Sensitivity of Calibrated Week-2 Probabilistic Forecast Skill to Reforecast Sampling of the NCEP Global Ensemble Forecast System

MELISSA H. OU, MIKE CHARLES, AND DAN C. COLLINS

National Centers for Environmental Prediction/Climate Prediction Center, College Park, Maryland

(Manuscript received 3 December 2015, in final form 11 March 2016)

ABSTRACT

CPC requires the reforecast-calibrated Global Ensemble Forecast System (GEFS) to support the production of their official 6–10- and 8–14-day temperature and precipitation forecasts. While a large sample size of forecast–observation pairs is desirable to generate the necessary model climatology and variances, and covariances to observations, sampling by reforecasts could be done to use available computing resources most efficiently. A series of experiments was done to assess the impact on calibrated forecast skill of using a smaller sample size than the current available reforecast dataset. This study focuses on the skill of week-2 probabilistic forecasts of the 7-day-mean 2-m temperature and accumulated precipitation. The tercile forecasts are expressed as being below-, near-, and above-normal temperature/median precipitation over the continental United States (CONUS). Calibration statistics were calculated using an ensemble regression technique from 25 yr of daily, 11-member GEFS reforecasts for 1986–2010, which were then used to postprocess the GEFS model forecasts for 2011–13. In assessing the skill of calibrated model output using a reforecast dataset with fewer years and ensemble members, and an ensemble run less frequently than daily, it was determined that reductions in the number of ensemble members to six or fewer and reductions in the frequency of reforecast runs from daily to once a week were achievable with minimal loss of skill. However, reducing the number of years of reforecasts to less than 25 resulted in a greater skill degradation. The loss of skill was statistically significant using only 18 yr of reforecasts from 1993 to 2010 to generate model statistics.

1. Introduction

It is well known that it is necessary to postprocess direct model output (DMO) from numerical weather prediction (NWP) models to improve forecast skill as a result of inherent model biases (Wilson et al. 2007; Hamill et al. 2004). Ensembles are often not well calibrated (Wilson et al. 2007), tending to be underdispersive (Hamill and Colucci 1998; Raftery et al. 2005), producing probabilistic forecasts with unreliable probabilities that are often overconfident (Hamill et al. 2004; Whitaker et al. 2006). Because raw ensemble members less accurately represent the uncertainty of the model with increasing lead time because of error growth attributable to chaos and model errors (Lorenz 1969), postprocessing is especially important at longer lead times such as medium-range forecasts and beyond week 1 (Hamill et al. 2004; Hagedorn et al. 2008; Cui et al. 2012). Uncalibrated ensemble forecasts (e.g., forecast temperatures) often produce probability forecasts with values that are consistently too high (Johnson and Swinbank 2009). Week-2 probabilistic temperature and precipitation forecasts derived from the raw, uncalibrated National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) have shown negative ranked probability skill scores (RPSSs) and very poor statistical reliability (Hamill et al. 2004, Whitaker et al. 2006).

Reforecasts (retrospective forecasts produced by a frozen version of an NWP model) have been commonly used to calibrate DMO and have been shown to significantly improve forecasts on various time scales (Hamill et al. 2004, 2008; Wilks and Hamill 2007; Hagedorn et al. 2012) by filtering the predictable signal from the unpredictable noise (Hamill et al. 2004). Hagedorn et al. (2008) found noticeable improvements in skill for the ECMWF and GEFS forecasts at day 4 and beyond, with the GEFS model benefitting the most from reforecast calibration. Reforecasts are also important to placing

Corresponding author address: Melissa H. Ou, NOAA/Center for Weather and Climate Prediction, Climate Prediction Center, 5830 University Research Ct., College Park, MD 20740.
E-mail: melissa.ou@noaa.gov

DOI: 10.1175/WAF-D-15-0166.1

real-time forecasts into the context of a model's historical forecasts. It is especially important to assess how rare or common a weather or climate event is within the context of the model climatology (Hamill et al. 2013, Lalaurette 2003).

A sufficiently sized reforecast dataset is required to obtain a robust sample of past forecast errors. Enough weather and climate scenarios must be captured by the hindcasts to reflect the range of possible outcomes in order to properly calibrate the real-time DMO, especially for nonnormal fields such as precipitation and for rare events (Hagedorn et al. 2008; Hamill et al. 2008; Hamill et al. 2013). Previous studies have shown that using longer-term reforecasts significantly improves forecast skill and issues related to underdispersive ensembles compared to using short, recent periods of forecasts for postprocessing DMO (Cui et al. 2012; Hagedorn et al. 2012).

The main goal of this study is to determine how subsampling reforecasts based on three parameters (the number of years, number of ensemble members, and frequency of model runs) impacts the skill of postprocessed week-2 forecasts over the continental United States (CONUS). This work provides an updated analysis of the sensitivity of medium-range forecast skill for temperature and precipitation using the 2012 GEFS model, which was needed since a previous similar study evaluated a much older GEFS version using 1998 model physics (Hamill et al. 2004). Other previous papers proposed evaluating whether a large training dataset would still reap large benefits with newer, high-resolution models with potentially reduced systematic errors (Hamill et al. 2004; Hagedorn et al. 2008). Additionally, many previous studies looked at the impact of training years and model run frequency of reforecasts on forecast skill, but few, if any, examine the sensitivity to various configurations of ensemble size (especially for the GEFS) beyond using just the control run versus a full set of members.

ESRL created previous reforecast datasets (including the one used in this study), but these datasets will now be produced at the NCEP operationally. To take advantage of the reforecast dataset, it must be produced by the same version as the real-time model, with the same model initialization and data assimilation methods, physics scheme, etc. (Hamill and Whitaker 2006; Hagedorn et al. 2012). When NCEP releases an upgraded version of the GEFS, a complementary reforecast dataset is needed to calibrate forecasts. Since the production of reforecasts requires considerable computational and human resources and would likely require resources shared with that of the operational real-time GEFS, it is desirable for NCEP to be able to produce the GEFS at a cheaper reforecast configuration without greatly reducing the skill of operational forecasts. This study was largely motivated by the need to find an optimal subset of reforecasts that decreases the resources needed to produce them but retains significant skill for forecasts produced at the weekly time scale. Some results and the main recommendations from this study were included in a collaborative white paper by Hamill et al. (2014), with sensitivity results from a number of national centers as well as a proposed configuration. The results from this white paper helped determine the configurations for NCEP's operational reforecast production.

Each day, CPC issues official daily 6–10- and 8–14-day probabilistic tercile forecasts of the mean near-surface temperature at 2 m (hereafter referred to as surface temperature) and accumulated precipitation over the CONUS and Alaska. The statistical–dynamical tool being examined in this study, referred to as the reforecast tool, is one of the most used guidance tools by CPC's forecasters. This tool uses the current GEFS reforecasts to perform calibration on real-time GEFS data to produce 6–10-day and week-2 forecasts of temperature and precipitation expressed as probabilistic terciles, similar to the official forecasts. A similar postprocessing method can be applied to any dynamical model with reforecasts. CPC is also using the reforecast tool to produce week-2 probabilistic forecasts of extremes, which serves as the main guidance for a new, experimental, probabilistic week-2 hazards forecast.

Below we first describe the datasets used in this study, as well as the framework of the sample size experiments, and an overview of the statistical techniques used for calibration and verification (section 2). Then, we present the results of the skill sensitivity tests, including seasonal variations in skill (section 3). Finally, we discuss our conclusions and recommendations regarding an optimal subset of reforecasts for calibrating week-2 temperature and precipitation forecasts (section 4).

2. Datasets and methods

Below, we describe the reforecast and verification datasets, explain the framework of the sample size experiments, and present the postprocessing and verification methodology used in this work.

a. Training data and forecast datasets

The reforecast dataset serves as both the training dataset as well as the source of forecasts being calibrated in this study. This dataset consists of 25 yr of once-daily, 11-member ensemble reforecasts from the legacy ''frozen'' 2012 version 10 of the NCEP GEFS model (Hamill et al. 2013). The resolution of the model

is T254L42 (~40 km at 40° latitude) during the first 8 days and T190L42 (~54 km at 40° latitude) from days 8 to 16. We use 25 yr of reforecasts to train the reforecast tool (1 December 1985–30 November 2010) and evaluate calibrated GEFS forecasts (valid dates centered at 1 January 2011–31 December 2013) for a 3-yr period following the training period. Therefore, the forecasts being verified are independent from the training data.

This study assesses week-2 (days 8–14) forecasts of 7-day mean 2-m temperature and accumulated precipitation. The tercile forecasts are expressed as three categories: below-, near-, and above-normal temperature/median precipitation, similar to CPC's official 8–14-day outlooks. The three categories are defined as being below the 33rd percentile, between the 33rd and 67th percentiles, and above the 67th percentile of the climatological distribution of observed temperature and precipitation, respectively. The tercile thresholds (33rd and 67th percentiles) were obtained using the climatology derived from the entire reforecast dataset (1986–2010). Our domain of interest is over the CONUS.

To calculate the calibration statistics, the daily analysis fields of the GEFS reforecast control run (considered the "day zero" of the model runs) were used as a proxy for the paired "observations" to the past forecasts to train the reforecast tool. The analyses were obtained by averaging four update cycles of the model (0000, 0600, 1200, and 1800 UTC) daily, where 0000 UTC is the initialization field from the Global Data Assimilation System (GDAS) reanalysis. Daily observations were converted to 7-day means, matching the format of forecasts used to train the reforecast tool. We elected to use the GEFS analyses as the observations for deriving the statistics for a few reasons. First, our group has experienced that there can be nonnegligible discrepancies in the forecast and observation grids associated with topographical features and sparse observation networks, and that by using a matching model grid analysis, the fields are more comparable. Compatible forecast and training observation grids produce smoother, more physically realistic calibrated forecasts (Fundel et al. 2010). Second, using currently available observation datasets to train the model would have suggested that we have a high degree of trust in the accuracy of the dataset, which is not necessarily the case. We confirmed this by comparing the skill of the reforecast tool using both observation- and analysis-based training data, and the difference was insignificant (results not shown).

The climatological mean of observations (used for deriving observation anomalies from the reforecast data) for each verification date was determined by pooling the observation data for 31 dates centered on each day of the year, and triangularly weighting dates such that the center date is given the greatest weight and weights decrease to zero at 16 days earlier and later.

It should be noted that the overall skill of the reforecast tool presented in this study is likely degraded (relative to the potential skill of a reforecast-based tool) as a result of a known land surface error in the 2012 GEFS model (Hamill et al. 2013), as well as changes in the datasets used for model initialization, contributing to significant model biases. Errors in land surface types and the initialization of soil conditions likely affect the response of the model to specific conditions and climate modes of variability, such as ENSO. However, these potential model biases are present in all subsets of the reforecast data used and should not impact the conclusions of this study on the relative skill of reforecast subsets.

b. *Verification datasets*

Week-2 mean surface temperature and precipitation forecasts and verifying observations used in this study are over the CONUS domain, including 205 stations for temperature and 100 for precipitation (Fig. 1) (there are fewer stations for precipitation as a result of having dry station statistics for only a limited number of stations readily available; for the details of the methodology, see the appendix). The gridded forecast data were interpolated to station data to match the format of the observational dataset. This was done by using the nearest-neighbor method, where the closest gridpoint value was used for a station.

The observational data used for verification is expressed in terms of which one of the three categories the observation fell in at each location, where the tercile thresholds are defined using the 30-yr climatology from 1981 to 2010 derived from the station observational dataset. The 7-day mean temperature and accumulated precipitation observed values were calculated using CPC's daily station observations (CPC precipitation/temperature tables; CPC 2016). This serves as an independent observation dataset for fairly assessing the skill of the forecasts (as opposed to using the reforecast model analyses as the observations).

This CPC station dataset was used for verification because of the following reasons. 1) It was desirable to use the same format of verification data as the other forecast guidance tools used at CPC (mostly in station format) to enable comparison to other tools in other evaluations. 2) NCEP reanalysis data are not updated with timeliness required for real-time verification, precluding these results from being consistent with real-time skill evaluation. Different datasets were used for verifying and training because we think it is important

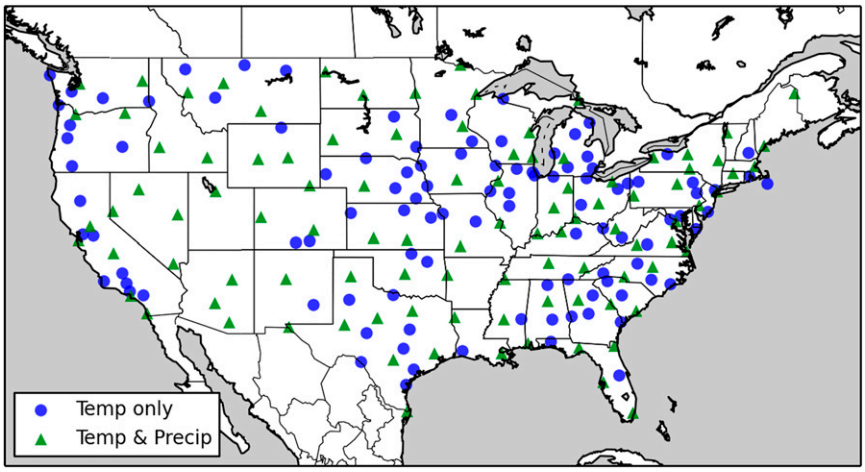

FIG. 1. Station locations of temperature and precipitation forecasts and observations verified in this study. Green triangles represent locations where both temperature and precipitation data are available. Blue circles indicate locations where only temperature data are available.

that an independent dataset is used for evaluation. This gives a more objective, fair benchmark of forecast skill expected in real-time operational forecasts. Previous verification tests we have done have shown that using the same training datasets for verification can inflate skill scores relative to real-time results.

c. *Sample size experiments*

To test the sensitivity of the forecast skill to the size of the reforecast dataset, we created 12 different configurations or subsets of reforecast data (including a control configuration that includes all data in the reforecast data pool) to calculate the calibration statistics. These subsets of data are aggregated based on varying three parameters of the reforecast configuration: 1) the number of years, 2) the number of ensemble members, and 3) the frequency of model runs (referred to as the model run frequency). Configurations were chosen to cover various possible combinations of reforecast parameters, including smaller configurations that may indicate ''breakpoints,'' in which the skill significantly drops from larger sample sizes. Table 1 describes the various configurations considered.

References to the number of ensemble members include the control run (e.g., 11 members refers to 10 members plus the control run). A configuration with only one member (configuration number 11) includes only the control run member. Subsets of training data with a model run frequency of once per week only include reforecast data from Thursdays and those with a twice per week frequency include Mondays and Thursdays. Thursday was selected to be consistent with other previous studies (Hagedorn et al. 2012), and Monday was used to approximate evenly skipped days.

TABLE 1. List of reforecast dataset configurations used for calibrating forecasts with the specifications of the three changing parameters of the dataset: number of years, number of ensemble members, and model run frequency.

Configuration No.	No. of years	Years	No. of ensemble members	Model run frequency
0 (control)	25	1986–2010	11	Daily
1	10	2001–10	11	Daily
2	25	1986–2010	6	Daily
3	25	1986–2010	11	Twice per week
4	25	1986–2010	11	Once per week
5	25	1986–2010	6	Once per week
6	10	2001–10	11	Once per week
7	10	2001–10	6	Once per week
8	18	1993–2010	6	Once per week
9	25	1986–2010	3	Once per week
10	10	2001–10	3	Once per week
11	25	1986–2010	1	Once per week

d. Methodology

The term postprocessing typically refers to the general concept of applying statistical corrections to DMO. There are a variety of widely accepted and practiced methods of postprocessing DMO, such as analog techniques (Hamill and Whitaker 2006), nonhomogeneous Gaussian regression (NGR), logistic regression, Gaussian ensemble dressing (Wilks and Hamill 2007), Bayesian model averaging (Raftery et al. 2005), ensemble kernel density model output statistics (EKDMOS; Glahn et al. 2009; Veenhuis 2013), and ensemble regression (ER; Unger et al. 2009), as well as newer techniques such as censored shifted gamma distributions (CSGD), which focuses on calibrating non-Gaussian forecast quantities (Scheuerer and Hamill 2015). ER was chosen as the postprocessing method for the reforecast tool because it retains more of the information from the individual model forecast member solutions with the benefit of only needing to derive regression coefficients using the ensemble mean. Basing regression statistics on only the ensemble mean allows statistics to be derived from a smaller number of ensemble members in the reforecast training dataset while allowing as many members as possible in the real-time model forecasts. The ER method accomplishes the following. 1) Model bias is corrected through removal of the model climatology. 2) Variance of model forecasts is corrected to observed variance in standardizing anomalies. 3) Uncertainty represented by the ensemble spread is corrected according to the mean correlation and mean ensemble spread to improve the reliability of the probabilities. 4) Low skill anomaly forecasts are damped such that the predicted probability distribution resembles climatology and tercile probabilities approach a third. EKDMOS has been used for many years by Meteorological Development Laboratory (MDL) to produce reliable forecasts, including out to week 2 (Glahn et al. 2009). ER shares many common characteristics with EKDMOS. Both techniques utilize linear regression to adjust ensemble kernel distributions. Since linear regression is probably the most used postprocessing technique over the history of objective weather and climate forecasting (Glahn et al. 2009), a benefit of using ER is that it is based on a well-established, relatively straightforward methodology, minimizing potential error attributed to using a more complex technique when focusing on sensitivity studies.

Using the ER method, 12 sets of statistics were generated using the training data for each of the reforecast configurations in Table 1. The calibration statistics were smoothed temporally using the triangular mean method (15 days before and after the center date), avoiding potential issues that arise from seasonally varying systematic errors (Hagedorn et al. 2012). We opted to use the same window for both temperature and precipitation, and felt that greater than 30 days may weaken the signals associated with typical synoptic patterns for the weekly period, such as including events from another season. Statistics were calculated for each calendar day, allowing the forecasts to be calibrated with a unique set of statistics daily. Reforecast configurations with skipped days (nondaily model run frequency) benefit from the compensation of using more training years because of the greater number of days that would be incorporated in each calendar day of the calibration statistics. Temporal smoothing of calibration statistics may reduce the impact of reductions in the number of ensemble members in reforecast subsets in this study, because the additional noise in the model ensemble mean with fewer ensemble members may be reduced through temporal averaging across multiple initialization days.

The following steps are taken (using the past forecasts and ''observations'') to derive the statistics for each of the reforecast configurations: 1) calculate the covariance between the forecasts and associated observations and the respective variances, 2) use the covariance and variances to calculate correlations between the ensemble mean forecasts and observations, and 3) use the correlation values and sample standard deviations from the forecasts and observations to calculate the regression coefficients. An extra step is performed for precipitation since ER assumes that the ensemble member errors are Gaussian distributed about each member. We do a log transformation of the precipitation amounts, such that the member error is assumed to be proportional to the precipitation amount, and the regression is between forecast and observed log-precipitation values.

Zero precipitation forecasts and observations are not separated from the regression. However, values below 1 mm (log precipitation = 0) were considered to be no measurable precipitation and assigned a log-precipitation value of −2 (or 0.01 mm). Errors for zero precipitation, either forecast or observed, are relative to a value of −2. In especially dry areas where zero or trace amounts of precipitation frequently occur, errors in the probability of precipitation above 1 mm will play a significant role in the estimated ensemble member errors.

To calibrate the individual ''real time'' forecast members using ER, the following procedure is performed for each of the reforecast configurations: 1) ensemble member forecast anomalies are calculated by subtracting the model climatological mean generated using the data included in the reforecast configuration, 2) standardized forecast anomalies are generated using

the model variance included in the reforecast configuration, 3) anomalies are multiplied by correlations previously derived from the ensemble mean and observations, 4) linear regression is applied to all 11 individual ensemble members using the derived statistics, and 5) all ensemble members are dressed with Gaussian kernel distributions (see, e.g., Hastie et al. 2009) to represent the expected error of the ''best member'' and describe the error in the ensemble distribution. The Gaussian kernel distributions about individual ensemble members are summed to form the full calibrated-probability distribution of the ensemble forecast, before calculating tercile probabilities. It should be noted that the model anomalies, variances, and covariances with observations were calculated based on the years and dates of each configuration.

Week-2 probabilistic temperature and precipitation forecasts were evaluated over the CONUS from 2011 to 2013 (3 yr) using three skill score metrics: RPSS, Heidke skill score (HSS), and reliability (Wilks 2006). We focus mainly on RPSS and reliability because these metrics assess the probabilistic aspect of the forecasts. Additionally, we review the skill of forecasts broken down by 3-month seasons to see if there is a seasonal component in the behavior of forecasts to reforecast subsampling. The statistical significance of skill score differences was determined using a bootstrapping method, resampling sample scores across the various reforecast configurations 10 000 times. Bootstrapping assumes that resampled statistics are drawn from independent and identical distributions to the evaluation period (2011–13), and so statistical significance is evaluated relative to the same period. This method was chosen because it is our intent to test the calibration of recent forecasts from prior reforecasts.

Reliability diagrams were constructed by using 10 bins of forecast probabilities. For each of these bins, a cumulative count across all categories of forecasts and observations is taken to obtain a reliability value:

$$\text{reliability} = (O_A/F_A) + (O_B/F_B) + (O_N/F_N), \qquad (1)$$

where subscripts B, N, and A represent the below-, near-, and above-normal (median) categories for temperature (precipitation); F is the number of forecasts of the specified category with a probability that is within the range of the probability bin being assessed; and O denotes the number of occurrences where the forecast within that probability bin correctly forecast that category. This format of the reliability score represents how frequent the correct category is forecast compared to the forecast probability, assessed cumulatively across the three categories for each probability bin. Reliability values are plotted as a function of the probability bin being assessed.

3. Results

a. Sensitivity to reforecast dataset size

Average HSS and RPSS results for temperature and precipitation are plotted as a function of decreasing configurations of the three parameters (number of training years, number of ensemble members, and model run frequency) to assess possible configuration thresholds that lead to a significant drop in skill (Fig. 2). Configurations were selected so that one reforecast parameter changes but the other two parameters have the same value, isolating the changes of a specific parameter. Vertical red lines indicate the first encountered decrease in skill with a significance level greater than or equal to 90% (going from configurations with greater sample size to less). Our results show that reducing the number of years of the reforecast dataset used for training the reforecast tool leads to the greatest drop in skill, while the least loss is associated with reducing the model run frequency. Similar findings were evident in the Hamill et al. (2004) evaluation of 6–10-day/week-2 reforecast-calibrated forecasts using the older 1998 version of the GEFS model, and the 2005 version of the ECMWF model (Hamill and Whitaker 2015). This minimal skill loss associated with skipping model runs may be attributed to forecast errors being highly correlated when the days are closer together, thus reducing the effective sample size (Hamill et al. 2008; Hagedorn et al. 2008). Hamill et al. (2004) actually found an improvement in skipping days in the reforecast sample, which was attributed to having a sample that spans a wider range of meteorological scenarios than those captured by daily forecasts with fewer years. Sensitivity studies performed by the MDL found skill improvement in skipping reforecast days as well when looking at days 1–8 wind and precipitation types (Hamill et al. 2014).

There is some skill sensitivity to using fewer ensemble members. Both HSS and RPSS show statistically significant degradation in skill when using fewer than six ensemble members and drastic skill loss when using only one member. Hamill and Whitaker's reforecast sensitivity study of the 2005 version of the ECMWF (Hamill and Whitaker 2015) found that three members was a sufficient number for calibrating the 6–10-day temperatures, with a range from five to seven members adequate for precipitation. There may be greater gains for the GEFS model using at least six members for temperature forecasts because the current, higher-resolution 2013 version of the ECMWF

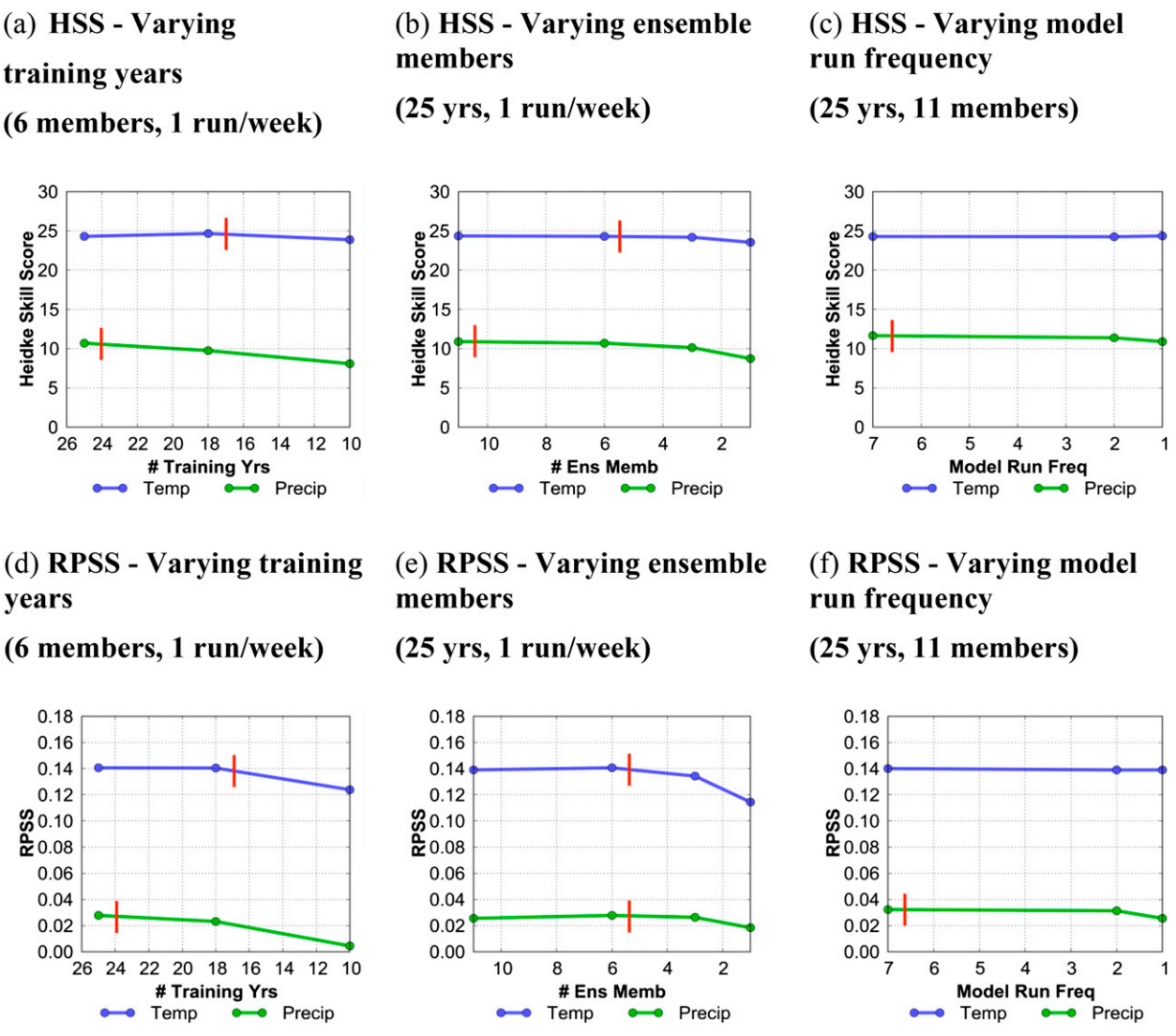

FIG. 2. (top) HSS and (bottom) RPSS of week-2, tercile category forecasts of surface temperature (blue lines) and precipitation (green lines) from the reforecast tool as a function of the (a),(d) number of years of training data, (b),(e) ensemble members, and (c),(f) model run frequency. Skill scores represent the average score across the CONUS for forecasts valid from dates centered on 1 Jan 2011–31 Dec 2013 (3 yr). Vertical red lines indicate the first encountered decrease in skill with a significance level greater than or equal to 90% (going from configurations with greater sample size to lower). Subsequent skill differences (both positive and negative) are significant.

model has superior skill to the 2012 GEFS. Hamill et al. (2004) showed that even though the control run was comparable to using the 15-member ensemble mean, there was a greater skill degradation for precipitation and week-2 forecasts. The benefit of adding more ensemble members decreases as the ratio of the predictable signal (i.e., ensemble mean anomaly) to the unpredictable noise (i.e., ensemble spread) increases (Hamill et al. 2004). By the week-2 time scale, this ratio likely becomes smaller as a result of the increasing unpredictable noise component, making it necessary to have at least a few members to capture information about the uncertainty.

Our results support the well-known fact that the skill of precipitation is more sensitive to reforecast configurations than temperature. For example, the reliability curves from our precipitation forecasts have greater spread among varying reforecast configurations than for temperature and there are quicker, steeper drops in skill with smaller reforecast configurations. The skill for precipitation forecasts starts to decrease when dropping to 18 training years, although temperature only shows skill loss when dropping to fewer than 18 training years. Similar to the findings of Hamill et al. (2004), the RPSS drops to nearly zero when only 10 training years are used for calibrating precipitation forecasts. RPSS indicates

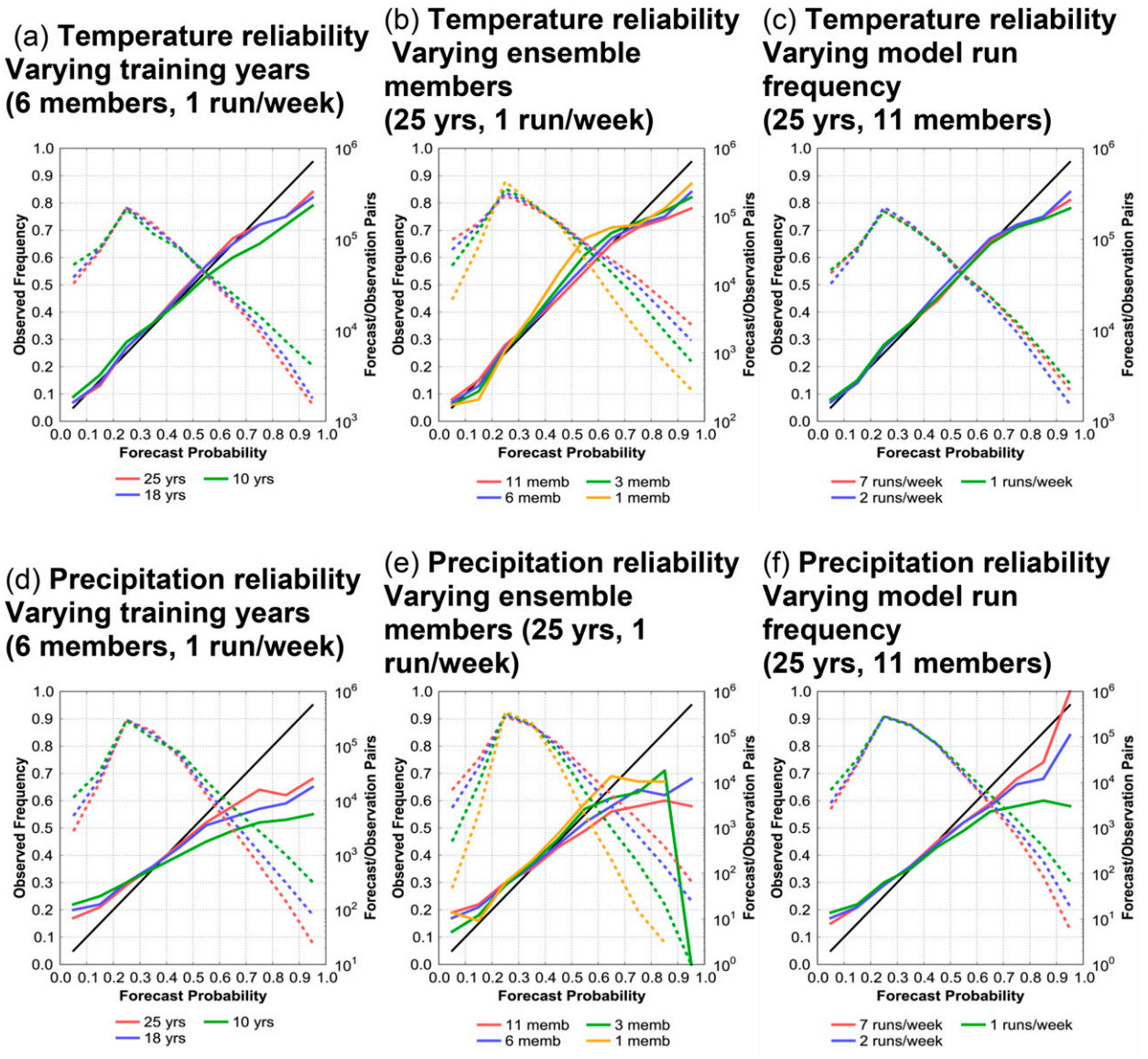

FIG. 3. Reliability diagrams for (top) temperature and (bottom) precipitation as a function of the (a),(d) number of years of training data, (b),(e) ensemble members, and (c),(f) model run frequency.

greater loss in skill for temperature than precipitation when dropping to fewer than six ensemble members and successively fewer members. On the other hand, there is minimal skill loss for both temperature and precipitation when reducing model runs from daily to once per week. More skill loss is observed for precipitation, compared to temperature, when dropping from two to one run per week. Precipitation is likely more sensitive to lower reforecast configurations because this quantity may be rare at many locations requiring longer periods of training data to adequately sample and calibrate precipitation events, compared to temperature, which is a continuous quantity (Müller et al. 2009).

Reliability diagrams (Fig. 3) yield results similar to those of the HSS and RPSS line plots (Fig. 2). In general, for both temperature and precipitation, dropping to 10 training years and fewer than three members significantly decreases the reliability. The exception to this is the reliability of precipitation forecasts associated with changing the number of ensemble members. In this scenario, the reliability for precipitation is similar between using one and three members. This is likely because of a statistical artifact attributed to fewer members producing a lower correlation estimate and lower probabilities as a result of greater noise in the ensemble mean. The lower forecast probabilities produced by fewer members may

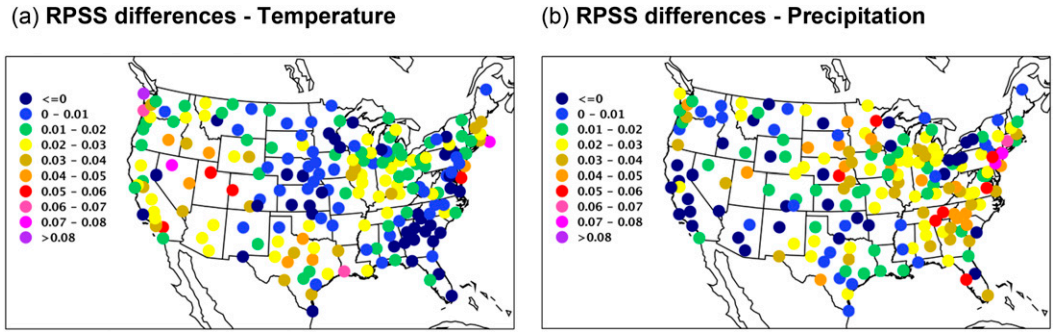

FIG. 4. Skill score differences at each station of RPSS averaged over 1 Jan 2011–31 Dec 2013 using 18 training years minus RPSS using 10 training years for (a) temperature and (b) precipitation forecasts.

coincidentally align more closely with the reality of the naturally high uncertainty of precipitation forecasts rather than intentionally and accurately estimating the true correlation and uncertainty.

Since our results show that the number of training years in the reforecasts impacts skill the most, we focus on this parameter to determine whether there are region-specific impacts to skill using subsamples of reforecasts for calibration. Because the drop in HSS and RPSS (averaged over the CONUS and all available forecast dates) is greatest for both temperature and precipitation when dropping from 18 to 10 yr (compared to from 25 to 18), we evaluate the spatial aspect of the skill change based on these configurations. The mean RPSS (across 2011–13) of calibrated forecasts using 10 training years is subtracted from those using 18 training years, at each station (Fig. 4; HSS not shown since the results are similar to those of RPSS). This allows us to evaluate the skill sensitivity of using fewer training years over various locations.

These skill maps show that there are many locations across the CONUS that suffer significant skill loss for both temperature and precipitation forecasts when decreasing to 10 training years. When comparing precipitation skill to temperature, precipitation forecasts experience more loss, both spatially and quantitatively. In terms of regional impacts, the greatest skill loss for temperature occurs over Texas and areas west of, and including, the Continental Divide. For precipitation, however, the most noticeable skill loss is mainly across the eastern half of the CONUS, especially along the Northeast coast. Overall, stations with the greatest skill loss were on the order of an HSS loss of 8 or greater (not shown), and an RPSS loss of 0.07 or greater. Some of these areas that benefit most from using more training years may be as a result of having clearly detectable systematic errors, in which the calibration corrects for statistical downscaling of the forecasts in addition to the broader scales resolved by the model, such as regions with complex terrain and coastal grid points (Hagedorn et al. 2012). These areas are typically harder to forecast for because of the diverse nature of their synoptic events and, thus, require more events in the reforecast dataset to capture the various flavors of the possible outcomes.

b. Seasonal skill sensitivity to reforecast dataset size

Skill evaluation was performed for the three years evaluated, partitioned by 3-month seasons—December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON)—to explore the potential seasonality in the impacts of lower reforecast configurations to forecast skill. Our findings show that there are seasonally based impacts from changing the reforecast configurations. Temperature and precipitation forecasts have the highest skill as well as the most skill sensitivity to changing reforecast configurations during MAM (Figs. 5 and 6), followed by DJF. JJA had the lowest skill and least sensitivity to configuration subsampling (only MAM results are shown). This result is evident in RPSS, HSS, and reliability. There is more loss in RPSS compared to HSS when using fewer training years and ensemble members, which may indicate that the probabilities are more affected than the forecast categories when using smaller reforecast configurations (Fig. 5).

Across the seasons (including those not shown), dropping to 10 training years and three members leads to the most noticeable skill loss in the temperature and precipitation forecasts. This is especially evident for MAM precipitation (Fig. 6c), where dropping to 10 training years results in lower reliability across most of the forecast probabilities (compared to, e.g., using 18 training years, where significant skill loss occurs at probabilities of 0.7 or greater). Model run frequency does not greatly alter the skill of the temperature forecasts (not shown). However, reliability diagrams do reveal degradation in precipitation forecasts when going from two runs to one run per week during MAM and SON, especially at probabilities greater than 0.7 (not shown).

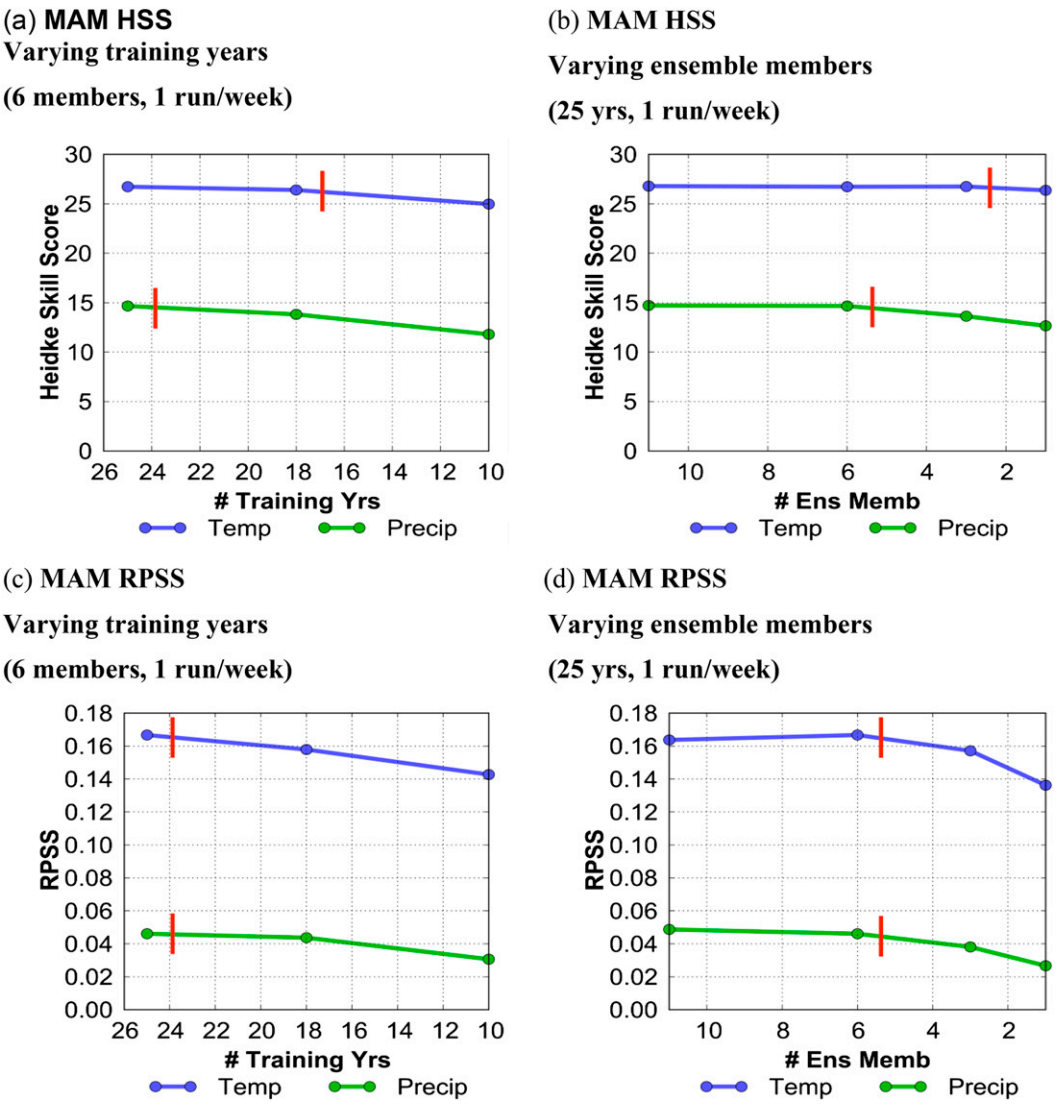

FIG. 5. As in Figs. 2a,b,d,e, but for March–May.

We also assessed reforecast configurations similar to Hamill et al. (2004) to compare the skill and skill sensitivity of the week-2 temperature and precipitation DJF forecasts from the newer 2012 version to the older 1998 version of GEFS (evaluated by Hamill) to identify potential differences in using different model versions. It should be noted that there are some differences between Hamill's study and ours that may account for some of the differences in the results. For example, our analysis uses 25 years and 11 members, daily, while Hamill's uses 23 training years, and 15 members daily. Differences aside, the RPSS from the 2012 GEFS seems to be slightly improved from those using the older 1998 GEFS, especially from configurations using the largest number of training years. Our DJF temperature (precipitation) forecasts using 25 training years yields an RPSS of 0.18 (0.05), which is a 0.03 (0.02) improvement over Hamill's results using 23 training years from the older GEFS. Our results may also indicate that the there is more skill sensitivity to reforecast sample size using the newer 2012 GEFS model compared to the 1998 version. Dropping from 25 to 10 training years yields a decrease in temperature RPSS of about 0.03, whereas using the 1998 GEFS showed a RPSS loss of 0.01 when using 9 instead of 22 years (Hamill et al. 2004). This difference may be attributed to a number of reasons, such as greater skill loss with a higher-resolution, improved model, or as a result of the years selected for verification. There is also a possibility that the difference when using two more training years in our

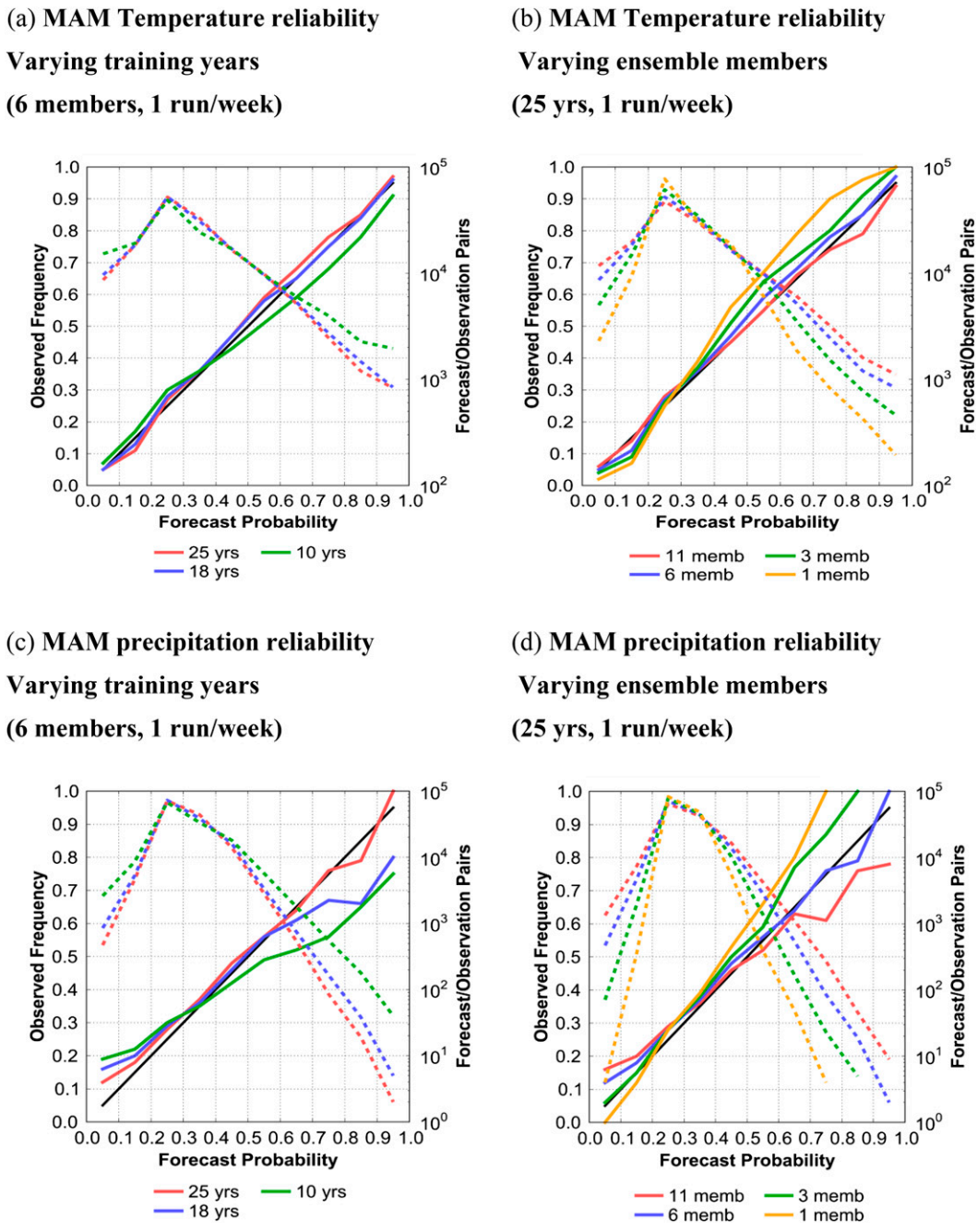

FIG. 6. As in Figs. 3a,b,d,e, but for March–May.

study, compared to Hamill's, contributed to more loss. Our study included fewer members than Hamill et al.'s (2004), which may support the concept that using a few less members does not impact the skill as much as using fewer training years. Precipitation forecasts experienced about a loss of 0.01 for both versions of the GEFS.

The regions that show the greatest skill improvement in DJF temperature forecasts (using 18 instead of 10 training years) occur over the eastern third of the CONUS and southern Texas, whereas for precipitation improvements are greatest across many areas west of the Continental Divide, the Southeast, and the Ohio valley (Figs. 7e,f). These areas differ greatly from the maps showing skill differences considering all seasons together (Fig. 4), as well as JJA (not shown), exemplifying the seasonal and regional variability in skill sensitivity. As a note of interest, the

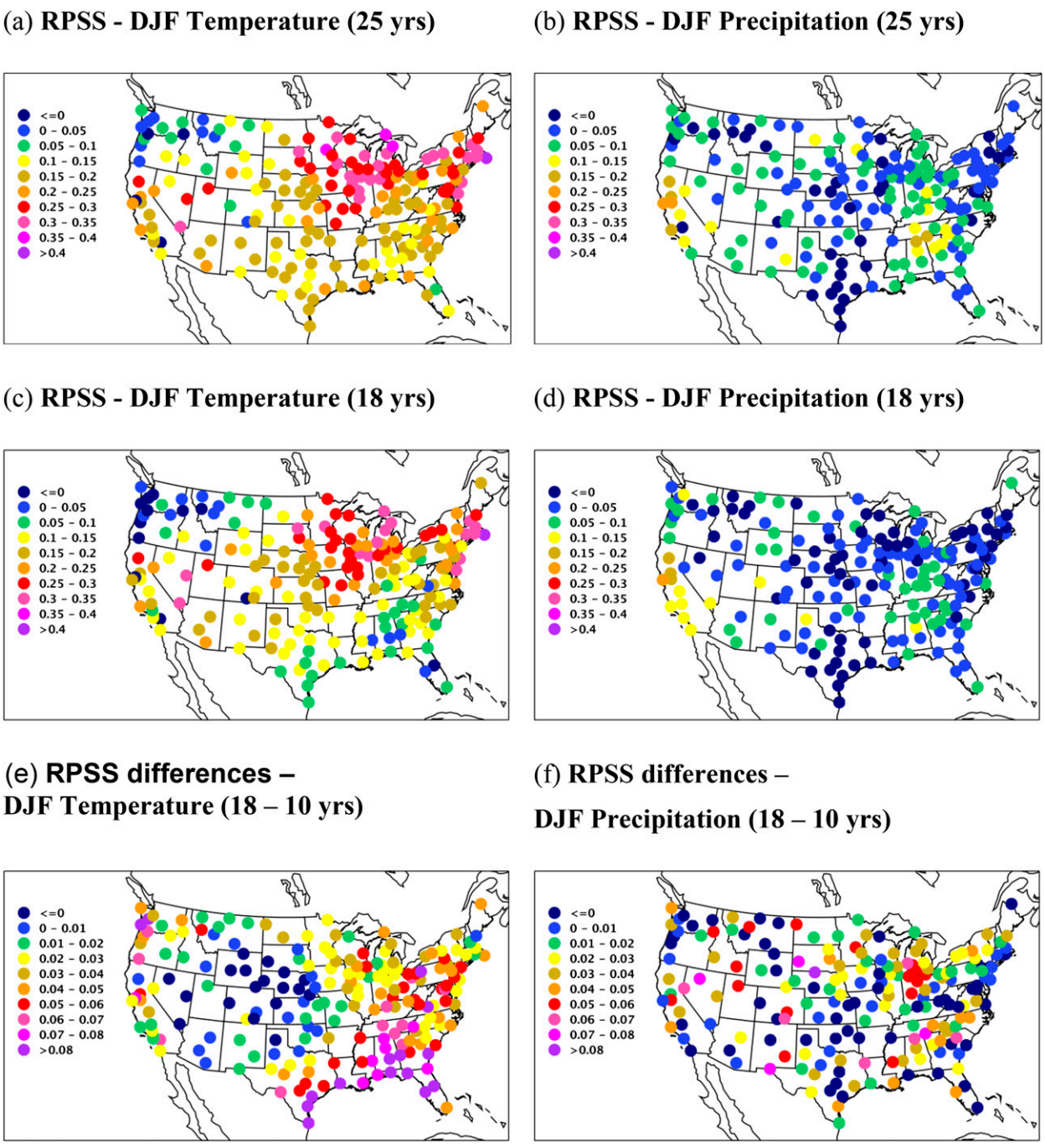

FIG. 7. RPSS at each station averaged over the evaluation period for (left) DJF temperature and (right) precipitation using (a),(b) 25 and (c),(d) 18 yr. Skill score differences using 18 training years minus 10 training years for DJF (e) temperature and (f) precipitation forecasts. Evaluation period is 1 Jan 2011–31 Dec 2013.

most skill gained by using the newer 2012 GEFS [compared to the 1998 GEFS skill results from Hamill et al. (2004)] is across the Upper Midwest and the Northeast region/mid-Atlantic coast for temperature, and over California and east of the Mississippi valley for precipitation. Using only 18 training years (Figs. 7c,d) actually leads to some areas (such as the south-central United States and the Southeast) having the same or lower skill compared to the 1998 GEFS, whereas 25 training years (Figs. 7a,b) produces many more locations with improved skill, especially for temperature forecasts.

4. Conclusions and recommendations

In summary, the reduction in the number of training years of the reforecast dataset leads to the greatest skill loss for week-2 surface temperature and precipitation forecasts, while model run frequency impacts skill the least. Our findings, in addition to those of a number of other studies (Raftery et al. 2005; Hagedorn et al. 2008, 2012), reinforce the importance of having a large number of training years consistent with the current forecast model for successful calibration, especially at longer forecast leads. The gain in skill by using reforecasts with many training years for calibration can rival the improvement that would take 5–10 yr of numerical modeling system development and model resolution upgrades (Wilks and Hamill 2007). A long training dataset is especially important for longer leads because more samples across diverse climate conditions are needed to identify the systematic bias as the error increases with lead time because of chaos (Hagedorn et al. 2008), with implications for the use of reforecasts for model postprocessing for week 3 and 4 forecasts. It may be possible to obtain reforecasts with diverse climate conditions from fewer years through the careful selection of the reforecast years; however, avoiding introducing other systematic biases through the selection criteria may be difficult.

We determine that fewer members are needed in reforecasts for the calibration to be effective, as long as the real-time run ensemble has more members (currently the 2012 GEFS has 21 real-time members) (Hamill and Whitaker 2015). In general, using reforecasts with only one model run per week instead of two seem to retain the most skill as the bias can still be calculated well, although it does lead to precipitation forecast probabilities that are much more poorly calibrated. The most significant drop in forecast skill occurs when decreasing the reforecast configuration to 10 training years and one ensemble member (i.e., the control run).

The regions most sensitive to reducing the reforecast configuration to 10 training years are Texas and areas west of, and including, the Continental Divide for temperature, and the eastern half of the CONUS for precipitation, especially along the Northeast coast. Regions impacted most by lower reforecast configurations differ depending on the temporal aggregation for evaluation, such as for various seasons. Overall, the RPSS and reliability diagrams show a greater degree of variance in skill than does HSS with changing reforecast configurations and are therefore better indicators of the sensitivity of skill to the sample size of the training dataset. This indicates the advantage of a larger sample of reforecasts for making improvements in reliability and resolution. The results from these skill metrics are essential because they evaluate the quality of the probabilistic attributes of the forecasts.

To retain week-2 surface temperature and precipitation skill, CPC provided the recommendation to NCEP that the reforecast dataset have as many years as possible (at the very minimum, 18 years), six ensemble members (five members plus a control run), and a model run frequency of once per week. Skipping days between forecast samples and only running the reforecast with a few members can significantly reduce the cost of the production of the reforecasts (Hamill et al. 2004) without compromising the forecast skill for the week-2 temperature and precipitation. However, CPC would prefer to have 30 yr of reforecast data, for two reasons. First, 30 yr would match the number of years used as a standard for climatology at CPC (and most other meteorological centers). This would allow the climatologies of the model forecast to be matched to our observation climatologies for more accurate calibration and avoid adjustments needed to determine the category thresholds. Second, there must also be enough reforecast data to sufficiently capture the systematic errors associated with a variety of synoptic events (Wilks and Hamill 2007). This is of importance to CPC because the reforecast tool is currently being used as the basis of the week-2 probabilistic extremes forecast tool, which serves as the main guidance product for the week-2 probabilistic hazards forecasts issued by CPC. For extremes, the importance of reforecast data with long, consistent model climatologies is well understood (Hagedorn et al. 2008; Vitart et al. 2008). There must be enough years to include a sufficient number of independent samples of rare weather events and patterns to properly calibrate the forecasts, especially when working with forecasts at longer leads (Hamill et al. 2006). Recent studies have shown that uncommon, high-impact forecast parameters such as heavy precipitation and high winds (e.g., greater than or equal to 10 kt, where 1 kt = 0.51 $m\,s^{-1}$) tend to be more sensitive to reforecast sample size than those that are more common such as light precipitation (Hamill et al. 2014). Since there is increasing focus on extreme events that are of high impact, it is important to ensure that reforecasts can be used to properly calibrate the model forecasts for accurate guidance. The hydrological community has also expressed a need for a long reforecast dataset spanning 30 yr or greater for calibrating and validating the characteristics of streamflow forecasts over a large sample of high-impact cases (Hamill et al. 2014).

This study has helped to inform NCEP of requirements for the reforecast configuration for upcoming updated datasets. Based on the requirements provided in the collaborative white paper (Hamill et al. 2014), NCEP will be producing reforecasts in the

TABLE A1. Threshold percentage of pentads with no precipitation (center column) and the estimated expected percent of correct forecasts by chance for each of the three categories of below, normal, and above, respectively for various climatological precipitation classifications (left column). Percentage of no precipitation is denoted as 'PNP'.

Climatological precipitation classification	Threshold percentage of pentads with no precipitation according to the dry station climatology	Expected percent of correct forecasts by chance for each of three categories (below, normal, above), respectively
Arid station	≥67%	PNP, 0, 1 − PNP
Semiarid station	≥34% and <67%	0.667, 0, 0.333
Normal station	<34%	0.333, 0.333, 0.333

upcoming months for the new operational GEFS (version 11), which was updated in December 2015. These will be produced with the configuration of 20 years, five ensemble members, and runs once every 4 days. Other previous work has also shown that at least 20 years of reforecasts sufficiently improves the forecast skill for many thresholds (including rare events), as well as the lead times for precipitation (Fundel et al. 2010). This study only included discrete selected configurations to cover a range of sample sizes, which is why the overall collective reforecast white paper recommendations (Hamill et al. 2014) slightly differ from the requirements as indicated by our results (e.g., we did not assess 20 training years specifically). This work and our current real-time reforecast tool forecasts (as of December 2015) use the legacy model runs from GEFS version 10, but we will be regenerating the calibration statistics using the reforecasts from the updated GEFS version 11 once the new dataset is available.

Acknowledgments. The authors thank ESRL for providing the reforecast data used in this study, Tom Hamill (ESRL) for his discussions regarding reforecasts and model postprocessing, and David Unger for his comments, edits, and contributions in the CPC reforecast tool project and evaluation.

APPENDIX

Dry Station Verification Methodology

CPC's forecast format requires that precipitation observations be placed in one of three equally likely categories based on climatology. A pentad with no reported precipitation (trace amounts are reported as zero in CPC's dataset) is always classified as "below normal." When the climatological probability of no precipitation at a given location exceeds 33% (defining a "dry" station), then it is no longer possible to classify observations into three equal categories, and adjustments to the verification methodology are needed. The following steps outline the treatment of dry stations in CPC verifications:

Step 1—create a climatology for the percentage of pentads with no precipitation for each month and each station based on the 1971–2000 period, which will be referred to as percentage no precipitation (PNP);

Step 2—classify a station as either arid, semiarid, or normal according to the definitions listed in Table A1;

Step 3—define a dry station as one in either an arid or semiarid location; then, reclassify all forecasts and observations at these locations as either above or below normal based on the following definitions:

- For arid stations,
  - all near-normal forecasts are converted to below-normal forecasts and
  - all near-normal observations are reclassified as below-normal observations;
- For semiarid stations,
  - all near-normal forecasts are converted to below-normal forecasts and
  - all near-normal observations are reclassified as below-normal observations; and
- For normal stations,
  - all forecasts and observations are left as they are.

Step 4—estimate the expected percent of correct forecasts by chance (last column in Table A1) for each of the three categories (below, normal, above) based on the dry station climatology;

Step 5—sum the expected correct percentages and the number of correct forecasts over all categories for each location prior to the calculation of the HSS, defined by

$$\mathrm{HSS}(\%) = 100 \times (H - E)/(T - E), \qquad \text{(A1)}$$

where H is the number of forecasts with the correct category, E is the expected number of correct forecasts by chance, and T is the total number of forecast–observation pairs. For regional summaries, summations over all stations (combining dry and normal stations) are performed prior to the calculation of the HSS.

REFERENCES

CPC, 2016: U.S. precipitation/temperature tables. Climate Prediction Center, accessed 20 February 2016. [Available online at http://www.cpc.ncep.noaa.gov/products/analysis_monitoring/cdus/prcp_temp_tables/index.shtml.]

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

Fundel, F., A. Walser, M. A. Liniger, C. Frei, and C. Appenzeller, 2010: Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts. *Mon. Wea. Rev.*, **138**, 176–189, doi:10.1175/2009MWR2977.1.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schuknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246–268, doi:10.1175/2008MWR2569.1.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, doi:10.1175/2007MWR2410.1.

——, R. Buizza, T. M. Hamill, M. Leutbecher, and T. N. Palmer, 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1814–1827, doi:10.1002/qj.1895.

Hamill, T. M., and S. J. Colucci, 1998: Evaluation of Eta–RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.*, **126**, 711–724, doi:10.1175/1520-0493(1998)126<0711:EOEREP>2.0.CO;2.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, doi:10.1175/MWR3237.1.

——, and ——, 2015: Exploring sample size issues for 6-10 day forecasts using ECMWF's reforecast data set. Accessed 14 August 2015. [Available online at http://www.esrl.noaa.gov/psd/people/tom.hamill/CTB_Hamill_Whitaker_ECMWF_results.ppt.]

——, ——, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

——, ——, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33–46, doi:10.1175/BAMS-87-1-33.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:10.1175/2007MWR2411.1.

——, G. T. Bates, J. S. Whitaker, D. R. Murray, M. Fiorino, T. J. Galarneau Jr., Y. Zhu, and W. Lapenta, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

——, and Coauthors, 2014: A recommended reforecast configuration for the NCEP Global Ensemble Forecast System. NOAA White Paper, 24 pp. [Available online at http://www.esrl.noaa.gov/psd/people/tom.hamill/White-paper-reforecast-configuration.pdf.]

Hastie, T., R. Tibshirani, and J. Friedman, 2009: Unsupervised learning. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, T. Hastie, R. Tibshirani, and J. Friedman, Eds., Springer Series in Statistics, Springer, 485–585, doi:10.1007/978-0-387-84858-7_14.

Johnson, C., and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quart. J. Roy. Meteor. Soc.*, **135**, 777–794, doi:10.1002/qj.383.

Lalaurette, F., 2003: Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quart. J. Roy. Meteor. Soc.*, **129**, 3037–3057, doi:10.1256/qj.02.152.

Lorenz, E. N., 1969: The predictability of a flow which possesses many scales of motion. *Tellus*, **21**, 289–307, doi:10.1111/j.2153-3490.1969.tb00444.x.

Müller, M., M. Kašpar, and J. Matschullat, 2009: Heavy rains and extreme rainfall-runoff events in central Europe from 1951 to 2002. *Nat. Hazards Earth Syst. Sci.*, **9**, 441–450, doi:10.5194/nhess-9-441-2009.

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

Scheuerer, M., and T. M. Hamill, 2015: Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Mon. Wea. Rev.*, **143**, 4578–4596, doi:10.1175/MWR-D-15-0061.1.

Unger, D. A., H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379, doi:10.1175/2008MWR2605.1.

Veenhuis, B. A., 2013: Spread calibration of ensemble MOS forecasts. *Mon. Wea. Rev.*, **141**, 2467–2482, doi:10.1175/MWR-D-12-00191.1.

Vitart, F., and Coauthors, 2008: The new VAREPS-monthly forecasting system: A first step towards seamless prediction. *Quart. J. Roy. Meteor. Soc.*, **134**, 1789–1799, doi:10.1002/qj.322.

Whitaker, J. S., X. Wei, and F. Vitart, 2006: Improving week-2 forecasts with multimodel reforecast ensembles. *Mon. Wea. Rev.*, **134**, 2279–2284, doi:10.1175/MWR3175.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. Academic Press, 627 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, doi:10.1175/MWR3402.1.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 1364–1385, doi:10.1175/MWR3347.1.