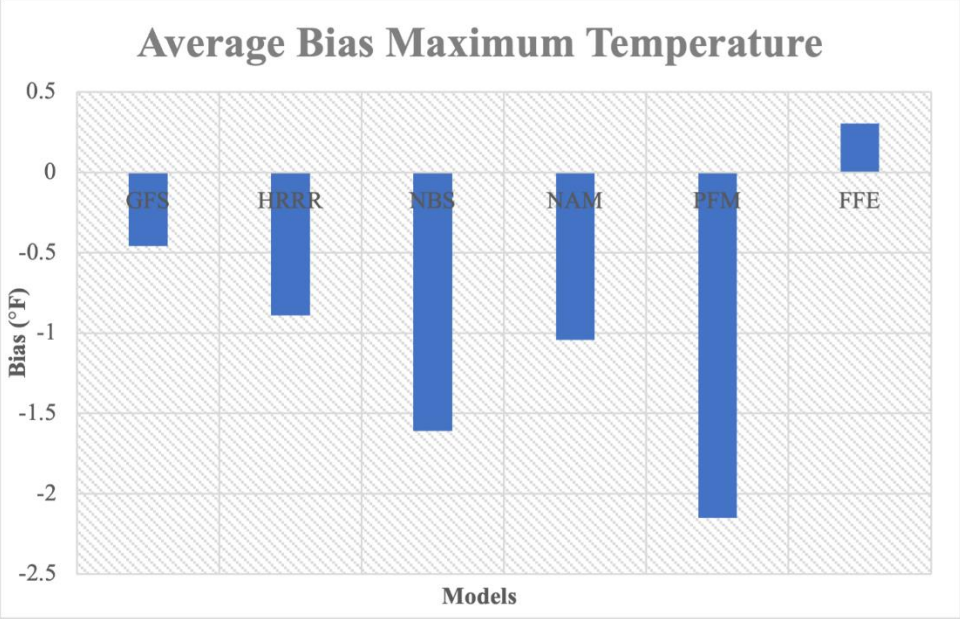

Figure 3. Average bias of maximum temperature forecasts

Figure 3, the average bias, allows for a different look at the resolving and bias reduction of the FFE's architecture. With a positive bias of .31, the FFE successfully removed all negative bias from the model input. With the closest competitor being the GFS at -.46 and the worst competitor being the PFM at -2.15, these results are promising and display a powerful post-processing technique for temperature output.

3.2 Minimum Temperature

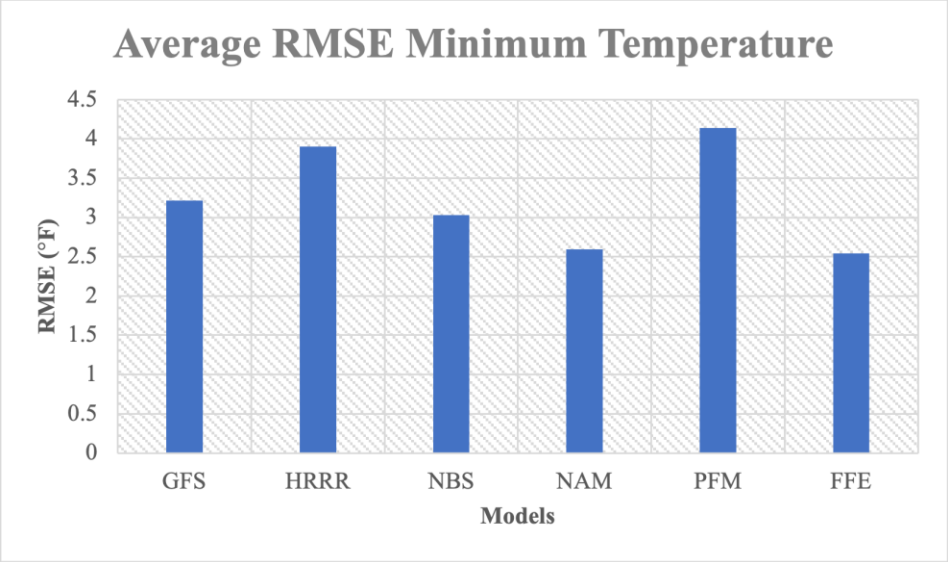

Figure 4. Average RMSE of minimum temperature forecasts.

Figure 4 shows the average RMSE for minimum temperature of all model input versus the FFE. It's clear that the FFE has an edge on its competitors though only with a small reduction in error, again, over the NAM. With the NAM versus the FFE being 2.6 and 2.54, respectively. Though when analyzing all inputs against the FFE, the error reduction of the ensemble is clearly superior.

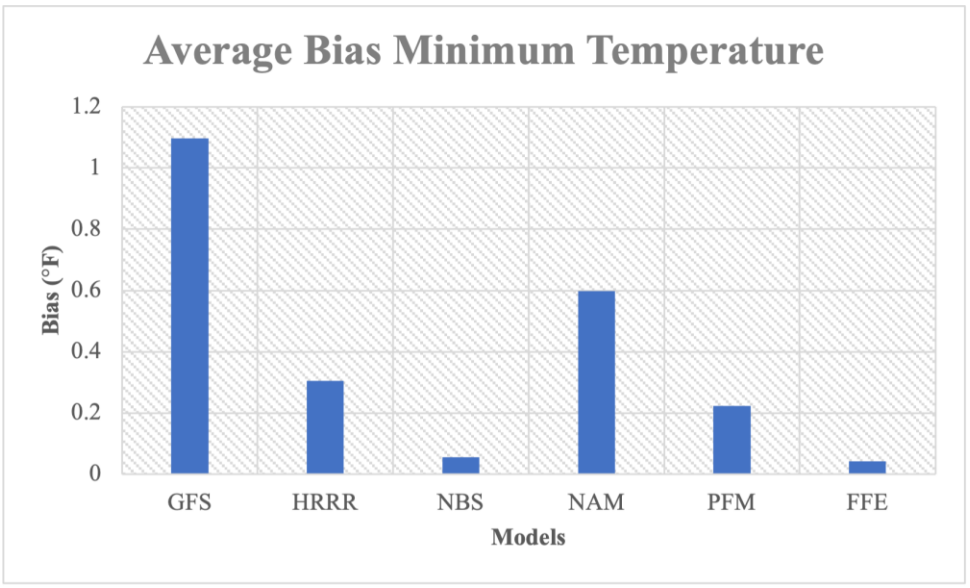

Figure 5. Average bias of minimum temperature forecasts.

Figure 5 displays the average bias of all of the models' minimum temperature. Looking over this performance metric illustrates a new story in which all guidance showed to have a positive bias. While the FFE, again, was able to almost completely remove all bias it was able to stay within the confines of a positive bias at .04 or, in other terms, not over-

correct for the significant biases within the model input. This is a significantly promising result as it shows both the powerful bias reduction and the stability of the model to not overcorrect and thus unintentionally underperform.

3.3 Maximum 2-Minute Wind

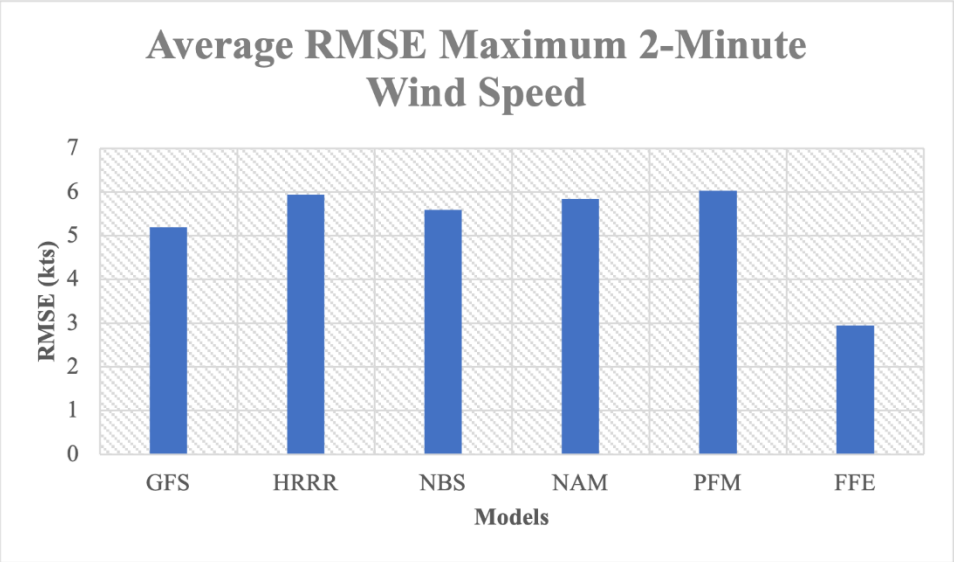

Figure 6. Average RMSE of maximum 2-minute wind speed forecasts.

Without a doubt one of the most promising results seen in the FFE's performance is the significant error reducing and bias removal of wind speed forecasts of all the input models. Figure 6 shows the true power of the FFE in this light with an RMSE of 2.95 and the closest competitor being the GFS at 5.19. The FFE's error reduction in wind speed forecasts, with respect to RMSE, very clearly shows a powerful post processing output technique.

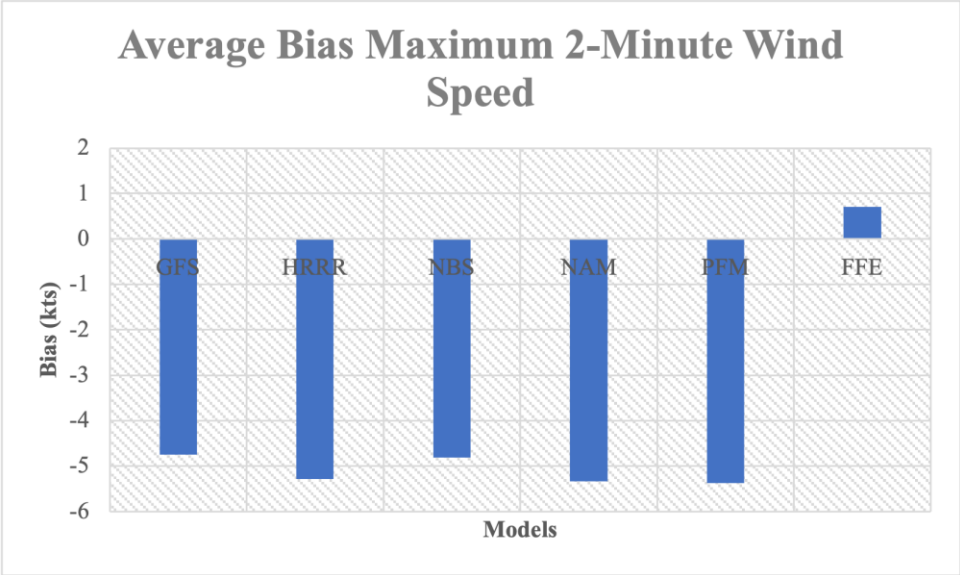

Figure 7. Average RMSE of maximum 2-minute wind speed forecasts.

Analyzing the average bias in Figure 7, the same story as RMSE comes to light. With a slight overcorrection in bias at a positive .71 and all model input being <-4.75, it is important to note the power of the FFE with regards wind

speed forecasts. Results of this magnitude bring to light the need for either far better post processing techniques of wind speed or a much larger in-depth review of attendant physics packages utilized in wind speed forecasts.

3.4 Quantitative Precipitation Forecasts

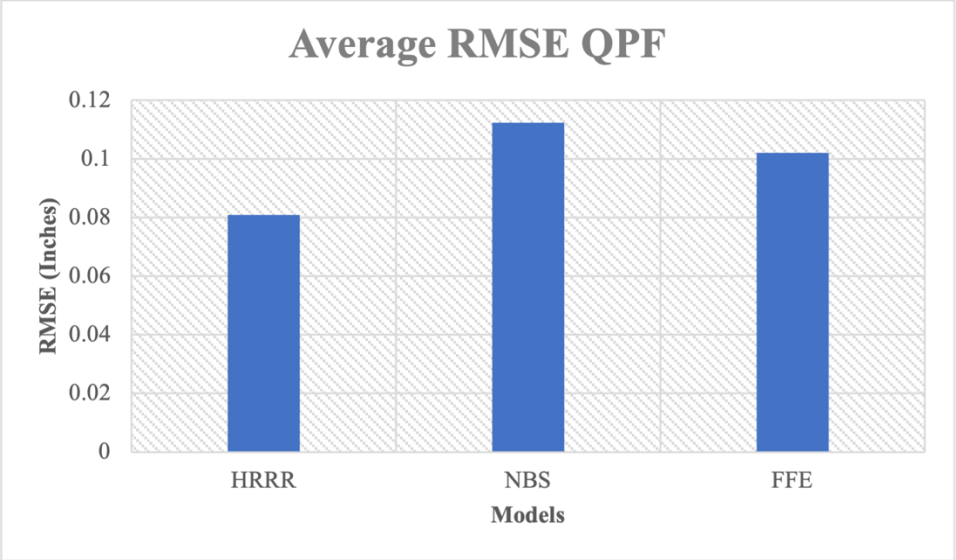

Figure 8. Average RMSE of QPF.

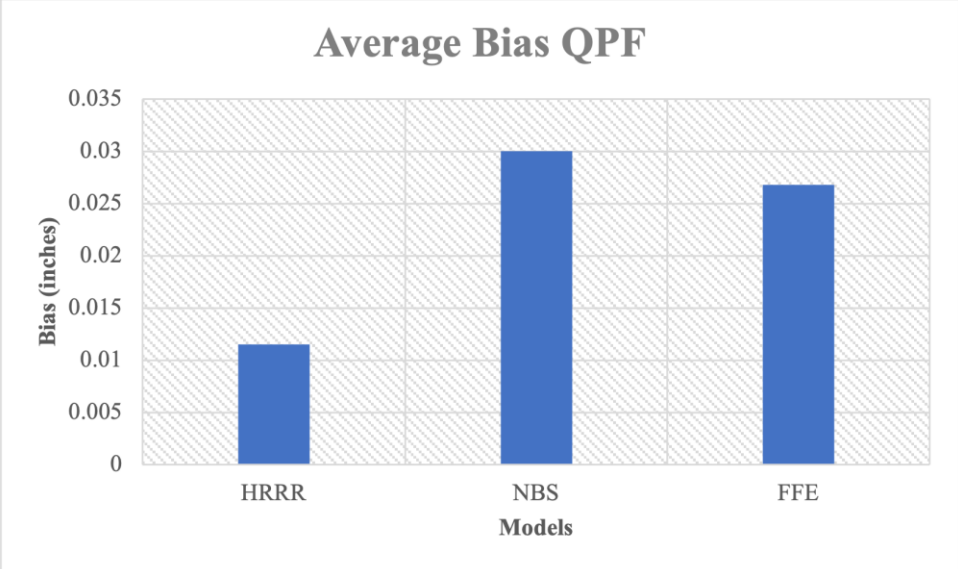

Figure 9. Average bias of QPF.

Possibly one of the most challenging parameters to tune a forecast model to is QPF. Due to the highly nonlinear and non-normal distributions of QPF datasets, QPF forecasting of the FFE and most certainly other output, is far from perfect. Displayed in Figure 8 and figure 9, a slight reduction in overall error and bias compared to the NBS is seen. Although the FFE does achieve marginally better error reduction than the NBS, the HRRR is noted to have narrowly better metrics with an RMSE of .08 and a bias of .01. With these results comes the necessity to highlight the deficiency in model input and training data—the HRRR's dataset begins in September 2019 and the NBS in November 2019. The shortened training dataset coupled with only two models as input, puts the FFE's QPF forecasts at a disadvantage.

Thus, author notes that these results should be taken with a grain of sand and will refine datasets and predictors in future work.

4. Conclusions

In this study, a novel ensemble weather forecast model composed of 950 individually trained NNs was utilized to produce point forecasts for maximum and minimum temperature, maximum 2-minute wind speed, and total accumulated precipitation across nine different US cities.

Results of this study, proved to be significant and overall, successful. With respect to developing postprocessing techniques for operational model output, results from this study show the necessity of implementing similar techniques to those found in this study.

In regard to future work many avenues are planned to refine and increase overall ensemble accuracy. This includes higher order ensemble filtering techniques[11] the implementation of a multimodel superensemble technique[12] for each algorithm, and finally in the later stages, a graphical timeseries FFE forecast.

5. Acknowledgements

I wish to thank my faculty advisor, Dr. Christopher Hennon for providing myself the council to succeed in this research. Many thanks to my colleague, Andy Hill, who aided in the implementation of this research and other attendant research. Finally, I would like to thank Dr. Bryan Blaylock of Utah University and the very valuable communication and help with implementing HRRR data mining.

6. References

1. Schmidhuber, J., 2015: Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117.
2. Vukicevic, T., I. Jankov, and J. McGinley, 2008: Diagnosis and Optimization of Ensemble Forecasts. *Mon. Wea. Rev.*, **136**, 1054–1074.
3. Murphy, A. H., 1993: What Is a Good Forecast? An Essay on the Nature of Goodness in Weather Forecasting. *Wea. Forecasting*, **8**, 281–293.
4. Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using Ensembles for Short-Range Forecasting. *Mon. Wea. Rev.*, **127**, 14.
5. Kuligowski, R. J., and A. P. Barros, 1998: Localized Precipitation Forecasts from a Numerical Weather Prediction Model Using Artificial Neural Networks. *Wea. Forecasting*, **13**, 11.
6. Manzato, A., 2013: Hail in Northeast Italy: A Neural Network Ensemble Forecast Using Sounding-Derived Indices. *Wea. Forecasting*, **28**, 3–28.
7. Hassan, A., A. Abbasi, and D. Zeng, 2013: Twitter Sentiment Analysis: A Bootstrap Ensemble Framework. *2013 International Conference on Social Computing*, 2013 International Conference on Social Computing (SocialCom), Alexandria, VA, USA, IEEE, 357–364.
8. MathWorks, 2021: Function Approximation and Nonlinear Regression — Functions. Accessed April 5, 2021, https://www.mathworks.com/help/deeplearning/referencelist.html?type=function&listtype=alpha&category=function-approximation-and-nonlinear-regression&blocktype=&capability=&s_tid=CRUX_topnav.
9. Blaylock, B. K., J. D. Horel, and S. T. Liston, 2017: Cloud archiving and data mining of High-Resolution Rapid Refresh forecast model output. *Comp. & Geo.* **109**, 43–50.
10. Herzmann, D., Iowa Environmental Mesonet, 2021: Archived Data & Plots. Accessed April 6, 2021, https://mesonet.agron.iastate.edu/archive/.
11. Anderson, J. L., and N. Collins, 2007: Scalable Implementations of Ensemble Filter Algorithms for Data Assimilation. *J. Atmos. Oceanic Technol.*, **24**, 1452–1463.
12. Krishnamurti, T. N., C. M. Kishtawal, D. W. Shin, and C. E. Williford, 2000: Improving Tropical Precipitation Forecasts from a Multianalysis Superensemble. *J. Climate,* **13**, 4217–4227.

Function Approximation and Nonlinear Regression — Functions.