NOAA
NCEP

NCEP OFFICE NOTE 520
https://doi.org/10.25923/6266-e822

A Machine Learning-Based Bias Correction Method for Global Forecast System Products

Linlin Cui, Jun Wang, Sadegh Sadeghi Tabas and Jacob R. Carley

US DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Weather Service
National Centers for Environmental Prediction
College Park, MD
February 2025

A Machine Learning-Based Bias Correction Method for Global Forecast System Products

Linlin Cui[1,2] (https://orcid.org/0009-0000-7997-9845)
Jun Wang[1] (https://orcid.org/0009-0007-9030-3417)
Sadegh Sadeghi Tabas[1,3] (https://orcid.org/0000-0001-9157-3397)
Jacob R. Carley[1] (https://orcid.org/0000-0003-4763-6666)

[1]NOAA National Weather Service, National Centers for Environmental Prediction, College Park, MD, USA (https://ror.org/00ndyev54)

[2] Lynker at Environmental Modeling Center, NOAA/National Centers for Environmental Prediction, College Park, MD, USA

[3] Axiom at Environmental Modeling Center, NOAA/National Centers for Environmental Prediction, College Park, MD, USA

National Centers for Environmental Prediction Office Note 520
February 2025

Recommended citation

Cui, L., J. Wang, S. S. Tabas, and J. R. Carley. (2025). A Machine Learning-Based Bias Correction Method for Global Forecast System Products. U.S. Dept. of Commerce, NOAA NCEP Office Note 520, 23 p. doi: 10.25923/6266-e822

Acknowledgements

The authors thank Alicia Bentley and Wei Li at the NOAA Environmental Modeling Center for internal reviews. The authors acknowledge the high performance computing resources provided by the NOAA Research and Development High Performance Computing Program. The project described in this article was supported by the Inflation Reduction Act and the NOAA Software Engineering for Novel Architectures (SENA) project. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce.

Copies of this report are available from

https://repository.library.noaa.gov/

ABSTRACT

Accurate numerical weather forecasting is essential for forecasters to make predictions. However, operational forecast models often exhibit systematic biases. In this study, we present a machine learning-based approach called BC-Unet, a convolutional neural network (CNN) based on the renowned U-Net architecture, to correct biases in National Centers for Environmental Prediction (NCEP) operational Global Forecast System (GFS) v16 products over the contiguous United States (CONUS). We incorporated the convolutional block attention module (CBAM) into the UNet architecture, which applies attention maps to the input feature map for adaptive feature refinement. We trained BC-Unet for operational GFS 2-m temperature with European Center for Medium-Range Weather Forecasts (ECMWF) Reanalysis 5 (ERA5) as ground truth, and subsequently used the model to correct biases during the test period. BC-Unet was assessed with both reanalysis data and observations. Results show that BC-Unet, when trained on a single forecast lead time from all forecast cycles (00 UTC, 06 UTC, 12 UTC, and 18 UTC), can be effectively applied to all other forecast lead times. Overall, BC-Unet reduced both root-mean-squared error (RMSE) and mean error (bias). When compared against ERA5, BC-Unet reduced the mean RMSE by up to 0.35 °C and cold bias by up to 0.8 °C in forecast cycle 00 UTC compared to GFS original forecasts, respectively. For all CONUS subregions, the BC-Unet was able to reduce biases and follows ERA5's diurnal variations.

1. Introduction

The accuracy of numerical weather prediction (NWP) is crucial for a variety of applications, including agricultural planning, energy management, transportation, and giving early warning of storms, heat waves, and natural disasters. However, despite advancements in NWP models, systematic biases in forecasts persist due to limited resolution or inaccurate physical parameterizations. In order to better utilize numerical model forecasts, developing postprocessing methods to reduce forecast biases is a standard practice.

Traditional bias-correction techniques, such as statistical postprocessing (Alerskans and Kaas, 2021), have been employed to address systematic errors in NWP outputs. These methods typically involve linear regression (Cheng and Steenburgh, 2007), moving average (McColor and Stull, 2008), and Kalman filters (Monache et al., 2011) to adjust model outputs based on historical observations. While effective to a certain extent, these conventional bias-correct methods often fall short in capturing complex, non-linear relationships between model forecasts and observations.

During the last few decades, the advent of machine learning (ML) has opened new avenues for bias correction, offering the potential to better account for intricate and non-linear dependencies. Among ML techniques, deep learning (DL) has shown significant promise. Deep learning algorithms, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), can learn from vast amounts of data to model relationships between predictions and truth more accurately. These approaches can adapt to changing conditions and provide more robust correlations compared to traditional statistical methods. For instance, Sha et al. (2020) applied convolutional neural networks for downscaling climate models and correcting biases in daily maximum and minimum 2-m temperature, achieving notable improvements over complex terrain. Similarly, Han et al. (2021) showed the deep learning method outperformed a traditional method, (i.e., anomaly numerical correction with observation) for 2-m temperature, 2-m relative humidity, 10-m wind speed, and 10-m wind direction from European Centre for Medium-Range Weather Forecast (ECMWF) Integrated Forecasting System global model.

The National Oceanic and Atmospheric Administration (NOAA) National Centers for Environmental Prediction (NCEP) Global Forecast System (GFS) has a longstanding issue of cold biases in the 2-m temperature field (Zheng et al., 2017). These biases have diurnal cycles and vary geographically. For example, Patel et al. (2021) compared the output from GFSv15 with observations at 210 airports across the United States and found a strong diurnal cycle in 2-m temperature errors conditioned on observed weather conditions (e.g., cloud cover amount). These errors are due to a large extent to boundary-layer and land-atmospheric coupling processes. The operational GFS was upgraded to v16 on 22 March 2021 (National Weather Service, 2021). One goal of the GFSv16 was to reduce the surface temperature biases. However, the Model Evaluation Group at NCEP's Environmental Modeling Center (EMC) has documented cold biases within GFSv16 as well (EMC Model Evaluation Group, 2020).

This study aims to explore the application of deep learning models for bias-correcting 2-m temperature of the GFSv16. The model used in this study is based on the U-Net (Ronneberger et al., 2015), which was originally developed for medical images.

2. Data and Method

In this study, we focus on the contiguous United States (CONUS) subdomain with the longitude expanding from 60°W to 130°W, and the latitude expanding from 25°N to 50°N. The total grid points are 101281 at 0.25° latitude/longitude resolution. Figure 1 shows the study domain and the terrain information. GFS and ECMWF Reanalysis v5 (ERA5) models are used for the period of 23 March 2021 to 31 May 2024, with 23 March 2021 to 31 December 2023 used for training, and 1 January to 31 May 2024 used for testing. For the training dataset, 10% of the dataset was randomly selected for validation. Because of the GPU memory limitation at the time, we started training the model with a single forecast hour (FH72). Systematic forecast errors are initially small, and grow linearly until Day 1 (FH24), then nonlinearly because of interacting with other systematic and random errors (Bhargava et al., 2018). FH72 could be a good starting point which not only avoids capturing only the initial errors, but also adds nonlinearity. As the GFS is run every 6 hours, the temporal resolution is 6-hourly. Then the model weights were used to correct forecast hours 6 to 240. To better understand the model's performance, NCEP Automated Data Processing (ADP) Global Surface Weather Observations from 1 January and 31 May 2024 were used to compare model output and bias-corrected results with observations.

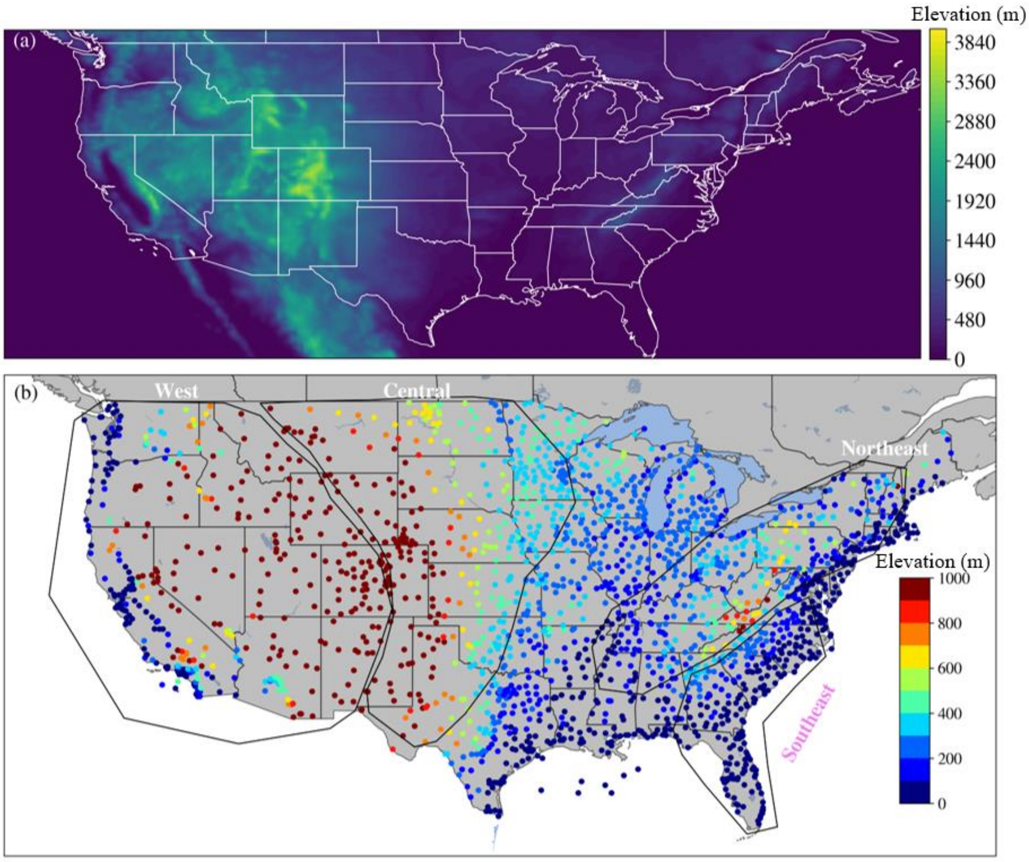

Fig. 1. (a) Study domain. Colors show orography (m) with a 0.25º by 0.25º resolution. (b) Observation stations with colors showing elevation (m) above sea level. Polygons show four CONUS subregions: west, central, northeast, and southeast CONUS.

a. Observations

NCEP ADP Global Surface Weather Observations (BUFR format) are composed of a global set of 2-m above ground level sensible temperature reports. The data includes land and marine surface reports received via the Global Telecommunications System. Data was downloaded from the NSF National Center for Atmospheric Research (NCAR) Research Data Archive. The stations were subset to isolate CONUS observations. For each station, 10-day observations were compared with each forecast cycle (00 UTC). If there is missing data within these 10-day observations, this forecast cycle will be discarded. The observations were resampled at the model's 6-h interval based on the nearest timestamp. The final CONUS station total was 2040 for the period 1 January to 31 May 2024.

b. GFS

The operational GFS is a global non-hydrostatic numerical weather prediction (NWP) model built with the NOAA Geophysical Fluid Dynamics Laboratory Finite-Volume Cubed-Sphere dynamical core (FV3; Lin, 2004, Putman and Lin, 2007, Harris et al., 2020) and the Grid-Point Statistical Interpolation (GSI) data assimilation system (Kleist et al., 2009, Kleist and Ide, 2015). The horizontal resolution is approximately13 km. In the vertical, there are 127 layers extending from the surface to 84 km. The GFS is run four times a day, with forecasts extending 16 days into the future. GFS forecast products are hourly for the first 120 hours, then every 3 hours for days 5–16 on a 0.25° latitude/longitude grid. We downsampled GFS output to 6 hour time intervals for forecast hours 6 to 240.

c. ERA5

The ground truth used in this study is the ERA5 reanalysis dataset (Hersbach et al., 2020), which is the fifth generation of ECMWF atmospheric reanalysis of the global climate. ERA5 assimilated 12-hour windows of observations (satellite and in-situ) with 4D-Var data assimilation system. The ERA5 data covers the period from January 1940 to the present, at 0.25° latitude/longitude resolution and 1 hour increments. The data was downsampled to 6 hour time intervals.

d. Model architecture

The U-Net is a type of CNN initially developed for biomedical image segmentation (Ronneberger et al., 2015). U-Net performs well on tasks where precise localization and segmentation of features are required. Its u-shaped structure (Figure 2), which includes an encoding (downsampling) path and decoding (upsampling) path, allows it to capture both the context and the detailed features of the input. The encoding path consists of repeated application of convolutions followed by max-pooling operations, which reduce the spatial resolution of the data while increasing the feature sizes. The decoding path, on the other hand, involves upsampling operations that restore the spatial resolution and combine them with the corresponding feature maps from the encoding path through skip connections. This structure allows U-Net to utilize both coarse and fine features for accurate prediction.

We kept the number of the downsampling and upsampling blocks the same as the original U-Net. The size of convolution kernels is 3-by-3 with batch normalization (Ioffe and Szegedy, 2015) and leaky rectified linear unit (LeakyReLU, Maas et al., 2013). We added the Convolutional Block Attention Module (CBAM, Woo et al., 2018) to the original U-Net. The attention mechanism can identify important features across channels and spatial regions (Trebing et al., 2021). The CBAMs are performed after each double convolution. However, the convoluted image is downsampled along the encoding path, which can keep the original feature. The image with the attention mechanism is reused in the corresponding upsampling part through the skip-connections (Figure 2). We refer to the bias-correction model used in this study as BC-Unet.

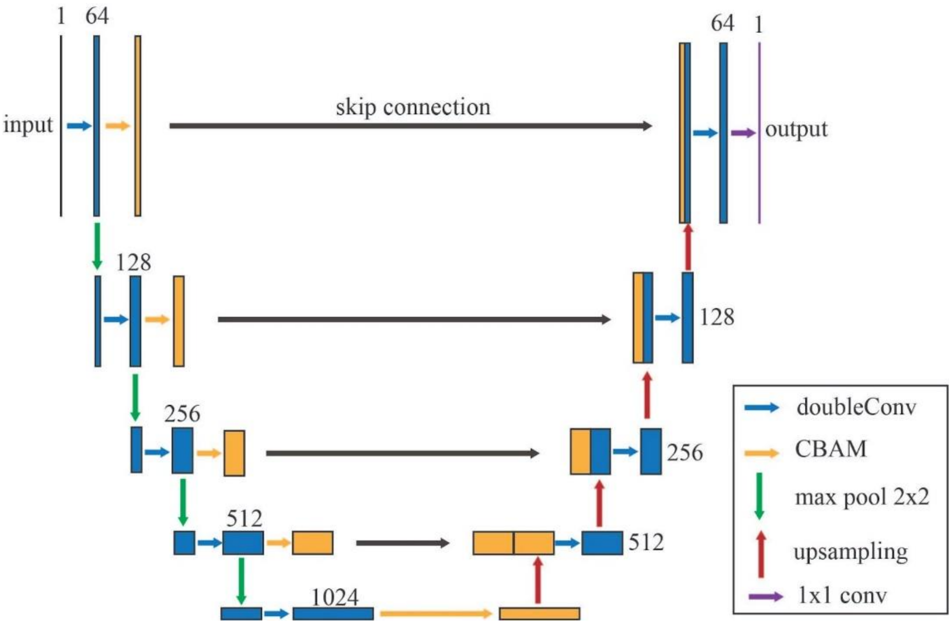

Fig. 2. The architecture of BC-Unet. Arrows represent the operation pass. Blue boxes are two 3-by-3 convolutional layers. Numbers above blue boxes are the number of convolutional channels. Green and red arrows represent max-pooling and up-sampling, respectively. Yellow boxes are refined images after applying CBAM. The black arrows represent layer concatenations.

e. Training

Mean squared error (MSE) is used as the loss function. The maximum epoch was set to 50. However, we employed an early stopping criterion to stop the training process when the validation loss did not increase in the last 30 epochs. Therefore, the maximum epoch was not reached. A learning rate scheduler was applied to reduce the learning rate by a factor of 0.1 when the validation loss did not increase for four epochs. The initial learning rate was set to 0.001. The Adam optimizer (Kingma and Ba, 2014) is used with default values. The training was done on a single NVIDIA Tesla P100 GPUs and took about three hours to finish. Training and validation losses are shown in Figure 3.

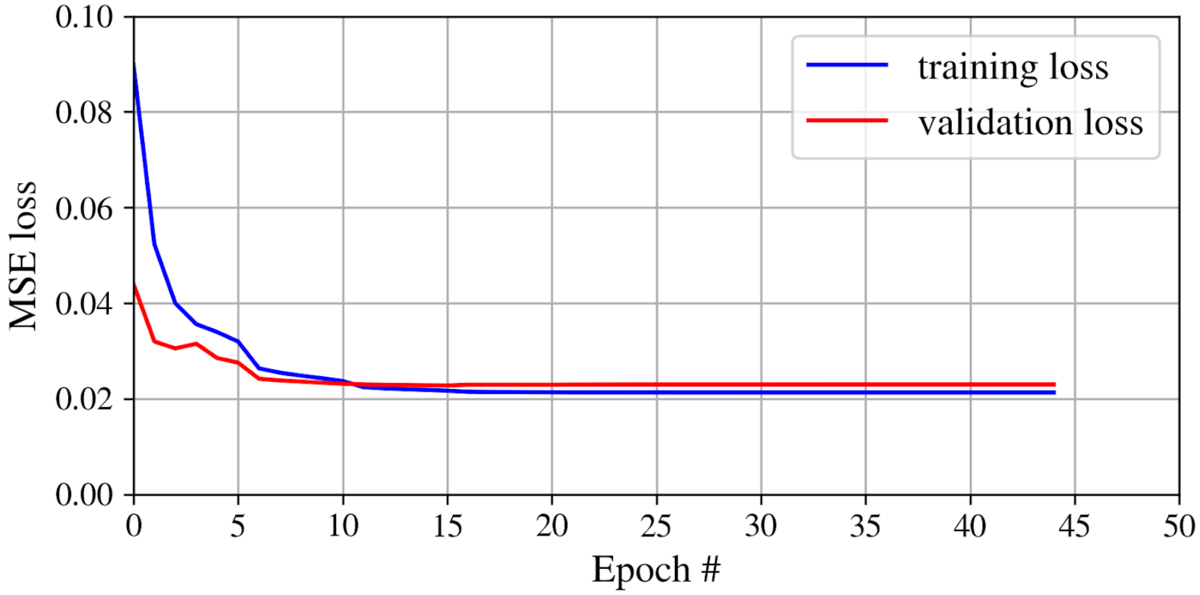

Fig. 3. Loss values as a function of the epoch.

f. Verification metrics

1) MODEL-MODEL COMPARISON

The performance of the bias correction model was evaluated with respect to the following metrics:

$$BIAS = \frac{1}{T \times N \times M} \sum_{t=1}^{T} \sum_{i=1}^{N} \sum_{j=1}^{M} (x_{t,i,j} - y_{t,i,j}) \quad (1)$$

$$RMSE = \sqrt{\frac{1}{T\times N\times M}\sum_{t=1}^{T}\sum_{i=1}^{N}\sum_{j=1}^{M}(x_{t,i,j} - y_{t,i,j})^2} \tag{2}$$

wherein $x_{t,i,j}$ and $y_{t,i,j}$ denote the raw/corrected forecast and ground truth at grid point (i,j) at time t, respectively. T is the total number of forecasts used in the testing, which is 608 forecasts for all cycles and 152 for each cycle, respectively. N is the dimension of the latitude, and M is the dimension of the longitude.

2) MODEL-OBSERVATION COMPARISON

The model outputs (GFS, BC-Unet, and ERA5) were interpolated to observation stations with the nearest-neighbor interpolation. The bias was calculated as follows:

$$BIAS = \frac{1}{T\times n}\sum_{t=1}^{T}\sum_{i=1}^{n}(x_{t,i} - y_{t,i}) \tag{3}$$

wherein $x_{t,i}$ and $y_{t,i}$ denote the raw/corrected forecast and observation at a station location i and at time t, respectively; n is the number of stations.

3. Results

a. Model-model comparison

To measure the model performance, we compared 2-m temperature from GFS and BC-Unet with reanalysis from ERA5 on the out-of-sample data from 1 January to 31 May 2024, a period which is not used in the training. The spatial and temporal mean bias and RMSE over CONUS for forecast hour 72 from all four cycles (00 UTC, 06 UTC, 12 UTC, 18 UTC) are shown in Table 1. The raw GFS has a cold bias of 0.33 °C over the CONUS. BC-Unet greatly reduced the CONUS cold bias to 0.04 °C. Systematic biases vary geographically in the GFS, with large cold biases over northeast and northwest, while large warm biases over Great Plains (Figure 4a). The BC-Unet performed well over these large bias regions, reducing the mean absolute error below 1 °C (Figure 4b). However, the BC-Unet overestimated 2-m temperature for the southwest of the CONUS. BC-Unet reduced RMSE all over the CONUS (Figure 5a, b). The overall mean RMSE was reduced by 0.4 °C. The model weights trained for forecast hour 72 worked well on other forecast hours. Both bias and RMSE were consistently reduced at all forecast hours (Figure 6a, b). The mean bias was more or less consistent from forecast hour 6 to forecast hour 72, then gradually increased in magnitude from forecast hour 72 to forecast hour 144 (Figure 6b).

Model	Bias (°C)	RMSE (°C)
GFS	-0.33	2.16
BC-Unet	-0.04	1.76

Table 1. Overall performance over CONUS with respect to bias, and RMSE for forecast hour 72.

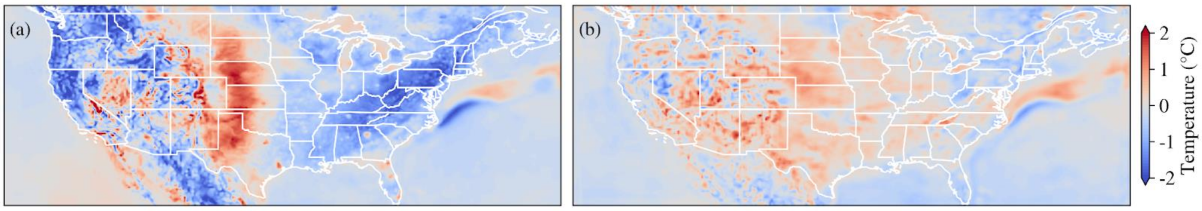

Fig. 4. Spatial distribution of mean bias of 2-m temperature for (a) original GFS, (b) corrected GFS for five months (January–May 2024) for forecast hour 72.

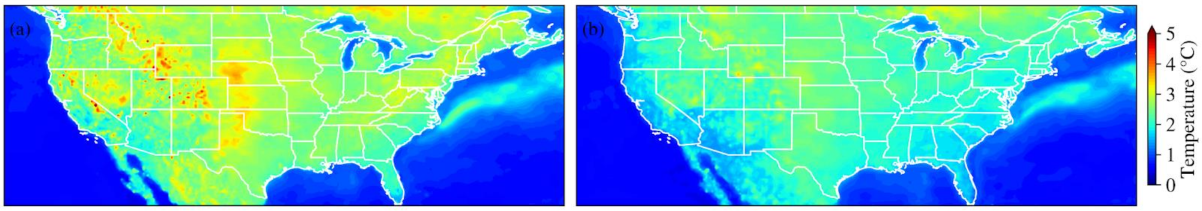

Fig. 5. Spatial distribution of mean RMSE of 2-m temperature for (a) original GFS, (b) corrected GFS for five months (January-May, 2024) for forecast hour 72.

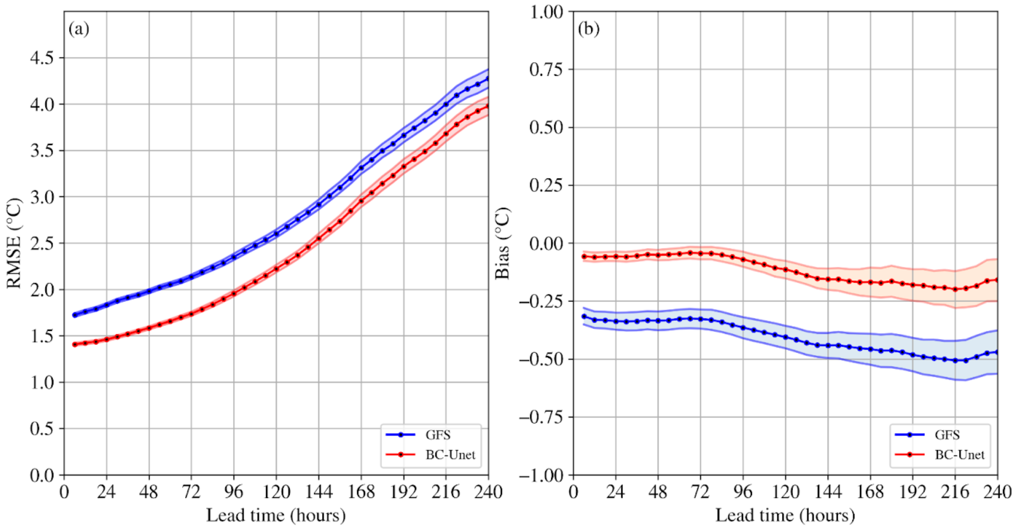

Fig. 6. Domain-averaged (a) RMSE and (b) bias compared with ERA5, as a function of 0- to 240-h forecast lead time for all forecast cycles averaged over 1 January through 31 May 2024. Blue lines are for the comparison between GFS and ERA5, and red lines are for the comparison between BC-Unet and ERA5. Shading extends from the 5th to 95th percentiles with the mean indicated by the solid line with markers.

Figure 7 presents the mean RMSE and bias for GFS and BC-Unet as a function of the model forecast hours from 6 to 240 for forecast cycle 00 UTC. The temperature biases exhibit a strong diurnal pattern. Compared with ERA5, GFS exhibits a noteworthy cold bias at the majority of valid times (Figure 7b), with a CONUS cold bias at 00 UTC as large as -1.0 °C, and a slight warm bias at 12 UTC for the first 6 days. BC-Unet reduced cold biases at 00, 06, and 18 UTC, with 00 UTC having the greatest reduction, up to 0.8 °C. For the short-range forecasts, weak warm biases at 12 UTC have also been corrected. It is worth noting that the BC-Unet did not remove the tendency of the cold bias to grow with increasing forecast lead time. The RMSE of GFS 2-m temperature increases with forecast lead time, from ~1.8 °C at FH06 to ~4.4 °C at FH240, with superposed diurnal variations. The BC-Unet reduced the RMSE considerably, from ~1.4 °C at FH06 to ~4.1 °C at FH240.

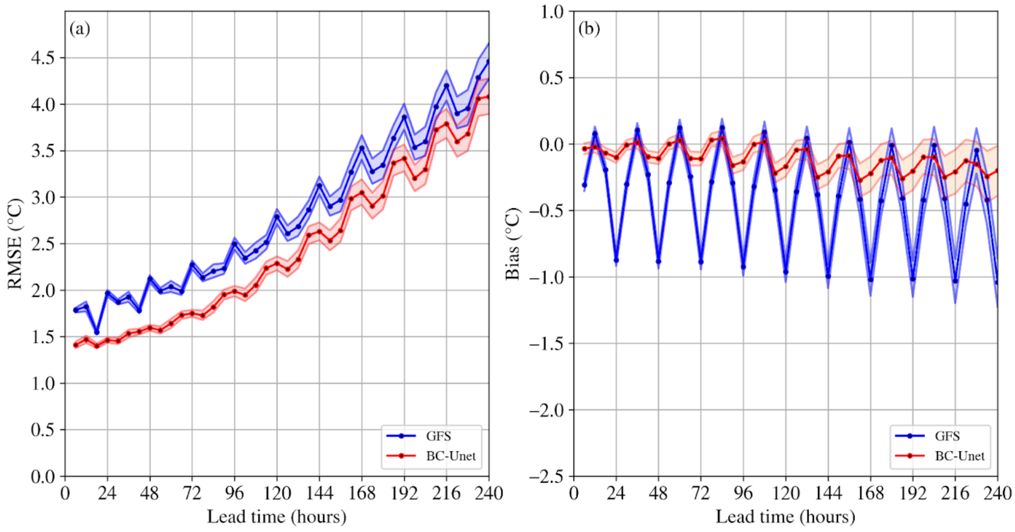

Fig. 7. Domain-averaged (a) RMSE and (b) bias compared with ERA5, as a function of 0- to 240-h forecast lead time for 00 UTC cycle averaged over 1 January through 31 May 2024. Blue lines are for the comparison between GFS and ERA5, and red lines are for the comparison between BC-Unet and ERA5. Shading extends from the 5th to 95th percentiles with the mean indicated by the solid line with markers.

b. Model-observation comparison

Figure 8a shows a comparison of mean 2-m temperature over CONUS in a function of forecast hour for 00 UTC cycles. There is a noteworthy reduction in CONUS 2-m temperatures from the late afternoon to evening (i.e., 18 UTC to 00 UTC) in the GFS forecast that is not present in ERA5 or observations. BC-Unet resolved this reduction in the GFS and matched well with ERA5 and observations. The mean bias of model-observation comparison shows a similar diurnal pattern to model-model comparison (Figure 8b), with GFS having stronger diurnal variations than ERA5. Both GFS and ERA5 exhibit cold biases, with the

GFS cold bias being considerably much larger. The largest daily cold bias in GFS forecasts occurs at 00 UTC, varying from 1.5 °C to 1.8 °C. The diurnal variations in ERA5 are less pronounced, with the largest daily cold bias (~0.2 °C) occurring at 18 UTC. Overall, the BC-Unet performs well in correcting GFS temperature biases over CONUS, especially for 00 UTC. BC-Unet has a slight 2-m temperature warm bias at 06 UTC valid times from day 1–5, in contrast to the cold bias seen in GFS forecasts at the same valid time.

Figures 9-12 displays the geographical distributions of the mean bias of 2-m temperature for the 06 (FH 6), 12 (FH 12), 18 (FH 18), and 00 (FH 24) UTC valid times, respectively, for the 00 UTC cycle. The GFS 2-m temperature biases exhibit notable spatial and temporal variations. At 06 UTC, GFS has large cold biases (< -1.5 °C) over the Appalachian Mountains, the valleys of California and Washington, and the Rocky Mountains, but large warm biases (> 1.5 °C) over the Great Plains (Figure 9a). ERA5 has relatively smaller biases in the range of 0.5 °C over the Appalachian Mountains and Great Plains, but a combination of both warm and cold biases over the Rocky Mountains (Figure 9c). The spatial distribution of BC-Unet is very close to that of ERA5. BC-Unet reduces the largest cold and warm biases over the Appalachian Mountains and Great Plains, respectively, but overestimates warm biases for some stations over the southeastern CONUS (Figure 9b). At 12 UTC, the GFS cold biases over the Appalachian Mountains are reduced from those at 6 UTC (Figure 10a). However, GFS warm biases over the Great Plains expands to the north and east at 12 UTC. As a result, the spatial-averaged bias over CONUS shows a warm bias in GFS (Figure 8). There is no significant change in ERA5 bias at 12 UTC compared with 06 UTC (Figure 10c). The bias of BC-Unet is still closer to the ERA5 than GFS (Figure 10b), with spatial-averaged bias over CONUS close to 0 °C (Figure 8b). At 18 UTC, GFS warm biases over the Great Plains and western CONUS have either decreased or become cold biases, and the cold biases over the Appalachian Mountains reappear (Figure 11a). ERA5 has relatively low biases over the Great Plains and eastern CONUS (Figure 11c). However, stations near bodies of water (e.g., Gulf Coast, East Coast, Great Lakes) show relatively large cold biases. This may be due to the 0.25^o resolution of ERA5 being unable to capture small-scale sea breezes along the coastlines. BC-Unet generates a 2-m temperature bias spatial pattern that is similar to ERA5 (Figure 11b). At 00 UTC, GFS has large cold biases over the majority of the CONUS except over the Great Plains (Figure 12a). ERA5 and BC-Unet share similar 2-m temperature bias spatial patterns with relatively small biases over the eastern CONUS (magnitudes less than 0.5 °C) and larger cold biases appear over the western CONUS (Figures 12b, c).

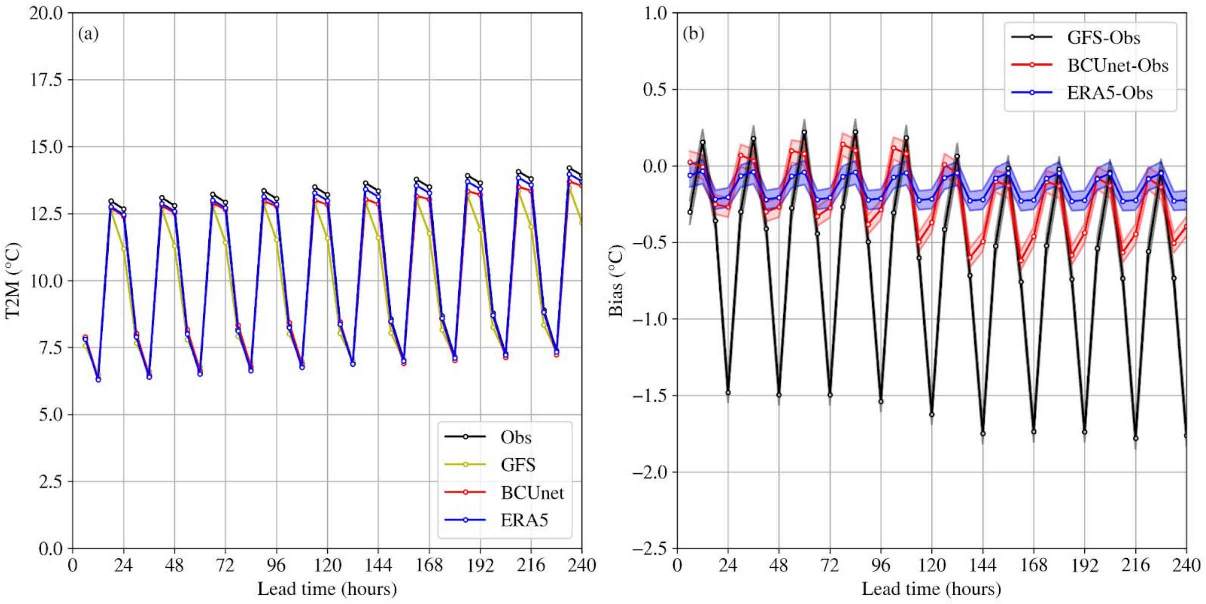

Fig. 8. Station-averaged (a) 2-m temperature (°C) and (b) bias (model minus observations, °C) as a function of 0- to 240-h forecast lead time for 00 UTC cycle averaged over 1 January through 31 May 2024. (a) Lines are mean 2-m temperature for observations (black), GFS (yellow), BC-Unet (red), and ERA5 (blue), respectively. (b) The black line is for the comparison between GFS and observation, the red line is for the comparison between BC-Unet and observation, and the blue line is for the comparison between ERA5 and observation. Shading extends from the 5th to 95th percentiles with the mean indicated by the solid line with markers.

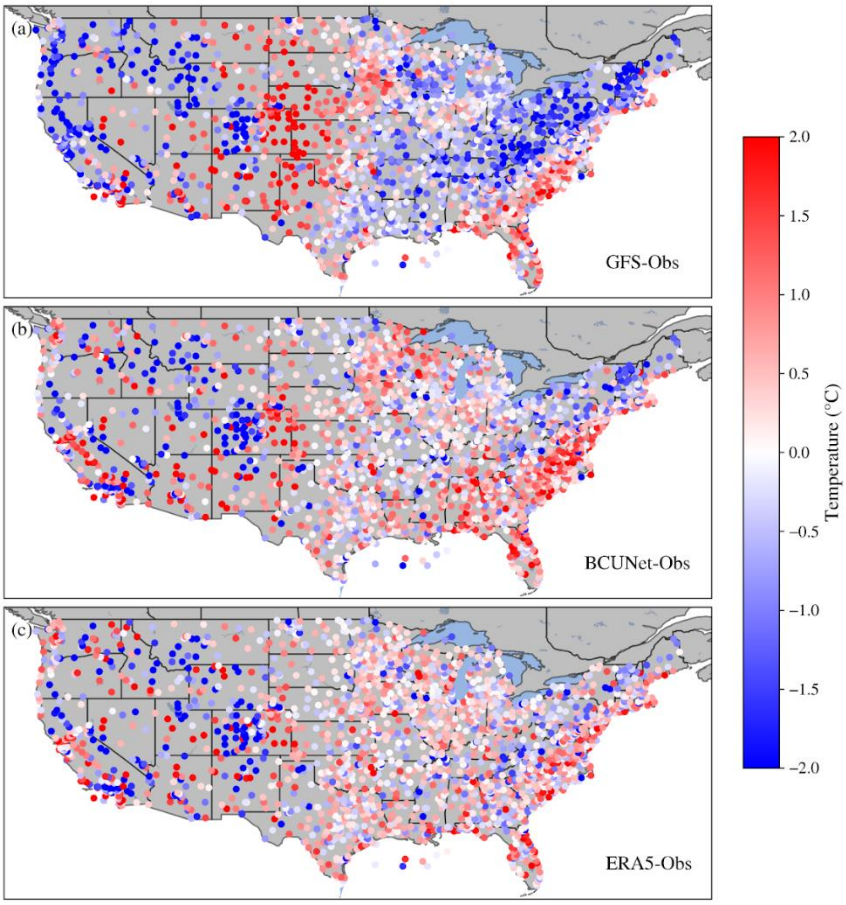

Fig. 9. Geographical distribution of biases (model minus observation) for forecast hour 06, valid at 06 UTC. Red/blue shading indicates that the model is too warm/cold in the 1 January 2024 to 31 May 2024 time frame.

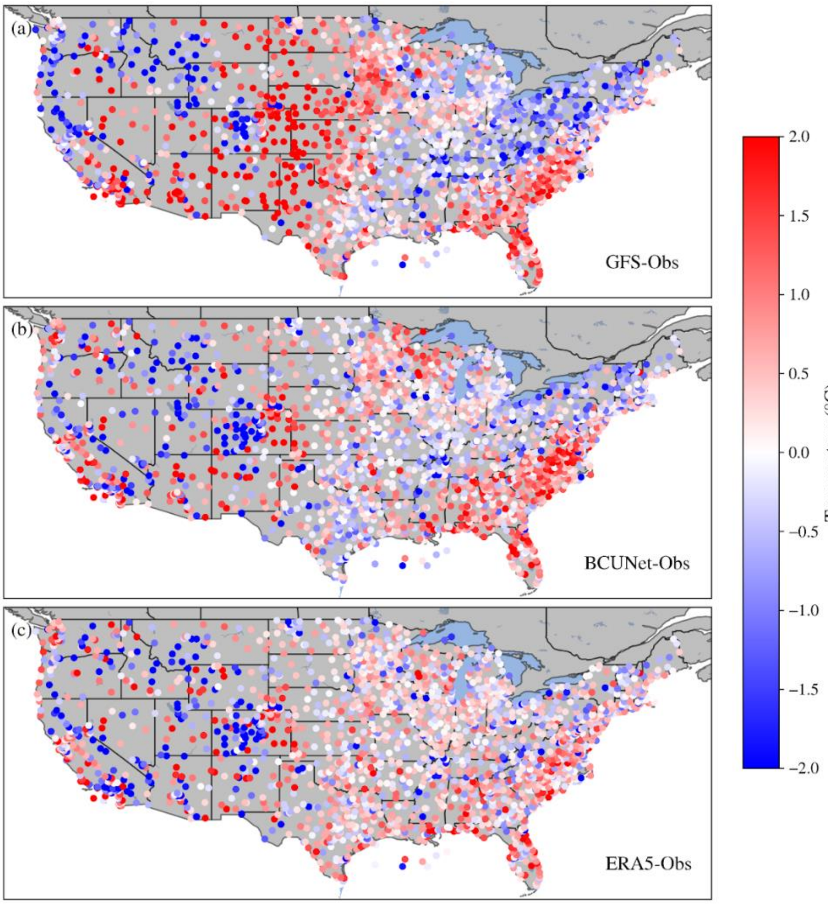

Fig. 10. Geographical distribution of biases (model minus observation) for forecast hour 12, valid at 12 UTC. Red/blue shading indicates that the model is too warm/cold in the 1 January 2024 to 31 May 2024 time frame.

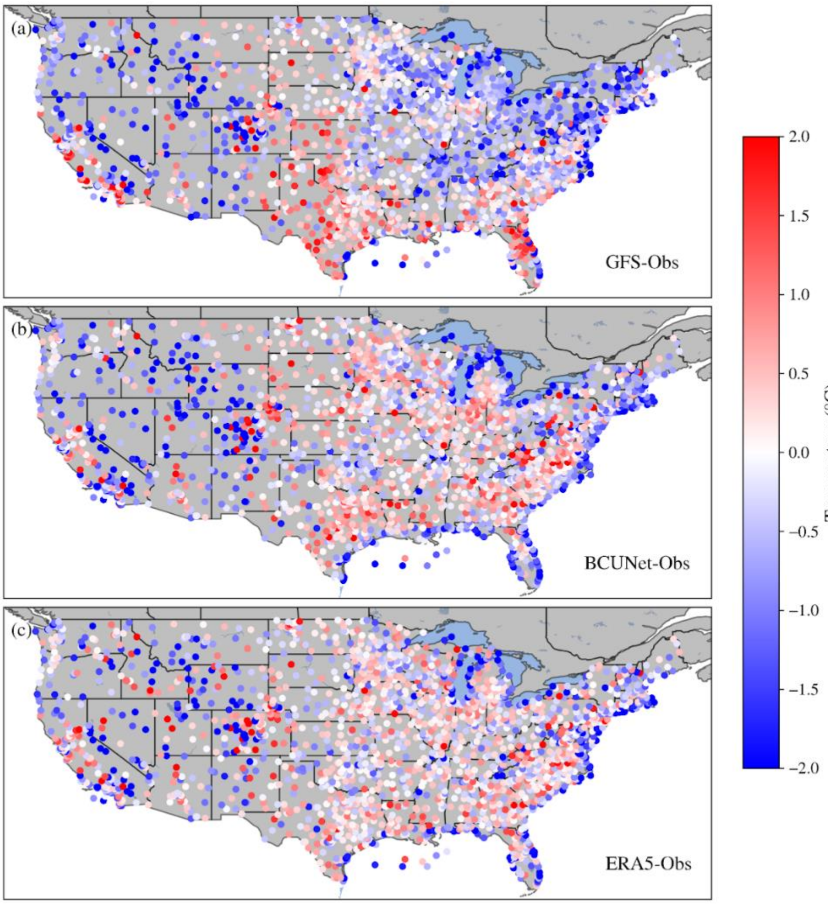

Fig. 11. Geographical distribution of biases (model minus observation) for forecast hour 18, valid at 18 UTC. Red/blue shading indicates that the model is too warm/cold in the 1 January 2024 to 31 May 2024 time frame.

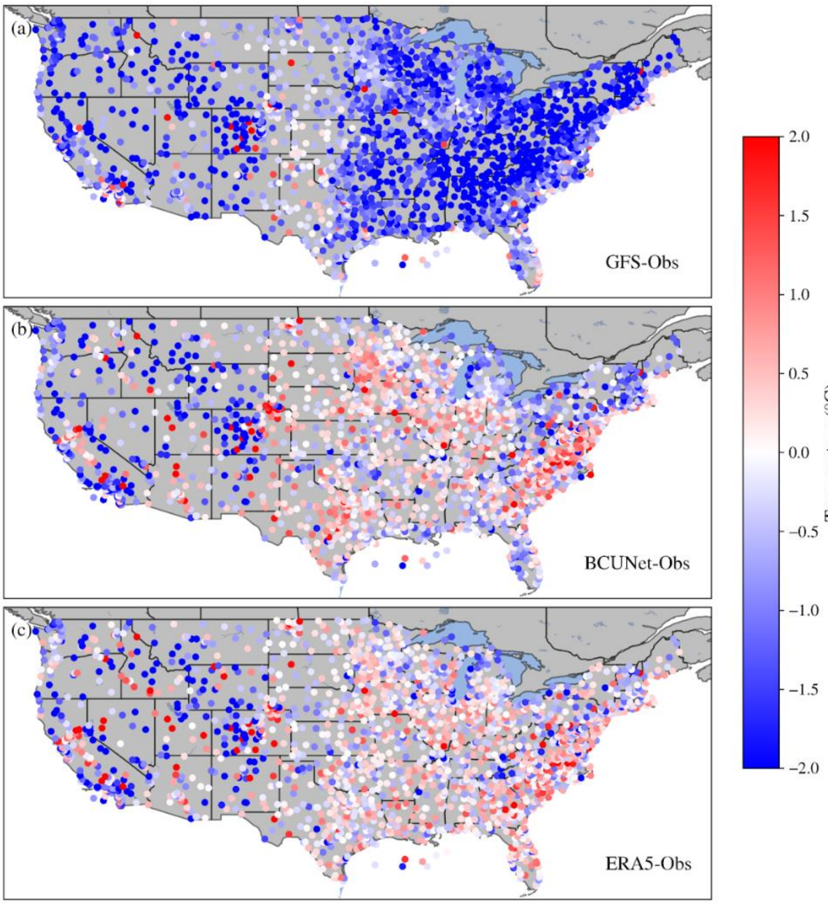

Fig. 12. Geographical distribution of biases (model minus observation) for forecast hour 24, valid at 00 UTC. Red/blue shading indicates that the model is too warm/cold in the 1 January 2024 to 31 May 2024 time frame.

To further explore the spatial distribution of 2-m temperature forecast biases, we calculated averaged biases for four CONUS subregions where obvious warm or cold biases from GFS were observed. These four subregions include the west, central, northeast, and southeast portions of CONUS (Figure 1b). Figure 13 presents station-averaged 2-m temperature bias as a function of forecast hours for each CONUS subregion. GFS shows strong diurnal variations in the 2-m temperature bias in all four subregions, whereas ERA5 has relatively weak diurnal variations in bias over the west and southeast CONUS subregions (Figures 13 a, d) and almost zero 2-m temperature bias over central and northeast regions (Figures 13 b, c). Over the western CONUS, both GFS and ERA5 exhibit cold biases (Figure 13a). GFS has close to zero bias at 12 UTC, but cold biases at other valid times, with the cold bias reaching -1.8 °C at 00 UTC. BC-Unet shows similar diurnal variations to ERA5, reducing the cold bias to -1.0 °C at 00 UTC. Over the central CONUS (which includes the Great Plains), the GFS bias varies between approximately -1.0 and 1.0 °C, with the largest cold bias at 00 UTC (late evening locally) and the largest warm bias at 12 UTC (early morning locally) (Figure 13b). BC-Unet has a slight warm bias and weak diurnal variations relative to the GFS over the central CONUS. In contrast, ERA5 exhibits almost no diurnal cycle. For the northeast CONUS subregion, GFS exhibits cold biases at all valid times, with the cold bias reaching -3.0 °C at 00 UTC for the longer forecast lead time (Figure 13c). BC-Unet is similar to ERA5 for the first five days with a 2-m temperature bias close to zero. The BC-Unet cold bias increases slightly at longer lead times, but is still considerably less than the GFS cold bias. Over the southeast CONUS subregion, GFS shows a weak warm bias at 06 UTC and 12 UTC (the overnight and early morning hours) and cold bias at 18 UTC and 00 UTC (the afternoon and evening). This diurnal cycle in 2-m temperature bias is similar to that of the central CONUS (Figure 13b), but the magnitude of warm bias is considerably smaller over the southeastern CONUS (Figure 13d). BC-Unet bias is much more similar to the ERA5 bias than the GFS bias over the southern CONUS, with relatively weak diurnal variations.

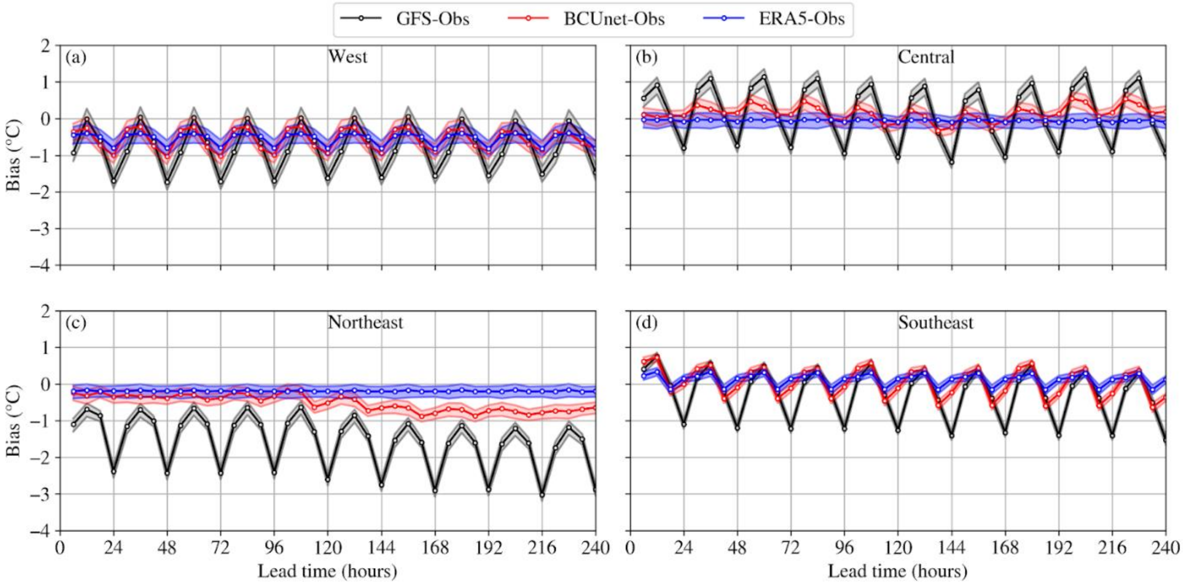

Fig. 13. Station-averaged bias (model minus observations) as a function of 0- to 240-h forecast lead time for 00 UTC cycles averaged over 1 January through 31 May 2024 for (a) west, (b) central, (c) northeast, and (d) southeast CONUS. Black lines show the comparison between GFS and observations, red lines show the comparison between BC-Unet and observations, and blue lines show the comparison between ERA5 and observations. Shading extends from the 5th to 95th percentiles with the mean indicated by the solid line with markers.

4. Discussion and Summary

NCEP GFS forecasts have longstanding systematic biases in 2-m temperature over land, which exhibits temporal and geographical patterns. The main reason for these biases is that near-surface temperature is controlled by a variety of processes, including radiative transfer, convection, diffusion, surface fluxes, and land-atmosphere coupling which may not be modeled correctly. Factors like humidity, cloud cover, and surface characteristics (vegetation, soil type, soil texture, etc.) can also impact the near-surface temperature (Sandu et al., 2020).

BC-Unet, a semantic-segmentation CNN, is proposed for correcting 2-m temperature biases in the operational GFS forecasts over CONUS. BC-Unet was trained with a single forecast hour (FH72) with GFS forecasts as inputs and ERA5 as ground truth. The performance of BC-Unet was evaluated with five months of forecast data from 1 January through 31 May 2024. For FH72, BC-Unet considerably reduced the 2-m temperature cold bias and RMSE when compared against ERA5. The domain-cycle-averaged statistics, the 2-m temperature RMSE and bias of GFS forecasts generally grow with forecast lead time and display strong diurnal variations. When focusing on just the 00 UTC cycle, BC-Unet reduced the mean 2-m temperature RMSE by up to 0.35 °C and the cold bias by up to 0.8 °C over CONUS.

BC-Unet, GFS, and ERA5 outputs were further compared with 2-m temperature observations. The domain-average 2-m temperature bias for the 00 UTC cycles examined shows the GFS has a large cold bias in the afternoon, evening, and overnight hours (18, 00, and 06 UTC), which may be related to the nocturnal stable boundary layer (Zheng et al., 2017). Geographically, GFS exhibits cold biases over western and northeastern CONUS at all valid times, while the 2-m temperature bias varies throughout the day over central and southeastern CONUS (with the largest cold bias at 00 UTC and largest warm bias at 12 UTC). ERA5 has a slightly negative to neutral bias over central and northeastern CONUS, exhibits a cold bias at all valid times over western CONUS, and a bias that fluctuates with the diurnal cycle over southeastern CONUS (with the largest cold bias at 18 UTC and largest warm bias at 12 UTC). For all four CONUS subregions, BC-Unet is able to reduce GFS 2-m temperature biases and more closely follows ERA5’s muted diurnal bias variations, which is not surprising, as BC-Unet is trained with ERA5 as ground truth. GFS and ERA5 have a pronounced cold bias over the western CONUS. The western CONUS contains complex orography including islands, basins, coastal areas, and mountains (Sha et al., 2020) and sparse observations compared with other CONUS subregions (Figure 1b). These factors may contribute to large cold biases in ERA5, preventing training on ERA5 from resolving the GFS cold bias in BC-Unet. BC-Unet could benefit from including observation data or training with a higher resolution analysis dataset, such as the UnRestricted Mesoscale Analysis (De Pondeca et al., 2011; Morris et al., 2020).

Another limitation in the current configuration of BC-Unet training is that the BC-Unet 2-m temperature bias still seems to grow as a function of the forecast lead time. This may be due to the fact that only a single forecast hour (F72) was used in the training of BC-Unet because of the GPU memory limitation at the time of this study. Future work could consider training on all forecast hours. Another possible solution could be using Long Short-Term Memory (LSTM), a type of recurrent neural network, which shows skill at time series forecasting.

While the results presented here address only 2-m temperature biases over CONUS, the model can be transferable to other crucial forecasting variables when NWP model errors and biases are also large, [e.g., 2-m relative humidity, Convective Available Potential Energy (CAPE), and Convective Inhibition (CIN)]. Given the training dataset size is small and there are fewer variables compared with ML models for global weather prediction, it is very efficient to train and inference the model.

Acknowledgments.

The authors thank Alicia Bentley and Wei Li at the NOAA Environmental Modeling Center for internal reviews. The authors acknowledge the high performance computing resources provided by the NOAA Research and Development High Performance Computing Program. The project described in this article was supported by the Inflation Reduction Act and the NOAA Software Engineering for Novel Architectures (SENA) project. The scientific results and conclusions, as well as any views or opinions expressed herein, are those of the authors and do not necessarily reflect the views of NOAA or the Department of Commerce.

Data Availability Statement.

GFS model output is available from NOAA via Amazon Web Services at https://noaa-gfs-bdp-pds.s3.amazonaws.com/index.html. ERA5 model output is available at https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=form. NCEP ADP Global Surface Observations is available at https://rda.ucar.edu/datasets/d461000/. Code are publicly available at https://github.com/NOAA-EMC/ML4BC.

REFERENCES

Alerskans, E., and Kaas, E., 2021: Local temperature forecasts based on statistical post-processing of numerical weather prediction data. *Meteorological Applications*, *28*(4), e2006. https://doi.org/10.1002/met.2006

Bhargava, K., Kalnay, E., Carton, J. A., and Yang, F., 2018: Estimation of Systematic Errors in the GFS Using Analysis Increments. *Journal of Geophysical Research: Atmosphere, 123*, 1626-1637. https://doi.org/10.1002/2017JD027423

Cheng, W. Y.Y., and Steenburgh, J. W., 2007: Strengths and Weaknesses of MOS, Running-Mean Bias Removal, and Kalman Filter Techniques for Improving Model Forecasts over the Western United States. *Weather and Forecasting, 22,* 1304-1318. https://doi.org/10.1175/2007WAF2006084.1

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*(5), 593-612. https://doi.org/10.1175/WAF-D-10-05037.1

EMC Model Evaluation Group, 2020: GFSv16 MEG Evaluation Overview. *Retrieved August 2, 2024, from https://www.emc.ncep.noaa.gov/users/meg/ gfsv16/pptx/MEG_9-24-20_GFSv16_MEG_Eval_Overview.pptx*

Han, L., Chen, M., Chen, K., Chen, H., Zhang, Y., Lu, B., Song, L., and Qin, R., 2021: A Deep Learning Method for Bias Correction of ECMWF 24–240 h Forecasts. *Advances in atmospheric sciences*, *38*(9), 1444-1459. https://doi.org/10.1007/s00376-021-0215-y

Harris, L., Chen, X., Zhou, L., and Chen, J.-H., 2020: The Nonhydrostatic Solver of the GFDL Finite-Volume Cubed-Sphere Dynamical Core. *NOAA technical memorandum OAR GFDL*; 2020-003. https://doi.org/10.25923/9wdt-4895

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*, 1999-2049. https://doi.org/10.1002/qj.3803

Ioffe, S., and Szegedy, C., 2015: Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv. https://arxiv.org/pdf/1502.03167.pdf.*

Kingma, D. P., and Ba, J., 2014, December 22: Adam: A Method for Stochastic Optimization. *arXiv. Retrieved September 20, 2024, from https://arxiv.org/abs/1412.6980*

Kleist, D. T., and Ide, K., 2015: An OSSE-Based Evaluation of Hybrid Variational–Ensemble Data Assimilation for the NCEP GFS. Part II: 4DEnVar and Hybrid Variants. *Monthly Weather Review*, *143*, 452-470. https://doi.org/10.1175/MWR-D-13-00350.1

Kleist, D. T., Parrish, D. F., Derber, J. C., Treadon, R., Wu, W.-S., and Lord, S., 2009: Introduction of the GSI into the NCEP Global Data Assimilation System. *Weather and Forecasting*, *24*, 1691-1705. https://doi.org/10.1175/2009WAF2222201.1

Lin, S.-J., 2004: A "Vertically Lagrangian" Finite-Volume Dynamical Core for Global Models. *Monthly Weather Review*, *132*, 2293–2307. https://doi.org/10.1175/1520-0493(2004)132<2293:AVLFDC>2.0.CO;2

Maas, A. L., Hannun, A. Y., and Ng, A. Y., 2013: Rectifier Nonlinearities Improve Neural Network Acoustic Models. Stanford AI Lab. *Retrieved September 20, 2024, from https://ai.stanford.edu/~amaas/papers/relu_hybrid_icml2013_final.pdf*

McCollor, D., and Stull, R., 2008: Hydrometeorological Accuracy Enhancement via Postprocessing of Numerical Weather Forecasts in Complex Terrain. *Weather and Forecasting*, *23*, 131-144. https://doi.org/10.1175/2007WAF2006107.1

Monache, L. D., Nipen, T., Liu, Y., Roux, G., and Stull, R., 2011: Kalman Filter and Analog Schemes to Postprocess Numerical Weather Predictions. *Monthly Weather Review*, *139*(11), 3554-3570. https://doi.org/10.1175/2011MWR3653.1

Morris, M. T., Carley, J. R., Colón, E., Gibbs, A., De Pondeca, M. S. F. V., and Levine, S., 2020: A Quality Assessment of the Real-Time Mesoscale Analysis (RTMA) for Aviation. *Weather and Forecasting*, *35*(3), 977-996. https://doi.org/10.1175/WAF-D-19-0201.1

National Weather Service, 2021: Upgrade NCEP Global Forecast Systems (GFS) to v16. National Weather Service. *Retrieved September 20, 2024, from https://www.weather.gov/media/notification/pdf2/scn21-20gfs_v16.0_aac.pdf*

Patel, R. N., Yuter, S. E., Miller, M. A., Rhodes, S. R., Bain, L., and Peele, T. W., 2021: The Diurnal Cycle of Winter Season Temperature Errors in the Operational Global Forecast System (GFS). *Geophysical Research Letters*, *48*, e2021GL095101. https://doi.org/10.1029/2021GL095101

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

Ronneberger, O., Fischer, P., and Brox, T., 2015: U-Net: Convolutional Networks for Biomedical Image Segmentation. *2015 Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, N. Navab et al., Eds., Lecture Notes in Computer Science, 9351, Springer International*, 234-241. https://doi.org/10.1007/978-3-319-24574-4_28

Sandu, I., and Coauthors, 2020: Addressing near-surface forecast biases: outcomes of the ECMWF project 'Understanding uncertainties in surface atmosphere exchange' (USURF). *https://www.ecmwf.int/en/elibrary/81202-addressing-near-surface-forecast-biases-outcomes-ecmwf-project-understanding*

Sha, Y., Gagne, D. J., West, G., and Stull, R., 2020: Deep-Learning-Based Gridded Downscaling of Surface Meteorological Variables in Complex Terrain. Part I: Daily Maximum and Minimum 2-m Temperature. *Journal of Applied Meteorology and Climatology*, *59*, 2057–2073. https://doi.org/10.1175/JAMC-D-20-0057.1

Trebing, K., Stanczyk, T., and Mehrkanoon, S., 2021: SmaAt-UNet: Precipitation Nowcasting using a Small Attention-UNet Architecture. *Pattern Recognition Letters*, *145*, 178-186. https://doi.org/10.1016/j.patrec.2021.01.036

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S., 2018: CBAM: Convolutional Block Attention Module. *arXiv. Retrieved September 20, 2024, from https://arxiv.org/abs/1807.06521*

Zheng, W., Ek, M., Mitchell, K., Wei, H., and Meng, J., 2017: Improving the Stable Surface Layer in the NCEP Global Forecast System. *Monthly Weather Review*, *145*, 3969-3987. https://doi.org/10.1175/MWR-D-16-0438.1