Comparing and Combining Deterministic Surface Temperature Postprocessing Methods over the United States

THOMAS M. HAMILL[a]

[a] *NOAA/Physical Sciences Laboratory, Boulder, Colorado*

(Manuscript received 13 February 2021, in final form 8 May 2021)

ABSTRACT: Common methods for the postprocessing of deterministic 2-m temperature (T_{2m}) forecasts over the United States were evaluated from +12- to +120-h lead. Forecast data were extracted from the Global Ensemble Forecast System (GEFS) v12 reforecast dataset and thinned to a $^1/_2°$ grid. Analyzed data from the European Centre/Copernicus reanalysis (ERA5) were used for training and validation. Data from the 2000–18 period were used for training, and 2019 forecasts were validated. The postprocessing methods compared were the raw forecast guidance, a decaying-average bias correction (DAV), quantile mapping (QM), a univariate model output statistics (uMOS) algorithm, and a multivariate (mvMOS) algorithm. The mvMOS algorithm used the raw forecast temperature, the DAV adjustment, and the QM adjustment as predictors. Forecasts from all the postprocessing methods reduced the root-mean-square error (RMSE) and bias relative to the raw guidance. QM produced forecasts with slightly higher error than DAV. DAV estimates were the most consistent from day to day. The uMOS and mvMOS algorithms produced statistically significant lower RMSEs than DAV at forecast leads longer than 1 day, with mvMOS exhibiting the lowest error. Taylor diagrams showed that the MOS methods reduced the variability of the forecasts while improving forecast-analyzed correlations. QM and DAV modified the distribution of forecasts to more closely exhibit those of the analyzed data. A main conclusion is that the judicious statistical combination of guidance from multiple postprocessing methods is capable of producing forecasts with improved error statistics relative to any one individual technique. As each method applied here is algorithmically relatively simple, this suggests that operational deterministic postprocessing combining multiple correction methods could produce improved T_{2m} guidance.

KEYWORDS: Bias; Regression analysis; Statistics; Forecast verification/skill; Numerical weather prediction/forecasting; Statistical forecasting

1. Introduction

Much of the attention in the recent literature on the statistical postprocessing of forecasts has shifted to the postprocessing of ensemble prediction system guidance and the production of skillful and reliable probabilistic forecasts. This is reflected in a recent textbook (Vannitsem et al. 2018) highlighting developments in this discipline. Despite the evolution in this direction, many weather prediction centers still produce deterministic forecast guidance from a variety of methods, especially forecasts of more statistically straightforward quantities such as surface temperature and particularly at shorter forecast lead times (days, not weeks). Hence, it is still of practical interest to operational weather prediction centers to understand the potential strengths and weaknesses of several plausible deterministic statistical postprocessing methods.

In this article we compare the characteristics of several algorithmically simple methods when applied to the statistical correction of 2-m above ground surface temperatures (T_{2m}) from a single forecast system. The algorithms are the decaying-average (DAV) bias correction (Cui et al. 2012), quantile mapping (QM; Hopson and Webster 2010; Voisin et al. 2010; Maraun 2013), and model output statistics (MOS) regression techniques (Glahn and Lowry 1972; Carter et al. 1989). While this is not an exhaustive list, these represent different techniques with different underlying correction principles, and each is used operationally in different contexts. In fact, in the U.S. National Weather Service, each of these is used. The DAV method is used in the National Blend of Models (NBM; Craven et al. 2020). QM is also used in the NBM for precipitation forecasts (Hamill et al. 2017; Hamill and Scheuerer 2018), and MOS is still used for station data postprocessing (e.g., Glahn et al. 2009).

At many prediction centers, improved products may also be generated through the combination of output from multiple prediction systems, with the implicit assumption that biases are quasi-independent and their average reduces error and provides more realistic estimates of forecast uncertainty. See, for example, Bougeault et al. (2010), Yamaguchi et al. (2012), and Liu and Xie (2014). Often the individual prediction systems' guidance are statistically adjusted as well, as in Vislocky and Fritsch (1995), Krishnamurti et al. (2000), Raftery et al. (2005), and Hamill (2012). While multimodel combination and calibration are valid approaches, the intent of this article is exploration of statistical adjustments applied to a single prediction system.

Long time series of forecasts and observations/analyses are ideal for postprocessing. Without them, approximations may be necessary, such as bolstering the training set with data from "supplemental locations" (Hamill et al. 2017) or pooling of training data over locations spanning large regions (Lowry and Glahn 1976). With newly available global reforecasts from version 12 of the NWS Global Ensemble Forecast System (Hamill et al. 2021, manuscipt submitted to *Mon. Wea. Rev.*; Zhou et al. 2021, manuscript submitted to *Wea. Forecasting*; H. Guan et al. 2021, submitted to *Mon. Wea. Rev.*), there is a

Corresponding author: Thomas M. Hamill, tom.hamill@noaa.gov

DOI: 10.1175/MWR-D-21-0027.1

long-enough training dataset that such approximations are not necessary for surface temperature, and each grid point can be processed using only that point's data for training, simplifying the algorithm development. In particular, this study independently evaluated the raw and postprocessed forecasts over a set of grid points 1/4° in size, separated by 1/2°, in a domain encompassing the contiguous United States (CONUS). The 2000–18 T_{2m} forecast and reanalysis data were used as training data, and the forecasts were validated during 2019. A full cross validation again was omitted to minimize computational expense (calculations were performed on a home desktop computer during COVID-19). These multiple postprocessing methods were evaluated with common (root-mean-square error, bias) metrics as well as those less commonly applied to weather predictions such as "Taylor diagrams" (Taylor 2001). The hope is that the results will guide the choice of algorithms and their possible combination in future operational weather postprocessing.

Below, section 2 discusses the data used in this study as well as the postprocessing methods and the methods of evaluation. Section 3 provides results, and section 4 concludes.

2. Data, postprocessing, and evaluation methods

a. Forecast data

Gridded T_{2m} reforecasts from the U.S. National Weather Service Global Ensemble Forecast System, version 12 (GEFSv12) were used in this study. The ensemble forecast system was described in Zhou et al. (2021, manuscript submitted to *Wea. Forecasting*), the reforecast data were described in H. Guan et al. (2021, submitted to *Mon. Wea. Rev.*) and the reanalyses used to initialize the reforecasts were described in Hamill et al. (2021, manuscript submitted to *Mon. Wea. Rev.*). Briefly, v12 of the GEFS provides a major system upgrade; the ensemble prediction system uses a new finite-volume dynamical core, there are major improvements to the deterministic and stochastic physics, and the grid spacing has been refined to ~25 km. Ensemble prediction skill is improved in many ways, as described in Zhou et al. (2021, manuscript submitted to *Wea. Forecasting*). The real-time ensemble is accompanied by a reforecast dataset spanning 2000–19 (which is available for free download from Amazon web services, https://noaa-gefs-retrospective.s3.amazonaws.com/index.html). During this period, for each day at 0000 UTC, a 5-member reforecast ensemble was generated to +16-day lead, one control, and four perturbed members. Once per week a larger 11-member ensemble was generated to +35 day, a control, and 10 perturbed members.

For this simple study, we examined only the deterministic control member from this reforecast ensemble. While data were available on a 1/4° grid, the data were subsampled to 1/2°, confined to a domain from 125° to 60°W and from 20° to 50°N. These choices were made to reduce the computational expense and storage requirements. The author has no reason to expect this subsampling to affect the results because the spatial correlation length scale of surface temperature errors is much larger than the grid spacing (Hamill and Scheuerer 2020). The domain encompassed the contiguous United States and included some of Mexico, southern Canada, and the Caribbean

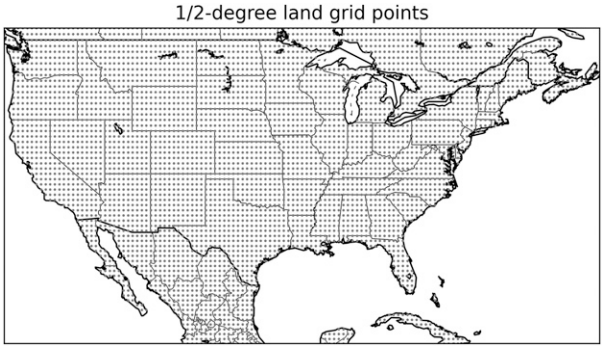

FIG. 1. Grid points (dots) used for the evaluation of postprocessing methods in this study.

(Fig. 1). Only the grid points in this domain that were denoted with >50% land in both the forecast and analyzed data were considered; there were some oddities where the forecast and analyzed data differed in their land–water classifications, and this profoundly affected the statistics. Forecasts were evaluated from +12- to +120-h lead time in time steps of 12 h. Beyond +120 h, deterministic forecasting is inappropriate.

b. Analysis data

Coincident "ERA5" reanalyses (Hersbach et al. 2020) from the European Centre for Medium-Range Weather Forecasts (ECMWF)/Copernicus Climate Service reanalysis were downloaded and used for statistical model training and validation. The data were extracted on a 1/4° grid and subsampled to the 1/2° grid, coincident in space with the forecast grid to reduce computational expense and storage. Data were extracted at 12-h intervals from the beginning of 2000 to the end of January 2020. ERA5 employs a T_{2m} analysis procedure using station observations, and it was thus deemed to be a reasonably trustworthy gridded reference product.

c. The decaying-average bias correction

This method will be abbreviated as "DAV" hereafter. The method has previously been described in Cui et al. (2012). The approach is quite simple, both algorithmically and in terms of implementation. The application developer chooses a value α that determines the weighting to apply to the most recent discrepancy between forecast and observation (or analysis). For a forecast date t for a particular forecast lead time and grid point, the DAV bias estimate is

$$\hat{b}_t^{\mathrm{DAV}} = (1-\alpha)\,\hat{b}_{t-1}^{\mathrm{DAV}} + \alpha(f_t - a_t), \tag{1}$$

where $\hat{b}_{t-1}^{\mathrm{DAV}}$ is the bias estimate at the same lead time and grid point but 1 day previous, and f_t is the sample forecast value of a random variable X_f and a_t is the sample analyzed value at date t. Some particularly appealing characteristics of DAV are as follows: (i) training may be conducted on-the-fly; one need not conduct a separate training, followed by validation. (ii) Because of this, storage of training data in an operational environment is not necessary. When considering high-resolution grids over large areas and spanning multiple forecast variables, multiple

lead times, and lengthy training periods, this storage can become quite large. Some disadvantages of the DAV method were discussed in Hamill (2018), in particular the difficulty in choosing an optimal value of α in the presence of time-varying unconditional bias.

The error of the DAV method was only slightly sensitive to the chosen value of α. Fig. 2 shows the RMSE of the DAV method during the 2000–18 training period as a function of α. Higher α produced the lowest RMSE for shorter leads, and smaller α for longer leads. Why? At the shorter leads, random error from chaotic processes are comparatively small, so fewer recent samples are sufficient to estimate the bias. Further, the biases may be related to quickly varying state aspects such as soil moisture, and thus past discrepancies between analysis and forecast in the more distant past are de-weighted. At longer leads with their larger random errors, the weighted averaging based on a longer time series of bias estimates (Cui et al. 2012) produced with the smaller α is beneficial. For the validation in 2019 against other techniques, the α that produced the lowest RMSE at each forecast lead time during the training period was chosen.

d. *Quantile mapping*

Let the cumulative distribution function (CDF) for the forecast at a particular gridpoint location and time be denoted by

$$F_f(V) = P(X_f \le V), \tag{2}$$

where X_f is again the temperature forecast random variable at time t, and V is a specific temperature value; $0.0 \le F_f(X_f)\,[= p] \le 1.0$. The CDFs for a given grid point and month are estimated from a long time series of data, described later. There is a one-to-one mapping between a forecast value and its cumulative probability. We define a quantile function that maps a cumulative probability p back to the forecast temperature variable:

$$X_f = F_f^{-1}(p). \tag{3}$$

The quantile-mapping (QM) procedure thus maps the forecast temperature sample to its cumulative probability in the climatological distribution of forecasts and then applies the quantile function (also known as the percent-point function) for the analyzed distribution:

$$\hat{a}_t^{\mathrm{QM}} = F_a^{-1}[F_f(f_t)]. \tag{4}$$

In this way QM estimates an analyzed value sharing the same cumulative probability relative to its analyzed climatological distribution as the sample forecast value to the climatological forecast distribution. The bias estimate is then $\hat{b}_t^{\mathrm{QM}} = f_t - \hat{a}_t^{\mathrm{QM}}$.

The CDFs for forecast and analyses at many grid points were characteristic of non-Gaussian distributions. After some experimentation, a three-component Gaussian mixture model was chosen to represent the CDFs instead of a one-component Gaussian or other parametric distribution. It used the python module scikit-learn.mixture. This module determined weights, means, and standard deviations associated with three Gaussian kernels whose weighted sum provided the closest fit to the empirical distributions of forecasts (or analyzed) data. An example of the fitted distributions and probability–probability (P–P) plots (Wilks 2011, his section 4.5.2) are provided in Figs. 3 and 4, respectively. At this grid point and at many others examined, the fitted CDFs appear to produce accurate parametric representations of the empirical CDFs. Different distributions were estimated for each grid point, forecast month, and lead time using 2000–18 data and the month of interest including the data from ±1 month. For example, +120-h forecast CDFs for the month of February are fit with January–March 2000–18 +120-h training data. See Wilks 2011 for interpretation of the P–P plots.

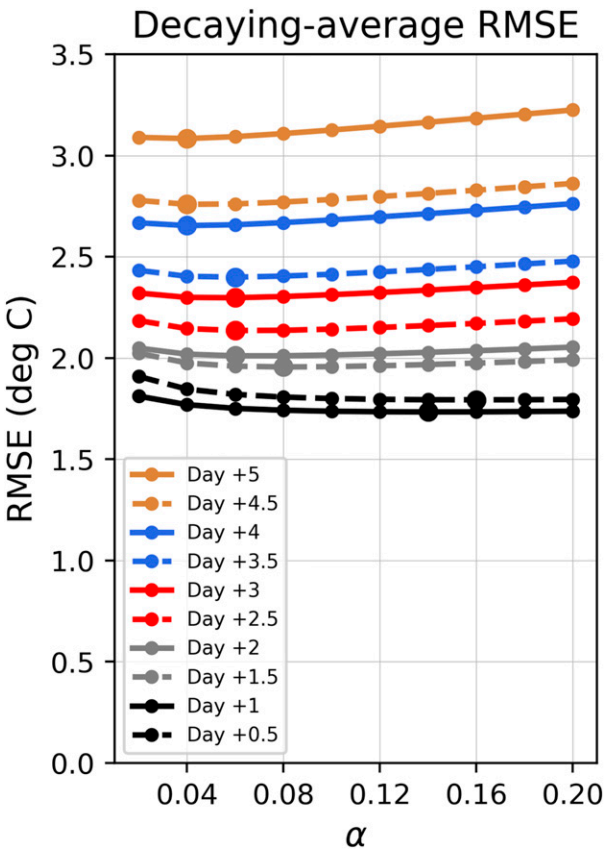

FIG. 2. Root-mean-square error of the decaying-average bias correction method as a function of α for various forecast lead times. The larger dot denotes the value with the lowest error in the training period.

e. *Univariate MOS*

Univariate MOS, or uMOS hereafter, is an application of simple linear regression to bias correction. This assumes that an estimate of the analyzed temperature may be determined through a regression equation of the following form:

$$\hat{a}_t^{\mathrm{uMOS}} = c_0 + c_1 f_t, \tag{5}$$

where c_0 and c_1 are the fitted intercept and slope. The error (or residual) $e_t = a_t - \hat{a}_t$ is commonly assumed to be normally distributed with zero mean. In practice, nonlinear relationships

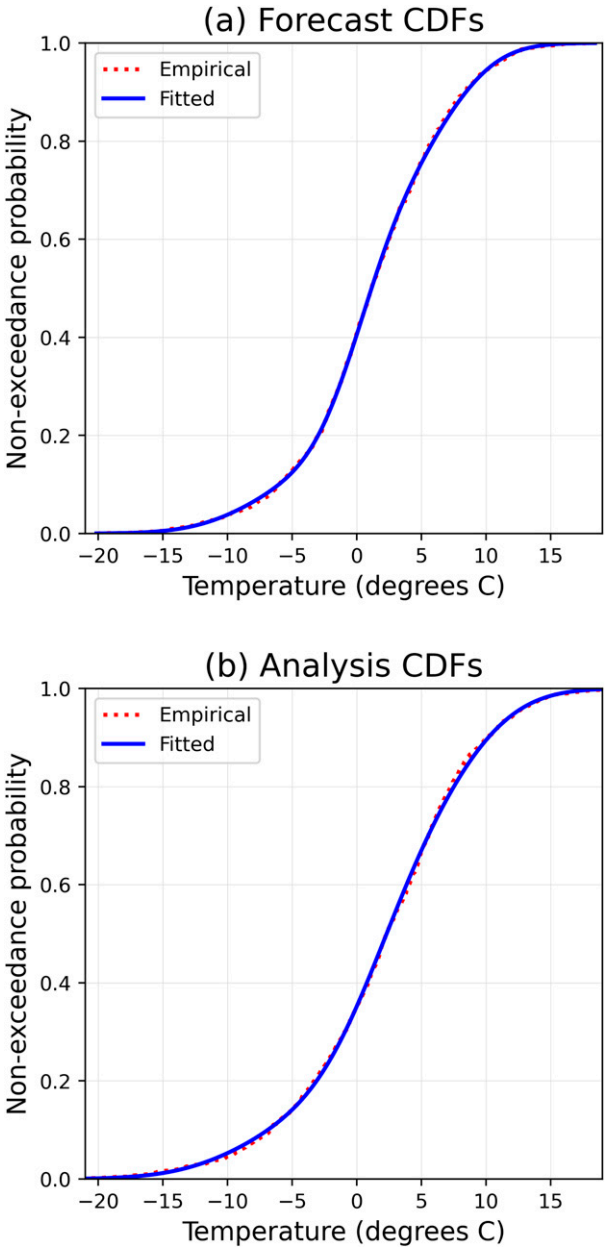

FIG. 3. Examples of empirical (dashed red, underneath) and fitted (blue, overtop) CDFs estimated with a three-component Gaussian mixture, here for January data at +24-h lead time near Boulder, CO (40°N, 105°W): (a) 2-m temperature forecast data and (b) corresponding analysis data.

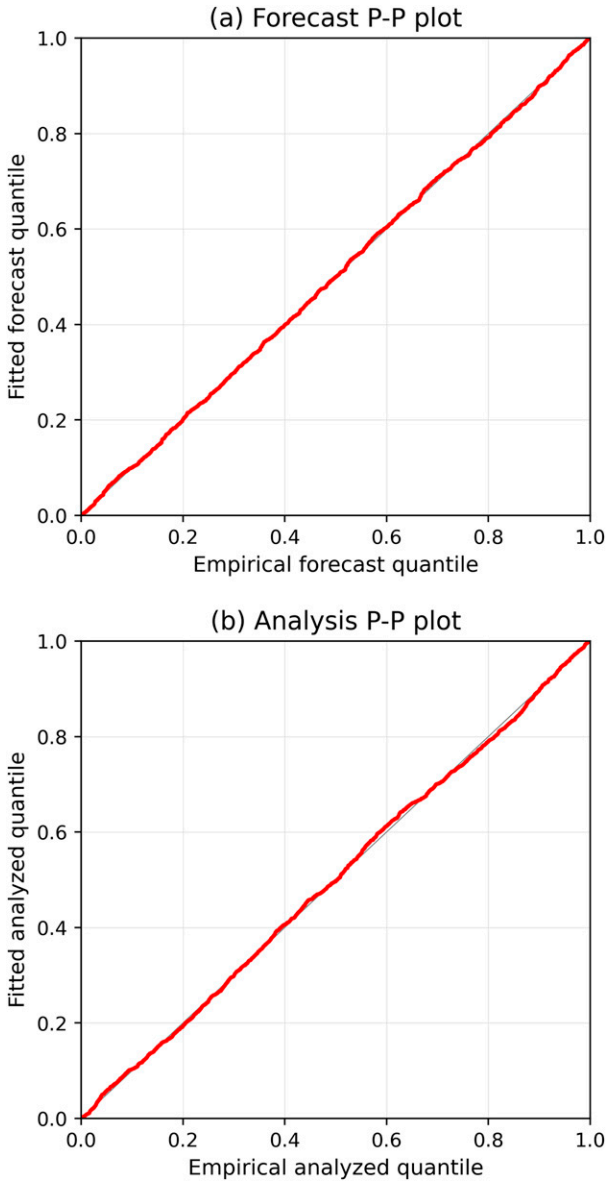

FIG. 4. Probability–probability (P–P) plots corresponding to the data in Fig. 3.

and heteroscedasticity were present at many grid points, as will be discussed in the results, but for generality, no gridpoint specific remedial measures such as power transformation of data were employed. Linear regression is reviewed in many texts, including Wilks (2011, his section 7.2.1). As with the QM, separate regression equations were fit for each grid point at each forecast lead time, month by month, using 2000–18 training data and a 3-month period centered on the month of interest. The implicit bias estimate was thus $\hat{b}_t^{\text{uMOS}} = f_t - \hat{a}_t^{\text{uMOS}}$.

f. *Multivariate MOS*

Multivariate MOS (mvMOS) using multiple forecast variables as predictors has a long heritage in the U.S. National Weather Service (Glahn and Lowry 1972; Carter et al. 1989) and in many other forecast agencies. Commonly, multiple forecast fields including variables above the surface are used as additional predictors, variables such as forecast cloud cover, thicknesses between constant pressure levels, and so forth. Application of this approach with the data at hand could be more challenging, for a screening regression approach to the selection of predictors might result in very different predictor choices for different parts of the domain, complicating interpretation. Rather than use this approach

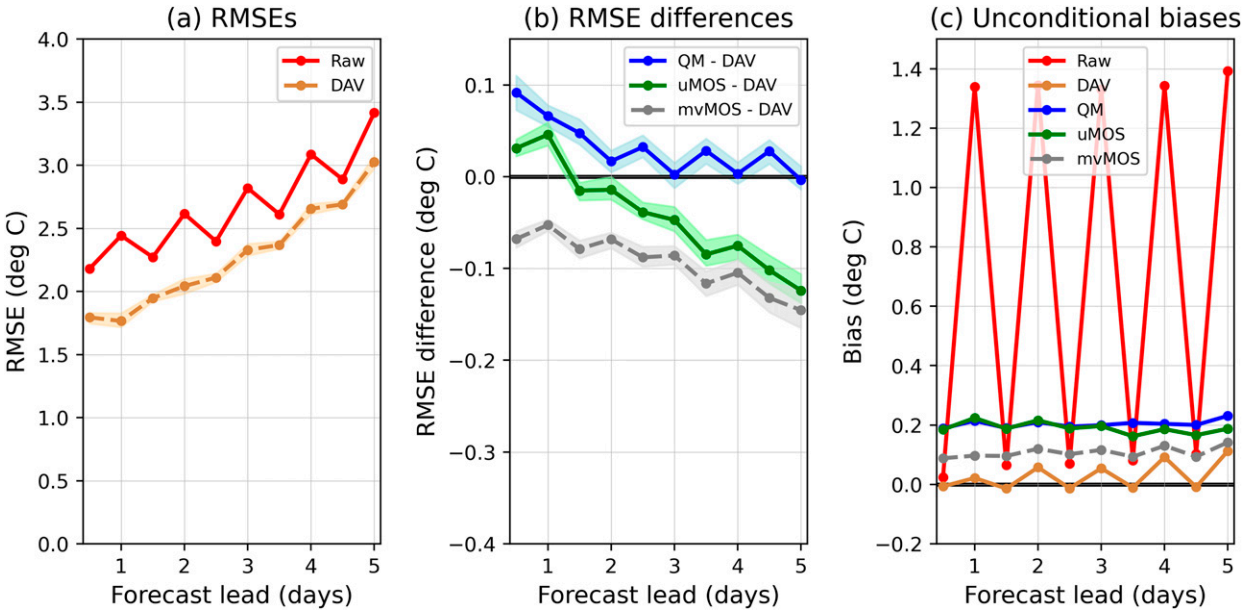

FIG. 5. Domain-averaged errors of raw forecasts and various bias-correction methods. (a) RMSE of raw (solid red) and DAV (dashed orange-brown) bias correction methods as a function of forecast lead time. The 5th and 95th percentile confidence interval of differences between the two forecasts are plotted as light orange-brown around the DAV method. (b) RMSE differences of the QM − DAV method (blue; lower is an improvement over DAV), uMOS (green), and mvMOS (gray). Confidence intervals are plotted in lighter-shade colors as in (a), but here the confidence intervals represent differences with respect to the DAV method. (c) Unconditional bias for raw forecasts and the various bias correction methods.

with its training and data management complexity, multivariate here implies something slightly different; instead of multiple forecast variables, the bias corrections from other approaches are used as predictors. Specifically, we estimate the analyzed state with a regression equation of the following form:

$$\hat{a}_t^{\mathrm{mvMOS}} = c_0 + c_1 f_t + c_2 \hat{b}_t^{\mathrm{DAV}} + c_3 \hat{b}_t^{\mathrm{QM}}. \tag{6}$$

This allows us to determine whether a method that uses information from alternative bias-correction approaches may improve the forecasts. No interaction terms were included. The implicit bias correction is $\hat{b}_t^{\mathrm{mvMOS}} = f_t - \hat{a}_t^{\mathrm{mvMOS}}$. The training data periods were the same as with uMOS, but a first sweep through the data was necessary to generate the DAV and QM bias estimates, which were then used as predictor data in the mvMOS regression analysis.

Previously, multimethod synthesis showed promise for probabilistic forecasts (Möller and Groß 2016; Bassetti et al. 2018; Yang et al. 2017; Baran and Lerch 2018).

g. *Verification methods*

Commonly used verification methods will be applied, focusing on error and bias. Root-mean square error (RMSE) statistics will be provided. For a bias-correction method with n samples across the grid and n_t times, the estimated RMSE for a given postprocessing method is

$$\widehat{\mathrm{RMSE}} = \left\{ \frac{1}{n_t n - 1} \sum_{t=1}^{n_t} \sum_{i=1}^{n} [\hat{a}_t(i) - a_t(i)]^2 \right\}^{1/2} \tag{7}$$

and the estimated unconditional mean bias (BIA) is

$$\widehat{\mathrm{BIA}} = \frac{1}{n_t n} \sum_{t=1}^{n_t} \sum_{i=1}^{n} [\hat{a}_t(i) - a_t(i)]. \tag{8}$$

Differences of RMSE for the various bias-correction methods were evaluated relative to the DAV method. The 5th and 95th percentile confidence intervals for these RMSE differences were generated with the paired block-bootstrap procedure described in Hamill (1999); 100 resamplings were performed.

Taylor diagrams were also generated as a way of understanding the forecast characteristics (Taylor 2001; Wilks 2011, their section 8.6.3). These diagrams were plotted in polar coordinates. The radial distance from the origin represented the ratio of the climatological standard deviations of forecast versus analyses. This was the mean forecast variability divided by the mean analyzed variability, where variability measured the standard deviation of the sample. The angle, computed clockwise from the 12 o'clock position, represented the forecast versus observed correlation. Gray lines denote lines of equal standardized RMSE. For this application of Taylor diagrams, a sample will be plotted for each forecast grid point, so that the potential variability of the error decomposition across the domain can be examined.

3. Results

Figure 5 provides RMSE and BIA statistics averaged over all land points within the domain. The DAV method provided a statistically significant decrease in RMSE relative to the raw guidance, and this reduction in error amounted to

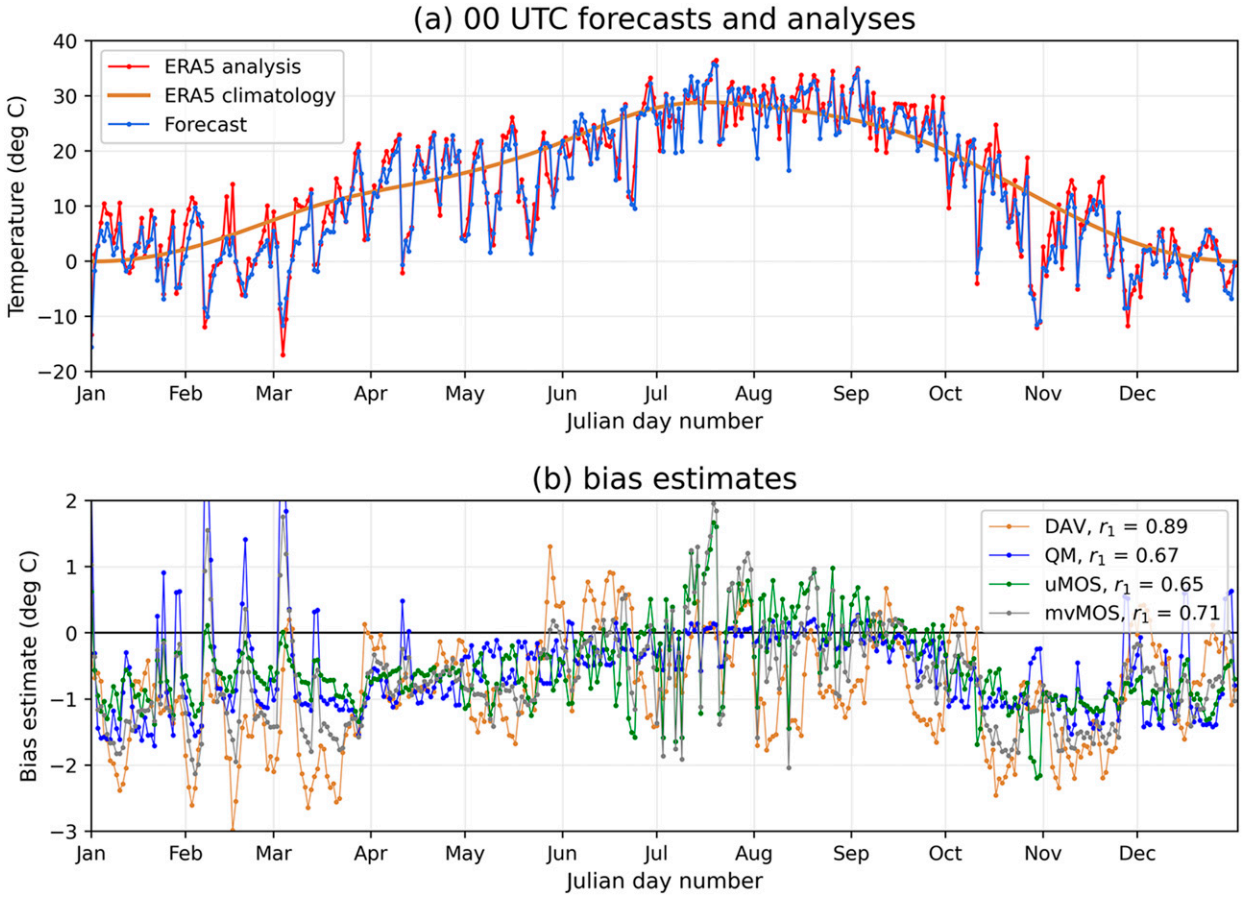

FIG. 6. The +24-h forecast and ERA5 analyzed time series of 0000 UTC data for a grid point near Boulder, CO, during 2019. (a) ERA5 analyses (red) and GEFSv12 forecasts (blue). (b) Bias estimates from various methods. One-day lag autocorrelations are provided in the inset legend.

1–3-day gain lead time; for example, the +3-day DAV forecasts were nearly as skillful as the +1-day raw forecasts. Figure 5b shows that QM generally produced forecasts with slightly higher RMSE than DAV, but the uMOS and mvMOS forecasts after +1.5 days provided a significant reduction in RMSE relative to DAV. Overall, the mvMOS forecasts produced the lowest domain-averaged RMSE, and this was generally true across seasons (not shown).

Domain-averaged unconditional biases were reduced in all methods (Fig. 5c), but the DAV method produced forecasts with the lowest unconditional bias. Apparently, the bias characteristics of 2000–18 training period were dissimilar to those of 2019, for QM, uMOS, and mvMOS all demonstrated slight warm biases. Despite their higher bias, both the uMOS and mvMOS algorithms were designed to minimize squared error with the training sample provided, which is likely why they provide slightly lower RMSE in the validation period than DAV at most leads, despite their larger biases. Alternative formulations of uMOS and mvMOS were also tested, wherein the forecast data were changed to be a deviation from a daily unconditional climatological mean analyzed temperature, which was estimated with a cubic-spline spline fit (not shown). These reduced the RMSE of these methods very slightly. However, to facilitate more direct comparison against the other methods, only the results using the unmodified forecasts are presented.

The reduced error of mvMOS is an interesting result, supported by other literature (Möller and Groß 2016; Yang et al. 2017; Bassetti et al. 2018; Baran and Lerch 2018). Different postprocessing methods have different strengths. To the extent that biases were unconditional on the forecast temperature and can be reasonably estimated with recent samples, then DAV performs well. The QM method does not attempt to minimize error but seeks the analyzed value associated with today's quantile in the forecast distribution, producing samples that should be draws from the analyzed distribution. Large mappings will occur if the cumulative probabilities differ between forecast and analyzed. The MOS methods by design minimize RMSE, but as will be discussed later, at the expense of other forecast characteristics.

How responsive were the various statistical adjustment techniques to day-to-day changes of weather? As an example, time series of +24-h forecast data for a grid point near Boulder, Colorado, are presented in Fig. 6. The top panel displays a time series of the +24-h lead GEFSv12 forecast and ERA5

analyzed data, as well as the ERA5 climatology, fitted with cubic splines as in Hamill and Scheuerer (2020). The bottom panel shows corresponding time series of the various post-processing methods. The 1-day lag autocorrelation coefficients are provided in the inset legend. Per its design, the DAV method changed the least from day to day, with the largest autocorrelations; it exhibits an inertia. The other methods' statistical adjustments were more responsive to the weather.

Let us consider more closely the rapid oscillations of the regression methods during July in Fig. 6b. Figure 7a shows the CDFs used in the quantile-mapping function at this location. The CDFs aligned closely with each other, so the mappings were quite modest, and only modest changes in bias estimates occurred from day to day during this month; however, in other months the QM corrections were more sensitive to the forecast temperature, such as in February. In contrast, the regression methods produced differing corrections for different forecast temperatures at this grid point during July. Figure 7b shows the 2000–18 training data for this location as well as the fitted uMOS regression curve. In this case, the one-size-fits-all regression approach, with no remedial measures to address issues such as heteroscedasticity, appeared to be a model shortcoming. The training data were in fact heteroscedastic, with larger differences between forecasts and observations at lower temperatures, suggesting uMOS accuracy could be further improved with remedial measures to alleviate heteroscedasticity. Further, the marginal distributions showed that the underlying data were multimodal in nature, with peak probability density at the higher temperatures; because of the larger number of samples with higher temperatures, the regression fit was more closely optimized to these samples. As a consequence, the regression model did not appear to provide a high-quality fit at the lower temperatures; in this instance when the forecast temperatures were comparatively low, the regression model predicted a cold forecast bias. The actual (forecast, analyzed) samples for July 2019 were presented in Fig. 7b as the bolder red points, several of which have colder forecast temperatures and predicted cold biases based on the regression line. With a daily change in forecast temperature from warm to cold, there was a corresponding change in the estimated forecast bias from too warm to too cold, and hence large oscillations occurred with the change in forecast temperatures from one day to the next.

Despite its challenges exhibited in Fig. 7, the uMOS and mvMOS methods did produce comparatively lower RMSE on average, but what other forecast characteristics did they have relative to the other methods? This can be examined in part with Taylor diagrams (Taylor 2001; Wilks 2011, their section 8.6.3). Figure 8 provides such diagrams for the +24-h forecasts for raw, DAV, QM, and mvMOS methods during the July–August–September 2019 period (other lead times and seasons had qualitatively similar results, not shown). See the references above for more interpretation of these diagrams. Differently colored dots denote the magnitude of the analysis standard deviation, i.e., the red dots denoted locations with little weather variability during the sample period while the brown dots were locations with the most weather variability.

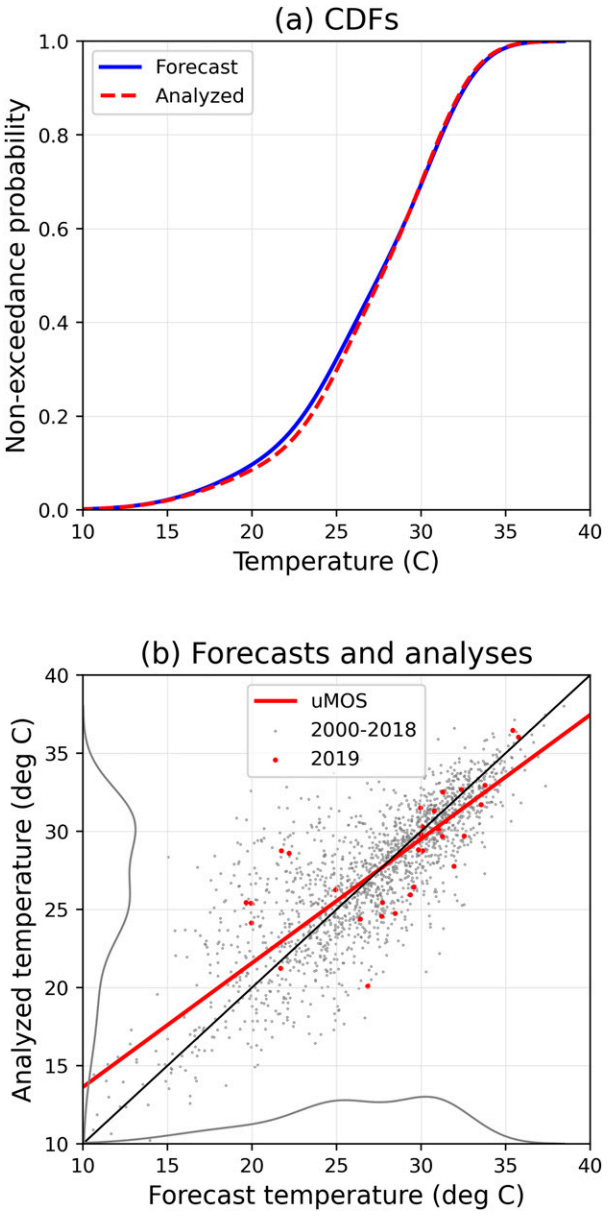

FIG. 7. (a) Fitted cumulative distribution functions (CDFs) used in the quantile-mapping procedure for +24-h lead forecasts at a grid point near Boulder, CO, during July. (b) Scatterplot of +24-h analyzed vs forecast analyzed 2000–18 training data for July at Boulder, CO (small gray dots), and marginal probability density functions (gray lines along each axis). The uMOS fitted linear regression line is presented in red, and the 2019 (forecast, analyzed) pairs are shown as the larger red dots.

The raw forecasts exhibited much scatter in the Taylor diagram standard deviation ratio, sometimes with the forecast sample during this season having more variability than the analyzed data, and sometimes less. These variations in the standard deviation ratio were muted only somewhat with the DAV method. The QM method, consistent with its goal of producing mappings that represented draws from the analyzed

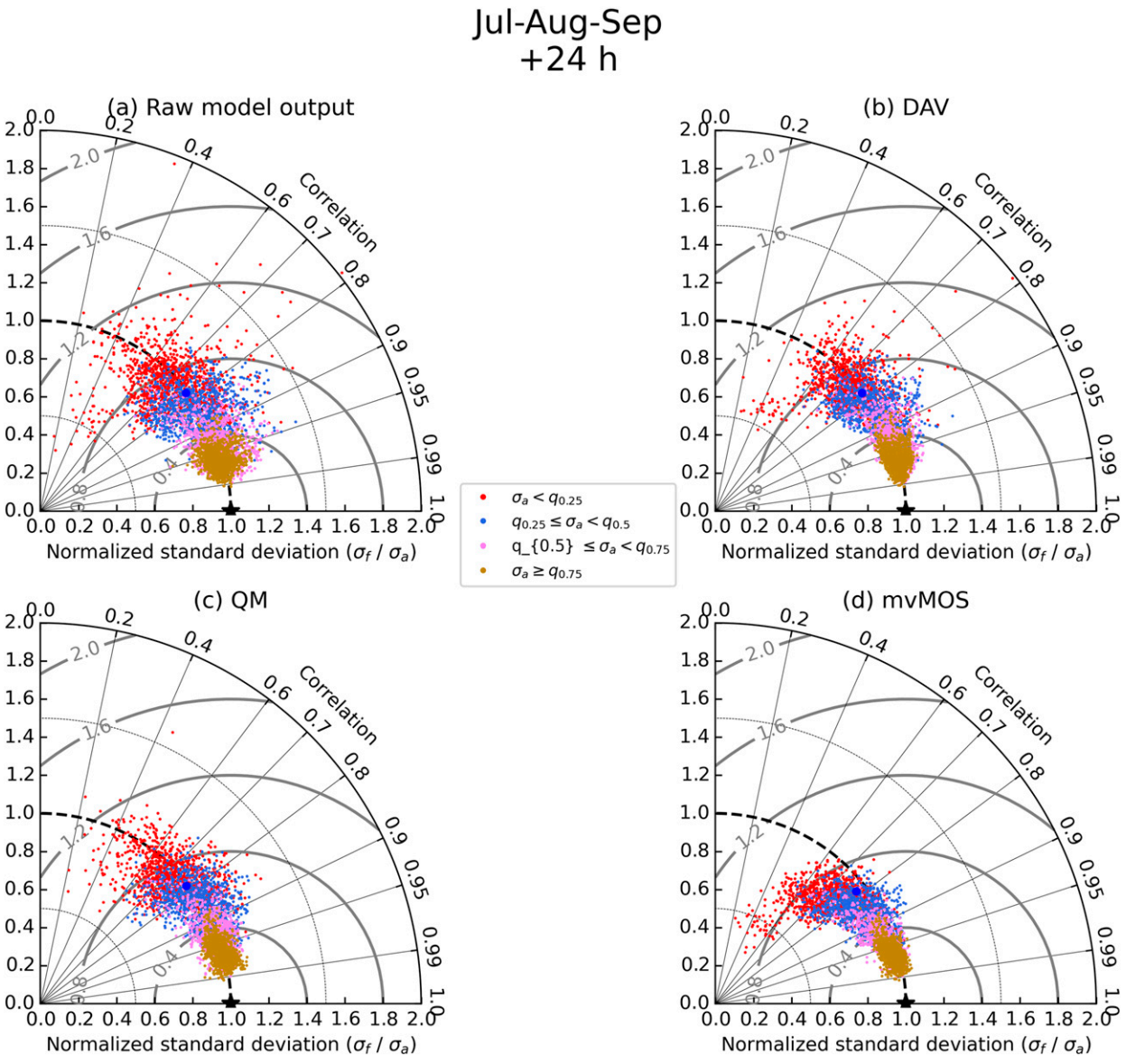

FIG. 8. Taylor diagrams for July–September and +24-h lead time for (a) raw forecasts, (b) DAV, (c) QM, and (d) mvMOS. A sample from each land grid point is plotted as a separate dot. The radial magnitude indicates the ratio of the sample forecast standard deviation during the season divided by the sample analysis standard deviation. Correlation increases clockwise from the 12 o'clock position (0.0) to the 3 o'clock position (1.0). Gray lines denote lines of equal standardized RMSE. Individual dots are colored by that grid point's sample analysis standard deviation σ_a. The dots' color legend is provided in the plot center; red dots are samples where the analysis standard deviation was in the lowest quartile of sorted samples, blue dots are for the second-lowest quartile, pink dots are for the second-highest quartile, and brown are for the highest quartile.

climatology, had a narrower range of standard deviation ratios that were more concentrated around the 1.0 ratio. The practical effect of this as a forecast procedure is that *this method retains more of the variability in the observations*, as it is designed to do. In contrast, the mvMOS procedure, especially for the forecasts at locations with smaller analyzed weather variability, produced less variability in the corrected forecasts than in the analyzed, as denoted by the ratio that on average was lower than 1.0. It is possible that a human forecaster, say, seeking to predict the magnitude of a warm or cold event, might prefer the QM guidance relative to one of the MOS procedures' guidance, given that the former retained more of the synoptic-scale variability.

4. Conclusions

This study provided an intercomparison of statistical post-processing methods applied to deterministic surface-temperature forecasts on a ½° grid over the CONUS and surrounding land regions out to +5-day lead time. The control member from the Global Ensemble Forecast System version 12 reforecast data

were used to provide the forecasts, 2000–18 for training and 2019 for validation. ECMWF reanalyses, specifically ERA5, were used for training and validation. The four methods that were considered were the decaying-average bias correction (DAV), quantile mapping (QM), a univariate model output statistics (uMOS), otherwise known as linear regression, and a multivariate MOS (mvMOS) that used the surface-temperature forecasts as well as bias estimates from DAV and QM as predictors. Except at the earliest leads, the MOS techniques produced forecasts with the lowest error with mvMOS providing errors lower than uMOS. Through an examination of Taylor diagrams, it was revealed that while the mvMOS reduced the error, especially at locations with low climatological variability across a season, it also reduced the variability in the postprocessed forecasts relative to the raw guidance. On the other hand, QM and DAV methods retained much of the seasonal variability in the raw forecasts. Which method a forecaster may prefer could depend on whether they are optimizing for RMSE (choose a MOS method) or for more realistic prediction of the magnitude of unusual events (choose DAV or QM). The DAV method produced bias corrections that were more consistent in time, while the QM and MOS techniques were more sensitive to the weather of the day.

A main conclusion is that because different postprocessing methods may have differing strengths and weaknesses, the judicious combination of them may be able to, in some metrics, provide guidance that is improved relative to any one on its own (this is similar to what has been found with multimodel ensemble combination). In particular, the mvMOS method here, which combined DAV, QM, and MOS approaches, produced guidance with the lowest RMSE. Since each postprocessing method is relatively straightforward to implement, an operational combination of these could be a practical solution that would provide modestly improved guidance for many customers. This is qualitatively similar to the reduction in error that many have previously found in multimodel ensemble combinations.

This study was not comprehensive; it considered only an area around the United States, and it used a long training dataset and considered only surface temperature, not other variables of interest such as winds or cloud cover or precipitation. Nonetheless, the optimistic results, confirmed by other supporting literature, suggest that the judicious combination of multiple, simple postprocessing methods may provide a practical way to reduce errors.

Acknowledgments. ERA5 reanalysis data from the Copernicus Climate Service were used in this study. Publication costs were provided by the NOAA Weather Program Office grant U8R2WRE-P00. Python library scikit-learn.mixture was used for distribution fitting.

REFERENCES

Baran, S., and S. Lerch, 2018: Combining predictive distributions for the statistical post-processing of ensemble forecasts. *Int. J. Forecasting*, **34**, 477–496, https://doi.org/10.1016/j.ijforecast.2018.01.005.

Bassetti, F., R. Casarin, and F. Ravazzolo, 2018: Bayesian nonparametric calibration and combination of predictive distributions. *J. Amer. Stat. Assoc.*, **113**, 675–685, https://doi.org/10.1080/01621459.2016.1273117.

Bougeault, P., and Coauthors, 2010: The THORPEX Interactive Grand Global Ensemble. *Bull. Amer. Meteor. Soc.*, **91**, 1059–1072, https://doi.org/10.1175/2010BAMS2853.1.

Carter, G. M., J. P. Dallavalle, and H. R. Glahn, 1989: Statistical forecasts based on the National Meteorological Center's numerical weather prediction system. *Wea. Forecasting*, **4**, 401–412, https://doi.org/10.1175/1520-0434(1989)004<0401:SFBOTN>2.0.CO;2.

Craven, J. P., D. E. Rudack, and P. E. Shafer, 2020: National Blend of Models: A statistically post-processed multi-model ensemble. *J. Oper. Meteor.*, **8**, 1–14, https://doi.org/10.15191/nwajom.2020.0801.

Cui, B., Z. Toth, Y. Zhu, and D. Hou, 2012: Bias correction for global ensemble forecast. *Wea. Forecasting*, **27**, 396–410, https://doi.org/10.1175/WAF-D-11-00011.1.

Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The gridding of MOS. *Wea. Forecasting*, **24**, 520–529, https://doi.org/10.1175/2008WAF2007080.1.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

——, 2012: Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Wea. Rev.*, **140**, 2232–2252, https://doi.org/10.1175/MWR-D-11-00220.1.

——, 2018: Practical aspects of statistical postprocessing. *Statistical Postprocessing of Ensemble Forecasts*, S. Vannitsem, D. Wilks, and J. Messner, Eds., Elsevier Press, 187–217.

——, and M. Scheuerer, 2018: Probabilistic precipitation forecast postprocessing using quantile mapping and rank-weighted best-member dressing. *Mon. Wea. Rev.*, **146**, 4079–4098, https://doi.org/10.1175/MWR-D-18-0147.1.

——, and ——, 2020: Improving ensemble weather prediction system initialization: Disentangling the contributions from model systematic errors and initial perturbation size. *Mon. Wea. Rev.*, **149**, 77–90, https://doi.org/10.1175/MWR-D-20-0119.1.

——, E. Engle, D. Myrick, M. Peroutka, C. Finan, and M. Scheuerer, 2017: The U.S. National Blend of Models for statistical postprocessing of probability of precipitation and deterministic precipitation amount. *Mon. Wea. Rev.*, **145**, 3441–3463, https://doi.org/10.1175/MWR-D-16-0331.1.

Hersbach, H., and Coauthors, 2020: The ERA5 global reanalysis. *Quart. J. Roy. Meteor. Soc.*, **146**, 1999–2049, https://doi.org/10.1002/qj.3803.

Hopson, T. M., and P. J. Webster, 2010: A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *J. Hydrometeor.*, **11**, 618–641, https://doi.org/10.1175/2009JHM1006.1.

Krishnamurti, T. N., C. M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendran, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate*, **13**, 4196–4216, https://doi.org/10.1175/1520-0442(2000)013<4196:MEFFWA>2.0.CO;2.

Liu, J., and Z. Xie, 2014: BMA probabilistic quantitative precipitation forecasting over the Huaihe basin using TIGGE multimodel ensemble forecasts. *Mon. Wea. Rev.*, **142**, 1542–1555, https://doi.org/10.1175/MWR-D-13-00031.1.

Lowry, D. A., and H. R. Glahn, 1976: An operational model for forecasting probability of precipitation—PEATMOS PoP.

Mon. Wea. Rev., **104**, 221–232, https://doi.org/10.1175/1520-0493(1976)104<0221:AOMFFP>2.0.CO;2.

Maraun, D., 2013: Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue. *J. Climate*, **26**, 2137–2143, https://doi.org/10.1175/JCLI-D-12-00821.1.

Möller, A., and J. Groß, 2016: Probabilistic temperature forecasting based on an ensemble AR modification. *Quart. J. Roy. Meteor. Soc.*, **142**, 1385–1394, https://doi.org/10.1002/qj.2741.

Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Taylor, K. E., 2001: Summarizing multiple aspects of forecast performance in a single diagram. *J. Geophys. Res.*, **106**, 7183–7192, https://doi.org/10.1029/2000JD900719.

Vannitsem, S., D. S. Wilks, and J. W. Messner, Eds., 2018: Statistical Postprocessing of Ensemble Forecasts. Elsevier Press, 347 pp.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output and statistics through model consensus. *Bull. Amer. Meteor. Soc.*, **76**, 1157–1164, https://doi.org/10.1175/1520-0477(1995)076<1157:IMOSFT>2.0.CO;2.

Voisin, N., J. C. Schaake, and D. P. Lettenmaier, 2010: Calibration and downscaling methods for quantitative ensemble precipitation forecasts. *Wea. Forecasting*, **25**, 1603–1627, https://doi.org/10.1175/2010WAF2222367.1.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. International Geophysics Series, Vol. 100, Academic Press, 704 pp.

Yamaguchi, M., T. Nakazawa, and S. Hoshino, 2012: On the relative benefits of a multi-centre grand ensemble for tropical cyclone track prediction in the western North Pacific. *Quart. J. Roy. Meteor. Soc.*, **138**, 2019–2029, https://doi.org/10.1002/qj.1937.

Yang, X., S. Sharma, R. Siddique, S. J. Greybush, and A. Mejia, 2017: Postprocessing of GEFS precipitation ensemble reforecasts over the U.S. Mid-Atlantic region. *Mon. Wea. Rev.*, **145**, 1641–1658, https://doi.org/10.1175/MWR-D-16-0251.1.