Systematic Error Analysis and Calibration of 2-m Temperature for the NCEP GEFS Reforecast of the Subseasonal Experiment (SubX) Project

HONG GUAN

SRG at NOAA/NWS/NCEP/EMC, College Park, Maryland

YUEJIAN ZHU

NOAA/NWS/NCEP/EMC, College Park, Maryland

ERIC SINSKY, WEI LI, AND XIAQIONG ZHOU

IMSG at NOAA/NWS/NCEP/EMC, College Park, Maryland

DINGCHEN HOU

NOAA/NWS/NCEP/EMC, College Park, Maryland

CHRISTOPHER MELHAUSER AND RICHARD WOBUS

IMSG at NOAA/NWS/NCEP/EMC, College Park, Maryland

(Manuscript received 8 June 2018, in final form 19 November 2018)

ABSTRACT

The National Centers for Environmental Prediction have generated an 18-yr (1999–2016) subseasonal (weeks 3 and 4) reforecast to support the Climate Prediction Center's operational mission. To create this reforecast, the subseasonal experiment version of the GEFS was run every Wednesday, initialized at 0000 UTC with 11 members. The Climate Forecast System Reanalysis (CFSR) and Global Data Assimilation System (GDAS) served as the initial analyses for 1999–2010 and 2011–16, respectively. The analysis of 2-m temperature error demonstrates that the model has a strong warm bias over the Northern Hemisphere (NH) and North America (NA) during the warm season. During the boreal winter, the 2-m temperature errors over NA exhibit large interannual and intraseasonal variability. For NA and the NH, weeks 3 and 4 errors are mostly saturated, with initial conditions having a negligible impact. Week 2 errors (day 11) are ~88.6% and 86.6% of their saturated levels, respectively. The 1999–2015 reforecast biases were used to calibrate the 2-m temperature forecasts in 2016, which reduces (increases) the systematic error (forecast skill) for NA, the NH, the Southern Hemisphere, and the tropics, with a maximum benefit for NA during the warm season. Overall, analysis adjustment for the CFSR period makes bias characteristics more consistent with the GDAS period over the NH and tropics and substantially improves the corresponding forecast skill levels. The calibration of the forecast using week 2 bias provides similar skill to using weeks 3 and 4 bias, promising the feasibility of using week 2 bias to calibrate the weeks 3 and 4 forecast. Our results also demonstrate that 10-yr reforecasts are an optimal training period. This is particularly beneficial considering limited computing resources.

1. Introduction

To provide seamless numerical guidance to a broad range of users and partners, the National Oceanic and Atmospheric Administration (NOAA) is extending its services from weather forecasts (week 1) and extended forecasts (week 2) to subseasonal forecasts (weeks 3 and 4) through the Next Generation Global Prediction System (NGGPS) project. The lack of memory of the atmospheric initial conditions and the effects of the atmosphere–land and ocean–sea ice interactions, which benefit weather forecasts and seasonal and longer forecasts, respectively, create particular challenges to the subseasonal forecasts (Johnson et al. 2014; Li et al. 2018). On the subseasonal time scale, the numerical model is a major driver for forecast error and skill. Thus, improvement in the

Corresponding author: Dr. Hong Guan, hong.guan@noaa.gov

DOI: 10.1175/WAF-D-18-0100.1

dynamical forecast system is critical to improving subseasonal forecast skill. In addition, statistical postprocessing improves the forecast quality after calibration and can also improve the forecast skill. Postprocessing is especially important for the subseasonal time scales due to larger forecast errors that exists at longer lead times.

Regarding the potential improvement of forecast skill in a dynamical forecast system, recent studies demonstrate that improvement in forecast skill can be derived from improved sea surface temperature (SST) forcing (Zhu et al. 2017), updated convection parameterization schemes (Vitart 2009; Zhu et al. 2018; Li et al. 2018), and new stochastic physics (Zhu et al. 2018; Li et al. 2018). These efforts have significantly improved the Madden–Julian oscillation (MJO) forecast skill and 500-hPa geopotential height forecast skill. Although the subseasonal forecast skill in the tropics and the large-scale circulation is improving, the weeks 3 and 4 forecasting of near-surface variables is still challenging. For example, the improvement in raw forecast skill for 2-m temperature and accumulated precipitation is marginal (Zhu et al. 2018). This suggests that developing a suitable postprocessing technique to calibrate the raw forecasts and further improve the forecast skill of near-surface variables is important on subseasonal time scales. Various postprocessing techniques have been proposed and applied to reduce systematic errors and improve the skill levels of probabilistic forecasts on weather and extended weather time scales. These techniques include Kalman filtering (Cheng and Steenburgh 2007), decaying averaging methods (Cui et al. 2012), logistic regression (Wilks and Hamill 2007), nonhomogeneous Gaussian regression (Gneiting et al. 2005), Gaussian ensemble dressing (Roulston and Smith 2003; Wang and Bishop 2005; Bishop and Shanley 2008), Bayesian model averaging (Raftery et al. 2005; Wilson et al. 2007), artificial neural networks (Yuan et al. 2007), and analog techniques (Hamill et al. 2013). Previous studies (Hamill et al. 2004, 2008; Cui et al. 2012; Guan et al. 2015; Guan and Zhu 2017; Ou et al. 2016) have revealed the importance of a hindcast (i.e., reforecast) for extreme weather forecasts or bias correction during week 1 or week 2. Thus, the hybrid decaying and reforecast bias-correction method (Guan et al. 2015) is being operationally applied to the North American Ensemble Forecast System (NAEFS; Candille 2009) in order to improve 1–16-day forecasts.

The major focus of this study is to analyze the spatial and temporal distributions of 2-m temperature bias and identify the saturation characteristics of 2-m temperature error. It is well known that numerical weather forecasting error grows with lead time. An understanding of the error saturation is crucial to further developing an inexpensive reforecast configuration and an effective bias-correction method for operational purposes. Creating a multiyear reanalysis and reforecast datasets requires considerable computational and human resources. It is desirable to produce a high quality forecast using fewer resources. To reach this goal, we determined the time scale when 2-m temperature error reached a saturated level and then address whether the week 2 bias in 2-m temperature can be used to calibrate weeks 3 and 4 forecasts. We also explore the impact of using inconsistent initial analyses on the weeks 3 and 4 forecasts and propose a solution (or analysis adjustment) when a consistent reanalysis dataset is not available for the entire period of study.

The forecast system and datasets are described in section 2. The temporal and spatial distributions of 2-m temperature bias and error saturation follow in section 3. In section 4, we develop weeks 3 and 4 bias-correction methods, including analysis adjustment. A summary and conclusions are given in section 5.

2. Forecast system and data

In May 2017, the National Centers for Environmental Prediction's (NCEP) Environmental Modeling Center (EMC) generated an 18-yr (1999–2016) reforecast dataset to support the NCEP Climate Prediction Center's (CPC) operational mission. With the exception of having a smaller ensemble size than the real-time forecast (1 control member and 10 perturbed members for reforecasts compared to 1 control member and 20 members for real-time forecasts), the Global Ensemble Forecast System (GEFS) used in the present study is the same as the one used by Zhu et al. (2018) and Li et al. (2018). The forecast system is based on the operational GEFSv11 (Zhou et al. 2017), but was upgraded in the following areas: 1) improved model uncertainty representation for the tropics through stochastic physical perturbations, including stochastic kinetic energy backscatter (SKEB; Shutts and Palmer 2004; Shutts 2005), stochastically perturbed parameterization tendencies (SPPTs; Buizza et al. 1999; Palmer et al. 2009), and stochastic perturbed humidity (SHUM; Tompkins and Berner 2008); 2) consideration of the impact of the ocean by using a two-tiered SST approach, which introduces bias-corrected CFSv2 forecast SST (Zhu et al. 2017); and 3) use of updated scale-aware convective parameterizations to improve model physics for tropical convection and MJO forecasts (Han et al. 2017). Each simulation was integrated for 35 days starting at 0000 UTC every Wednesday. The resolution of the model is $T_L574L64$ (~34-km horizontal spacing) during the first 8 days and $T_L382L64$ (~55-km horizontal spacing) for the remaining lead times. The forecast dataset was bilinearly interpolated onto 1° × 1° latitude and longitude grids from the model native resolution.

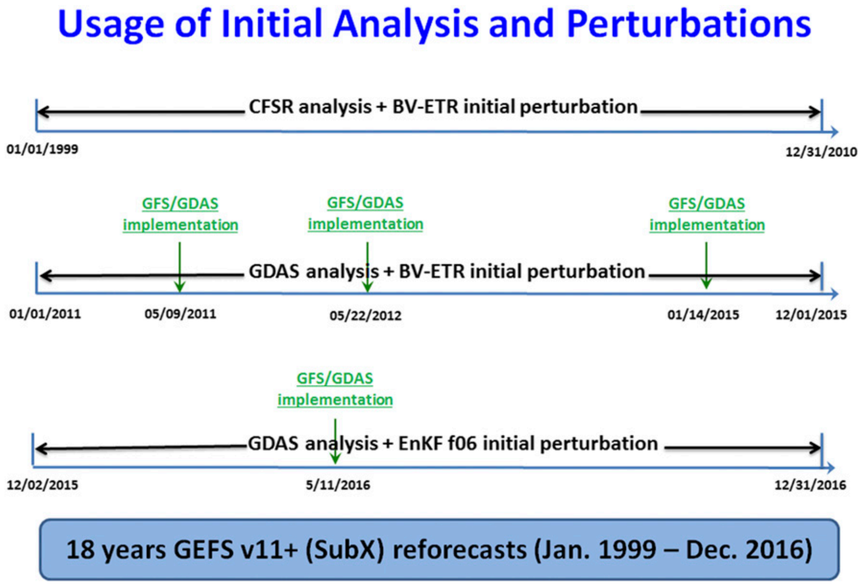
Usage of Initial Analysis and Perturbations
CFSR analysis + BV-ETR initial perturbation
01/01/1999
12/31/2010
GFS/GDAS implementation
GFS/GDAS implementation
GFS/GDAS implementation
GDAS analysis + BV-ETR initial perturbation
01/01/2011
05/09/2011
05/22/2012
01/14/2015
12/01/2015
GFS/GDAS implementation
GDAS analysis + EnKF f06 initial perturbation
12/02/2015
5/11/2016
12/31/2016
18 years GEFS v11+ (SubX) reforecasts (Jan. 1999 – Dec. 2016)

FIG. 1. Evolution of the initial analyses and perturbations during the 18-yr GEFS reforecast period (1 Jan 1999–31 Dec 2016). There were four GFS/GDAS upgrades after switching to the GDAS analysis: 9 May 2011, 22 May 2012, 14 Jan 2015, and 11 May 2016.

Similar to Zhu et al. (2018), the forecast skill levels are defined relative to an NCEP–NCAR 40-yr reanalysis (Kalnay et al. 1996) climatology.

Creating a full set of consistent reanalysis data, including the invariant data assimilation system (and model) and observation systems, is an important part of the reforecast process. A reforecast with initial conditions from a different analysis system would produce a different bias to the forecast. However, the frequent updating of the forecast model, satellite data, or analysis system makes running a reanalysis impractical in operations because generating a multiyear reanalysis is computationally expensive. As illustrated in Fig. 1, this study utilizes two major sets of existing analysis data because a consistent 18-yr reanalysis is unavailable during 1999–2016. The Climate Forecast System Reanalysis (CFSR; Saha et al. 2010) and NCEP operational Global Data Assimilation System (GDAS) [varied generations of the hybrid Gridded Statistical Interpolation (GSI)—ensemble Kalman filter (EnKF)] were the two analysis datasets used as the model's initial conditions during 1999–2010 (CFSR-12) and 2011–16 (GDAS-5), respectively. The analysis data are consistent prior to 2011 and, subsequently, vary with the GFS/GSI/EnKF upgrades after merging during the GDAS period. Note that a new surface roughness formulation in the Global Forecast System (GFS) upgrade of 9 May 2011 (Zheng et al. 2012) led to a significant change in 2-m temperature analysis and forecasts for the arid regions. The current study also provides an opportunity to assess the impact of using initial conditions from different analysis systems on weeks 3 and 4 forecast.

The breeding vector and ensemble transform with rescaling (BV-ETR) technique (Wei et al. 2008) was used to produce initial perturbations for the period of 1 January 1999–2 December 2015 and the hybrid 3D-Var–EnKF data assimilation system was used afterward. The studies of Zhou et al. (2016) show that the initial perturbation could impact the ensemble spread significantly but it has less impact on the ensemble mean errors and forecast skills. Furthermore, the impact on the spread is only limited to the shorter forecast lead times (week 1; Zhou et al. 2016). Therefore, inconsistent perturbation schemes may have a negligible impact on the weeks 3 and 4 forecasts due to the short memory of the atmosphere (Zhu 2005; Song and Mapes 2012).

The full reforecast dataset (1999–2016) will be used for systematic error analysis. In an effort to perform the calibration of the recent forecasts using historical information, the reforecasts from 1999–2015 will be used for calibration and the 2016 data will be withheld as an independent dataset. Therefore, the forecasts being verified during 2016 are independent from the 17-yr training dataset.

3. Bias analysis

To calculate the bias, the analysis fields of CFSR-12 and GDAS-5 were used as an approximate truth. Bias is defined as the difference between the 11-member ensemble mean forecast and the analysis at the time the

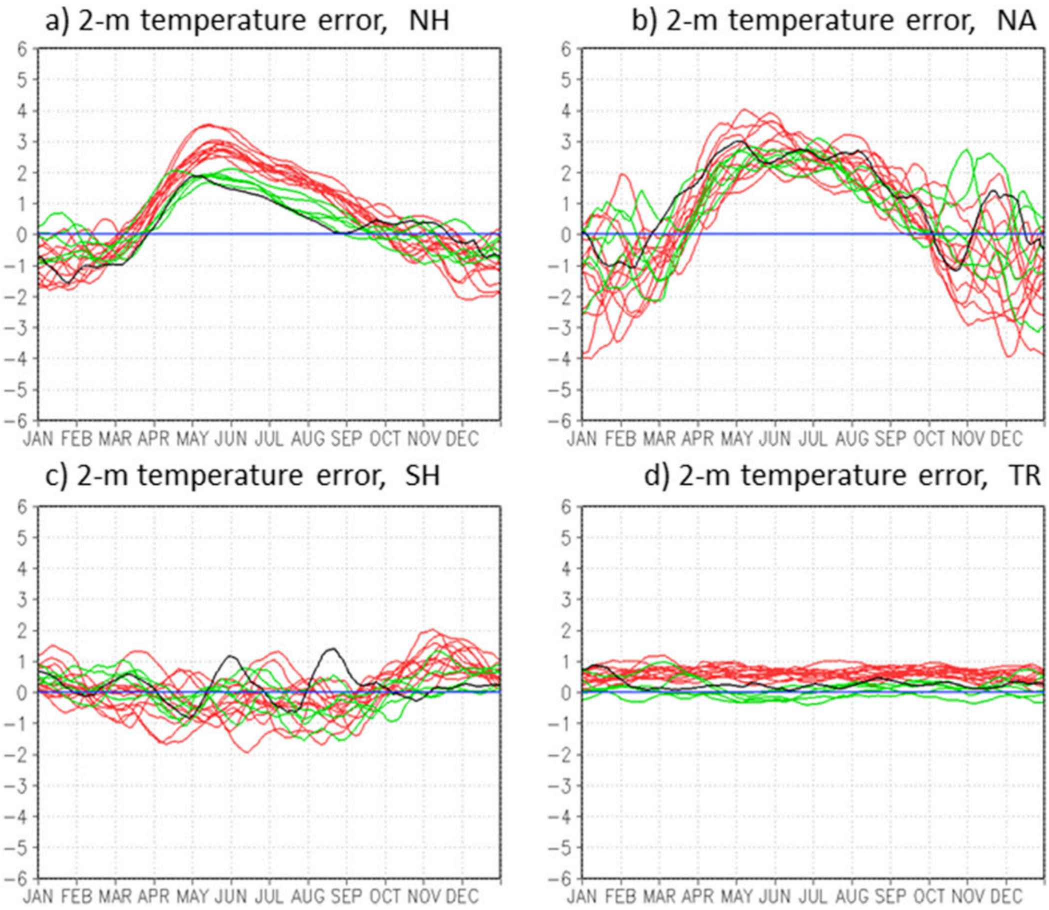

FIG. 2. Time series of 2-m temperature forecast errors for weeks 3 and 4 over (a) the NH, (b) NA, (c) the SH, and (d) the TR domains. Each curve represents one particular year. Red and green curves indicate the errors for the CFSR-12 and GDAS-5 periods, respectively. Thick black curves indicate the errors for 2016.

forecast is valid. The biases of week 2, week 3, week 4, and weeks 3 and 4 are days 8–14, 15–21, 22–28, and 15–28 averaged forecast errors at 0000 UTC, respectively. To calculate the bias climatology from the 18-yr (1999–2016) weekly reforecast dataset, we use a time window of 31 days, centered on the day being considered and leading to a total training dataset of 18 yr × 4–5 samples yr^{-1} = 72–90 samples for each grid point and each forecast time. Any forecast initiated within the 31-day time window falls in the sample. The sensitivity test on the length of time window in Guan et al. (2015) already shows that the 31-day option is an optimal window.

a. Bias distribution

We show the land-only 2-m temperature errors (i.e., bias) over the Northern Hemisphere (NH) (Fig. 2a), North America (NA) (Fig. 2b), the Southern Hemisphere (SH) (Fig. 2c), and the tropics (TR; 20°S–20°N) (Fig. 2d). The errors over the NH (SH) and NA (the TR) display a strong (weak) seasonal dependence. A warm bias is prevalent for the warm season (April–September) for both the NH and NA. It is also evident over NA that the interannual variability of the bias is larger during boreal winter than during other seasons; this alludes to the relatively low predictability of winter-related physical processes. During winter, the ability of the model to forecast 2-m temperature depends significantly on its ability to determine (or assimilate) snow characteristics (Kazakova and Rozinkina 2011; Lavaysse et al. 2013). It has been found that the northern Great Plains, southern Canadian prairies, and the northeastern United States experience high interannual and intraseasonal variability in snow cover and depth (Robinson 1996; Frei and Robinson 1999; Robinson and Frei 2000; Klingaman et al. 2008). Therefore, it is possible that the large interannual variability of 2-m temperature bias over NA during boreal winter was directly associated with the variability of the snow characteristics. It is also noted that in the 2017 GFS upgrade, maximum snow albedo has been adjusted and the snow cover fraction and snow albedo have been unified in Service Change

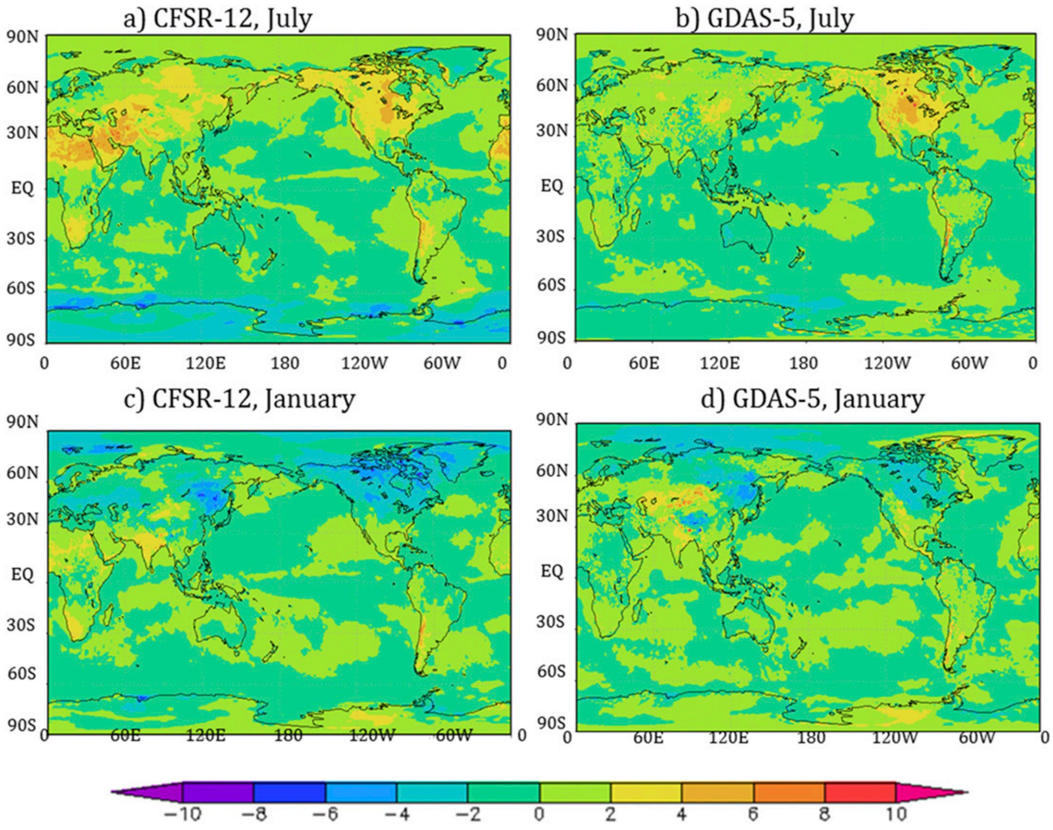

FIG. 3. Spatial distributions of 2-m temperature bias during weeks 3 and 4 (30-day running mean) for July during (a) CFSR-12 and (b) GDAS-5 and January during (c) CFSR-12 and (d) GDAS-5.

Notice (SCN) 1767 (NOAA 2017), which may have some impact on 2-m temperature forecasts during winter. These statements were not explicitly tested in the present study and need to be confirmed in the future.

Figure 2a reveals the tendency to have a larger (slightly larger) warm (cool) bias during the CFSR-12 than during the GDAS-5 for the warm season (cool season) over the NH. The systematic difference in bias characteristics between the two assimilation periods is also noted over the TR (Fig. 2d). The warm bias is prevalent for the CFSR-12, while the bias is near zero for the GDAS-5. To find the cause of the systematic difference between the two analysis system periods, we compare the spatial distributions of global 2-m temperature errors between the CFSR-12 and GDAS-5 for July and January (monthly average) in Fig. 3. In July (Figs. 3a,b), the large difference between the CFSR-12 and GDAS-5 periods occurs near the Sahara and Middle Eastern desert areas. This may be largely attributed to the modification of the surface roughness length formula in the 2011 GFS upgrade (Zheng et al. 2012) that led to a larger change in 2-m temperature analyses and forecasts over arid and desert regions. This speculation will be further discussed in section 4. In January (Figs. 3c,d), the difference between the CFSR-12 and GDAS-5 periods is relatively small except near Kazakhstan, where the opposite biases are noted with a positive bias for the GDAS-5 period and a negative bias for the CFSR-12 period.

b. *Saturation analysis of 2-m temperature errors*

It is well known that forecast error grows with lead time and asymptotically reaches a saturated state. We show the error growth of 2-m temperature forecasts with lead time for the full reforecast period (1999–2016) over NA (Fig. 4a), the NH (Fig. 4b), the SH (Fig. 4c), and the TR (Fig. 4d) domains (land only). For all domains, errors quickly grow within the first 10 days and gradually saturate over weeks 3 and 4. The absolute errors (ABSEs; dotted curve) for NA (Fig. 4a), the NH (Fig. 4b), the SH (Fig. 4c), and the TR (Fig. 4d) are ~79%, 77%, 75%, and 73% of the root-mean-square error (RMSE; solid curve), respectively, at saturation (e.g., day 28 or at the end of week 4). Chai and Draxler (2014) pointed out that RMSE should have the same magnitude as ABSE when the error variance is zero

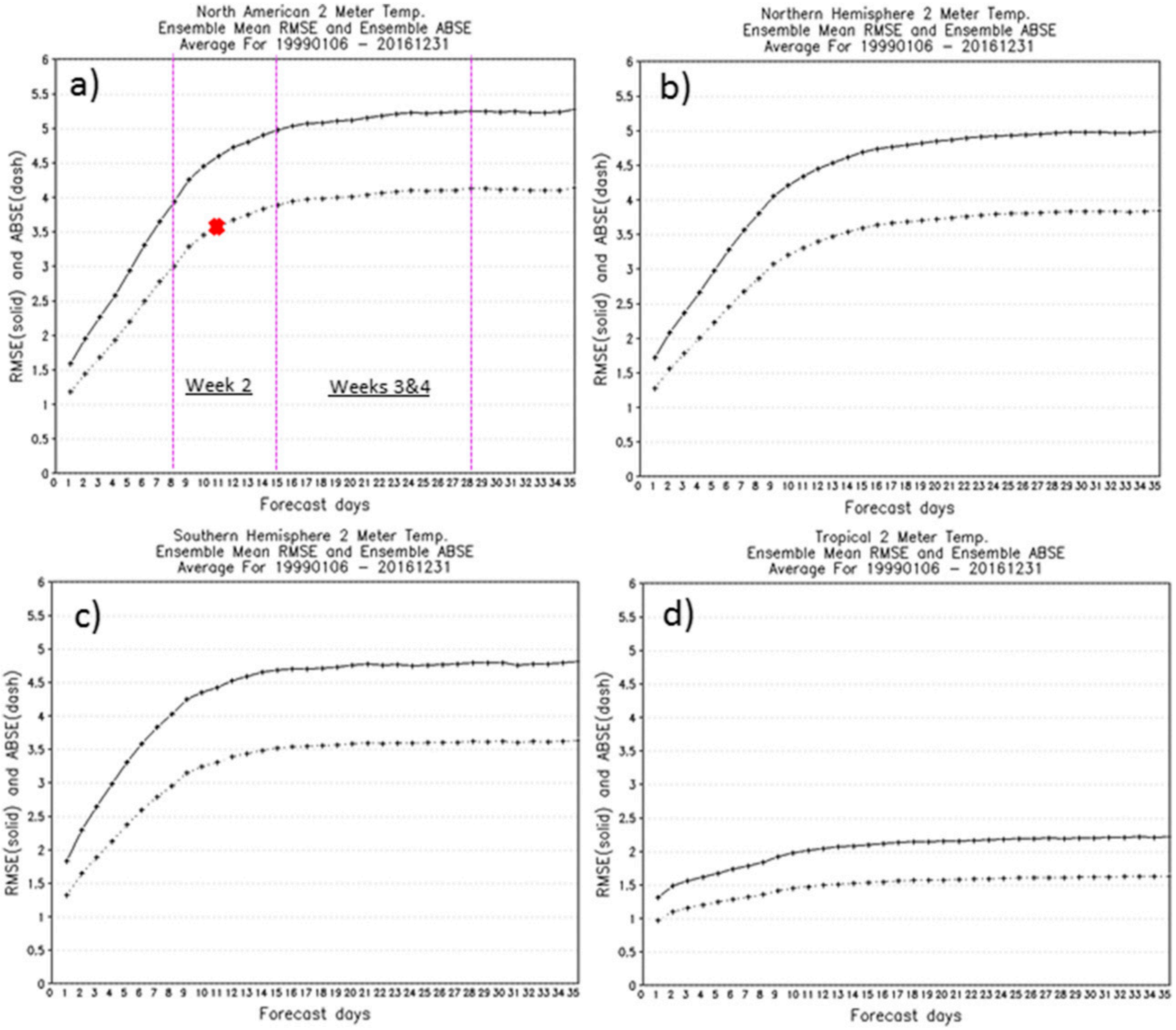

FIG. 4. Domain average (land only) 2-m temperature RMSE (solid curve) and ABSE (dashed curve) during 1999–2016 for (a) NA, (b) the NH, (c) the SH, and (d) the TR from 0 to 35 days.

(i.e., the error is uniformly distributed). In the present study, the contributions of the error variance to the RMSE are less than ~21%, ~23%, 25%, and 27% for NA, the NH, the SH, and the TR, respectively. In general, the errors at saturation over NA are slightly (significantly) larger than those over the NH and SH (TR). On average, the ABSEs for the day 11 (midday of week 2) forecast over NA, the NH, the SH, and the TR are about 88.6%, 86.6%, 91.2%, and 92.5% of their saturation values, respectively. It is understood that the time scale of error saturation is strongly dependent on the geographical area. For example, the time scale for the land error saturation (weeks) is shorter than that for the ocean error saturation (months; Song and Mapes 2012). Our preliminary analysis shows the error saturation time is shorter over the southern contiguous United States (CONUS) than the northern CONUS for both summer and winter (not shown). A detailed diagnosis for the reasons causing this difference is reserved for future work.

To identify if error patterns change with forecast lead times, the global 2-m temperature mean error is compared among weeks 2, 3, and 4 during July and January (Fig. 5). It is evident that the error patterns have nearly fixed geographical structures with lead times in both the summer month (July; Figs. 5a,c,e) and winter month (January; Figs. 5b,d,f). The error magnitudes are shown to be similar during weeks 2, 3, and 4 except over northern NA and Europe, where the errors at weeks 3 and 4 are noticeably more negative than that at week 2 during January. Longer saturation times during

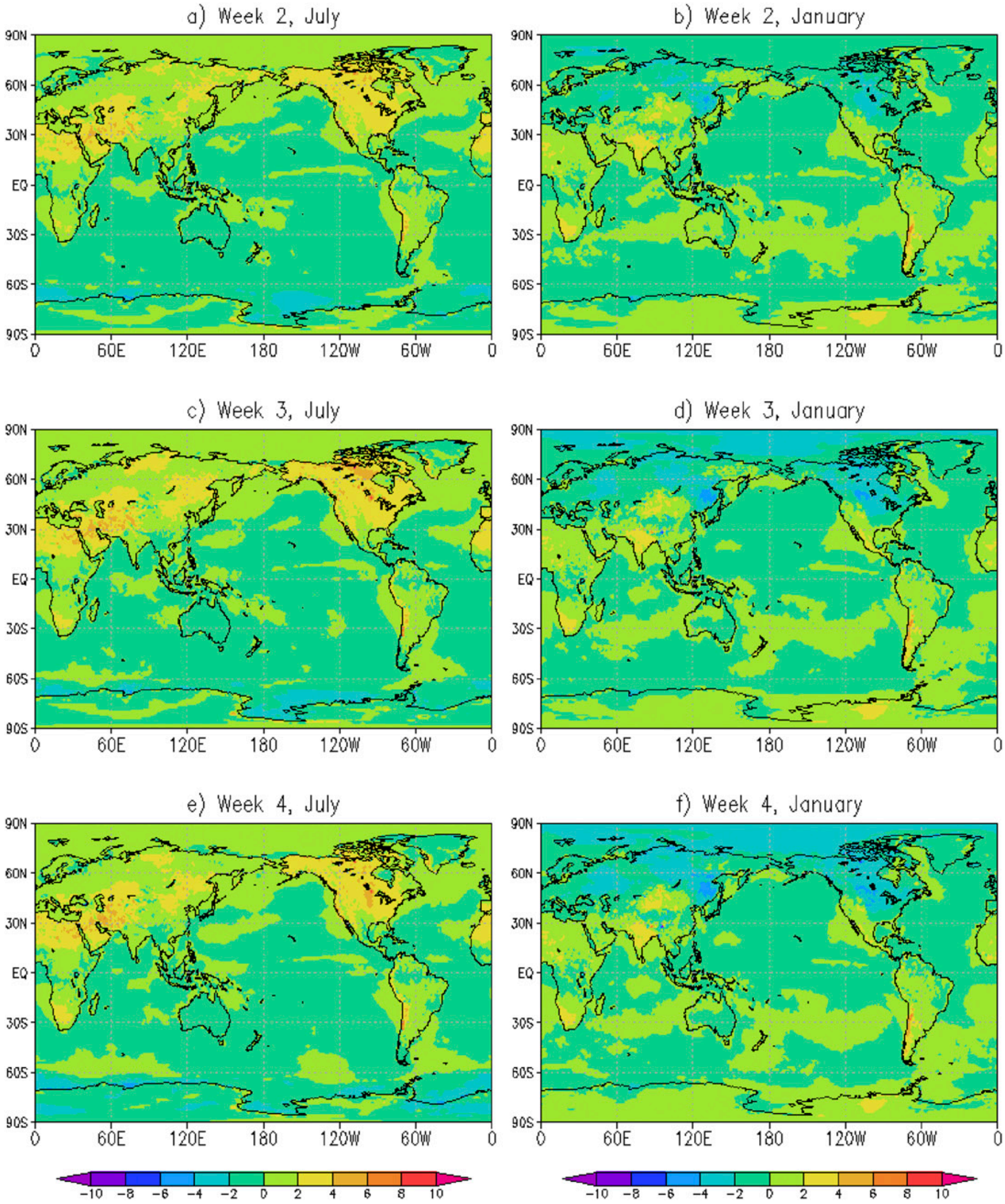

FIG. 5. Spatial distribution of 2-m temperature mean error (i.e., bias) calculated using 30-day running means over 18 years (1999–2016) for July during week (a) 2, (c) 3, and (e) 4 forecasts and January during week (b) 2, (d) 3, and (f) 4 forecasts.

the winter at the high latitudes have been linked to some larger system (i.e., polar vortex system) with more thermal or mechanical inertia (Song and Mapes 2012).

To assess the impact of using different initial conditions to produce 2-m temperature forecasts, the evolution of yearly time series is examined for 24-, 120-, and 480-h forecasts for the NH (land only) in Fig. 6. During the beginning of the model integration (24 h; Fig. 6a), the 2-m temperature forecast for the GDAS period (green curves) is systematically warmer than for the CFSR period (red curves) between July and October. The impact of using different initial conditions to produce 2-m temperature forecasts is lessened by the 120-h forecast (Fig. 6b) and eventually negligible by weeks 3 and 4 (480 h; Fig. 6c). This also implies that the observed difference in the weeks 3 and 4 bias (Figs. 2a,d) between the two analysis periods must mainly come from an inconsistent reference analysis.

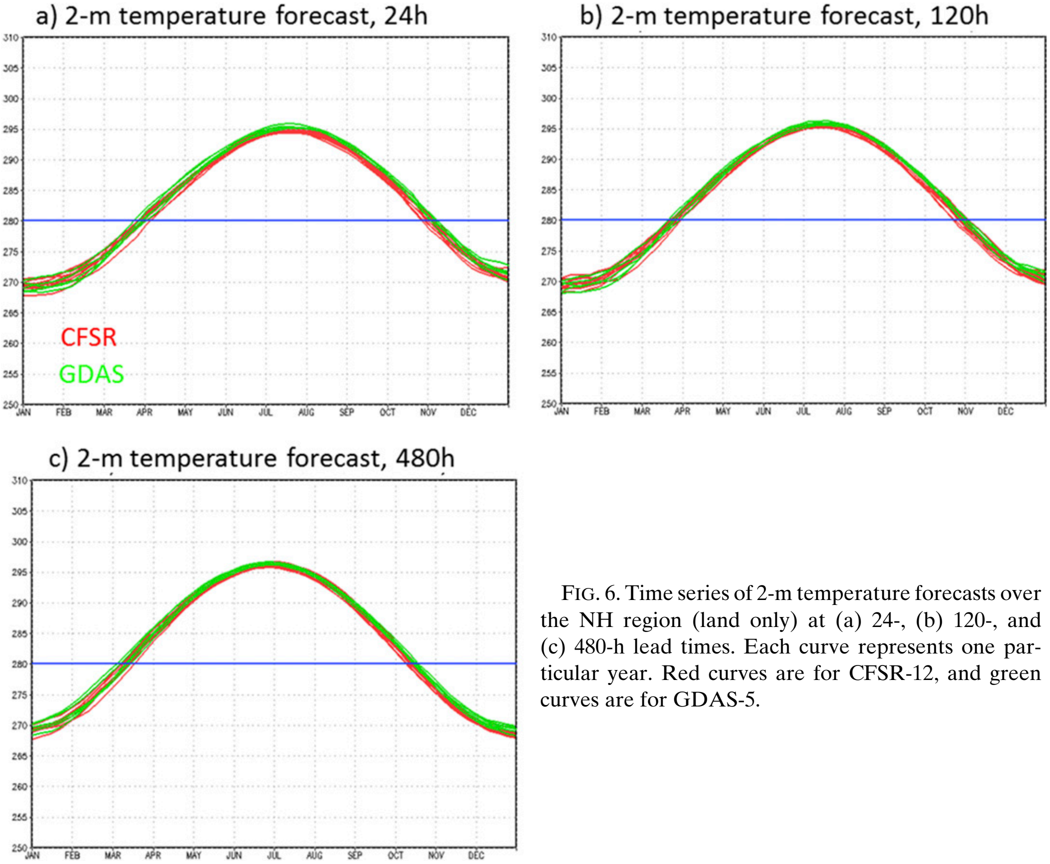

FIG. 6. Time series of 2-m temperature forecasts over the NH region (land only) at (a) 24-, (b) 120-, and (c) 480-h lead times. Each curve represents one particular year. Red curves are for CFSR-12, and green curves are for GDAS-5.

4. Bias correction for weeks 3 and 4

a. Methodology and analysis adjustment

The bias-corrected forecast F for each grid point (i, j) for weeks 3 and 4 (t_{w34}) is obtained by subtracting the weeks 3 and 4 average bias $b_{i,j}(t_{w34})$ at the same grid point from the raw forecast $f_{i,j}(t_{w34})$ using the following expression:

$$F_{i,j}(t_{w34}) = f_{i,j}(t_{w34}) - b_{i,j}(t_{w34}). \quad (1)$$

We can also apply week 2's average bias $b_{i,j}(t_{w2})$ to calibrate the weeks 3 and 4 forecast. The week 2 and weeks 3 and 4 biases are two 7-day (days 8–14) and one 14-day (days 15–28) bias at 0000 UTC to match a validated forecast period, respectively. For example, if we want to calculate the bias to calibrate the weeks 3 and 4 forecast that initialized at 0000 UTC 16 January 2016 (i.e., validation period of 0000 UTC 31 January–0000 UTC 13 February 2016), we can either calculate the difference between the weeks 3 and 4 forecast initialized at 0000 UTC 16 January 2016 and the analysis data for the corresponding validation period (i.e., 0000 UTC 31 January–0000 UTC 13 February 2016), or calculate the average of two week 2 biases, which validate during 0000 UTC 31 January–0000 UTC 6 February 2016 (initialized at 0000 UTC 23 January 2016) and 0000 UTC 7 February–0000 UTC 13 February 2016 (initialized at 0000 UTC 31 January 2016), respectively. The reason for the latter method is that when we use week 2 bias (which is a 7-day average) to calibrate the weeks 3 and 4 forecast, we need to make sure the sample size of the bias is consistent with the sample size of the forecast days. In other words, we do not want to use a 7-day bias to calibrate a 14-day forecast. The average biases of a certain validation date for a certain lead time is the average of biases during 1999–2015 within a time window of 31 days centered on that validation date. For example, the average bias of the weeks 3 and 4 forecast initialized on 16 January (validated on 31 January–13 February) is the averaged difference between the forecast and analysis for the validation period across 1999–2015. To test the sensitivity of the forecast skill to the number of training years, we also compare the calibrated forecast using the reforecast bias from the most recent 5- (2011–15), 10- (2006–15), and 17-yr (1999–2015) training datasets to evaluate the 2016 forecasts.

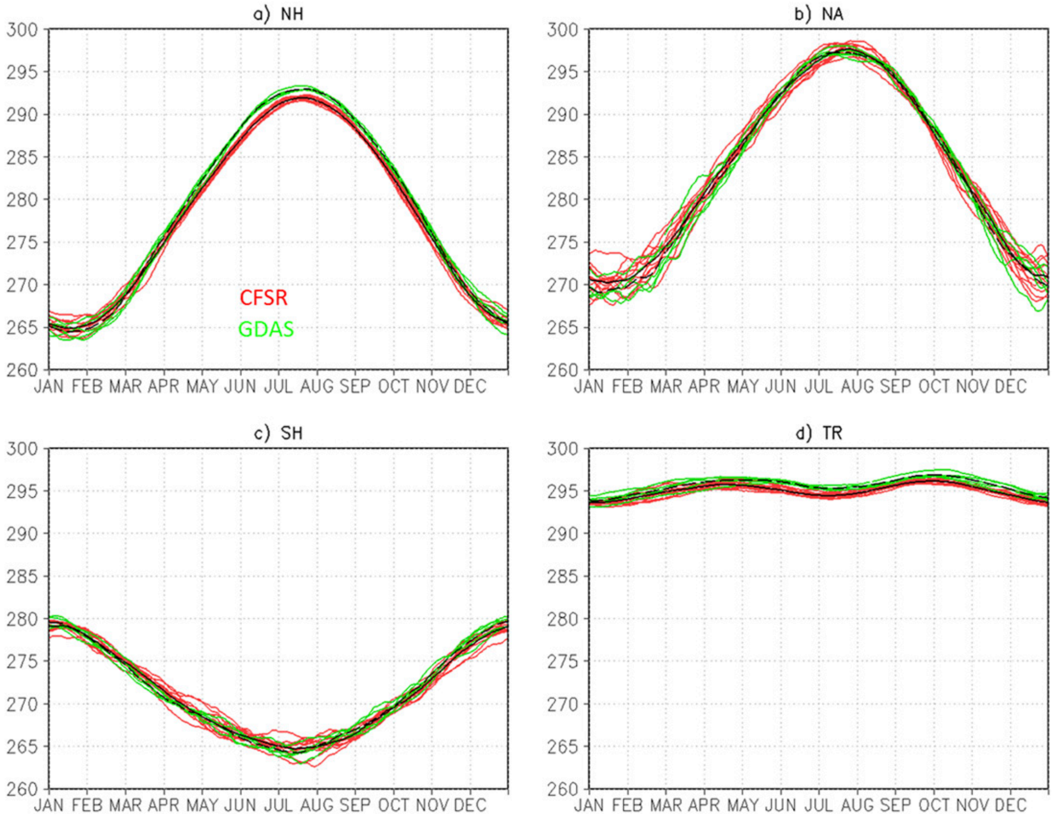

FIG. 7. The time series of year-by-year 2-m temperature analyses for (a) the NH, (b) NA, (c) the SH, and (d) the TR (land only). Red curves are for CFSR-12, and green curves are for GDAS-5. Black solid curves are the averages for the CFSR-12 period, and black dashed curves are the averages for the GDAS-5 period.

The calibration of the ensemble forecast system is evaluated via the RMSE (Zhu and Toth 2008) and rank probability skill score (RPSS; Wilks 2011). The RPSS is frequently used for evaluating the performance of probabilistic forecasts (Ou et al. 2016; Melhauser et al. 2016; Zhu et al. 2017), which measures the improvement of a multicategory forecast relative to a reference analysis. The higher the RPSS, the more skillfully the probabilistic system performs.

As noted in Fig. 2, there is a systematic difference in 2-m temperature bias between the CFSR-12 and GDAS-5 periods for the NH and TR domains, which most likely arose from inconsistent reference analyses. To test this hypothesis, we examine a land-only year-by-year analysis for four geographic domains [1) the NH, 2) NA, 3) the SH, and 4) the TR] in Fig. 7. As expected, the analysis difference between the two assimilation periods is evident for the NH and TR domains, with a maximum difference of more than 1°C (Figs. 7a,d). The black curves represent the averages for each analysis period. Note that both domains encompass the desert and arid regions of North Africa and the Middle East, the regions most affected by the 2011 GFS upgrade (Zheng et al. 2012). Figures 7a and 7d also reveal that the 2-m temperatures are systematically warmer during the GDAS-5 period than the CFSR-12 period for the NH and TR. A warmer reference analysis during the GDAS-5 period (Fig. 7a) induces a smaller forecast warm bias (Fig. 2a), assuming that the forecast is less dependent on the initial analysis for weeks 3 and 4.

To make a consistent reference analysis from 1999 to 2015, it is necessary to adjust the early CFSR analysis $a_{i,j}$. We first calculate the CFSR-12 and GDAS-5 averaged analyses for each grid point (i,j) ($a_{i,j}^{12\mathrm{y}}$ and $a_{i,j}^{5\mathrm{y}}$, respectively) and then apply the difference ($a'_{i,j}$) to the first 12-yr analysis as follows:

$$a'_{i,j} = a_{i,j}^{12\mathrm{y}} - a_{i,j}^{5\mathrm{y}} \quad \text{and} \tag{2}$$

$$a_{i,j}^{\mathrm{adj}} = a_{i,j} - a'_{i,j}\,. \tag{3}$$

Note that an "analysis adjustment" $a_{i,j}^{\mathrm{adj}}$ is based on our assumption (Fig. 6c) and previous work (Zhu 2005) that the weeks 3 and 4 forecast errors have a negligible impact from the initial conditions. The climate trend cannot be well estimated in this study because a full

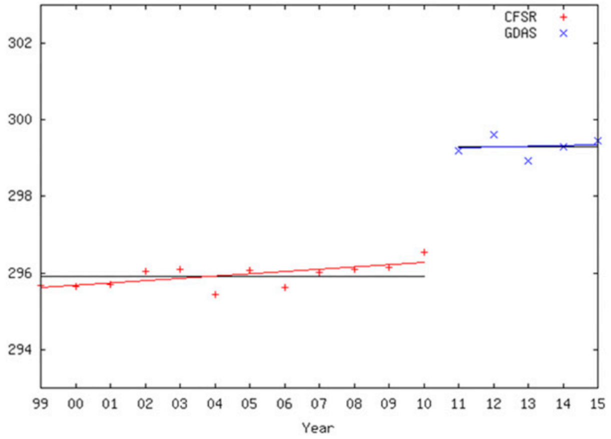

FIG. 8. Domain-averaged 2-m temperature analyses for the North Africa and Middle East regions during 1999–2015. Red plus signs (+) represent individual years during the CFSR-12 period and blue crosses (x) represent individual years during the GDAS-5 period. The red (blue) solid line represents the line of best fit for the CFSR-12 (GDAS-5) period. Black lines are the averaged values for the corresponding two periods.

consistent set of the CFSR analysis or GDAS analysis for the studied period (1999–2015) is not available. However, a comparison of domain-averaged 2-m temperatures for the North African and Middle East regions during 1999–2015 does illustrate that the considerable differences (~3.4°C) between the two analysis periods is mainly caused from the inconsistent analysis. This is indicated by a sharp increase in 2-m temperature in 2011 (Fig. 8). In contrast, the actual trend in 2-m temperature during the CFSR-12 (red line) or GDAS-5 (blue line) analyses is relatively minor.

To demonstrate the consistency of forecast errors after analysis adjustment, domain-averaged 2-m temperature errors (i.e., bias), without and with analysis adjustment, are presented in Fig. 9. The analysis adjustment mitigates the inconsistency of 2-m temperature bias for the NH (Figs. 9a,b) and TR (Figs. 9c,d). The adjustment has a small impact for the SH and NA (not shown). In the next section, we will examine the bias correction (i.e., calibration) without and with analysis adjustment.

b. Calibrating the 2016 forecasts using the 17-yr training dataset

The week 2 and weeks 3 and 4 biases for 0000 UTC without and with analysis adjustment were used to calibrate the weeks 3 and 4 forecasts. The forecast RMSE (Fig. 10a) and RPSS (Fig. 10b) for 0000 UTC

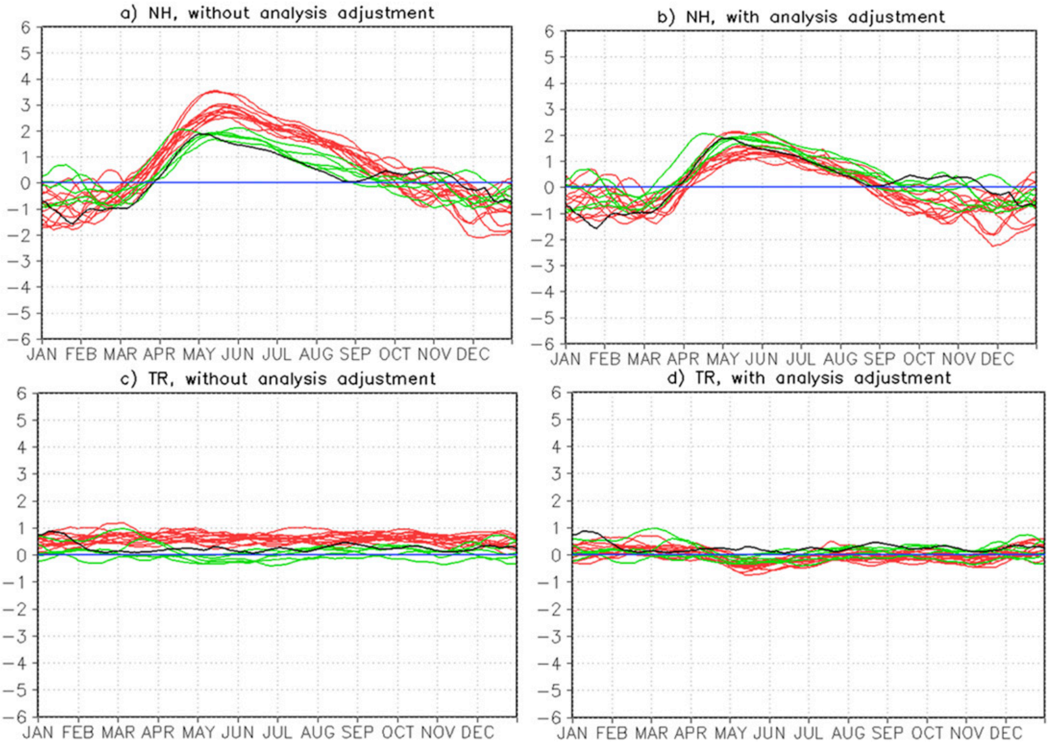

FIG. 9. Time series of 2-m temperature forecast errors (i.e., biases) during weeks 3 and 4 for the (a),(b) NH and (c),(d) TR domains, without and with analysis adjustments, respectively. Each curve represents one particular year. Red curves indicate the errors for CFSR-12, and green curves indicate the errors for GDAS-5. Black lines indicate errors for 2016.

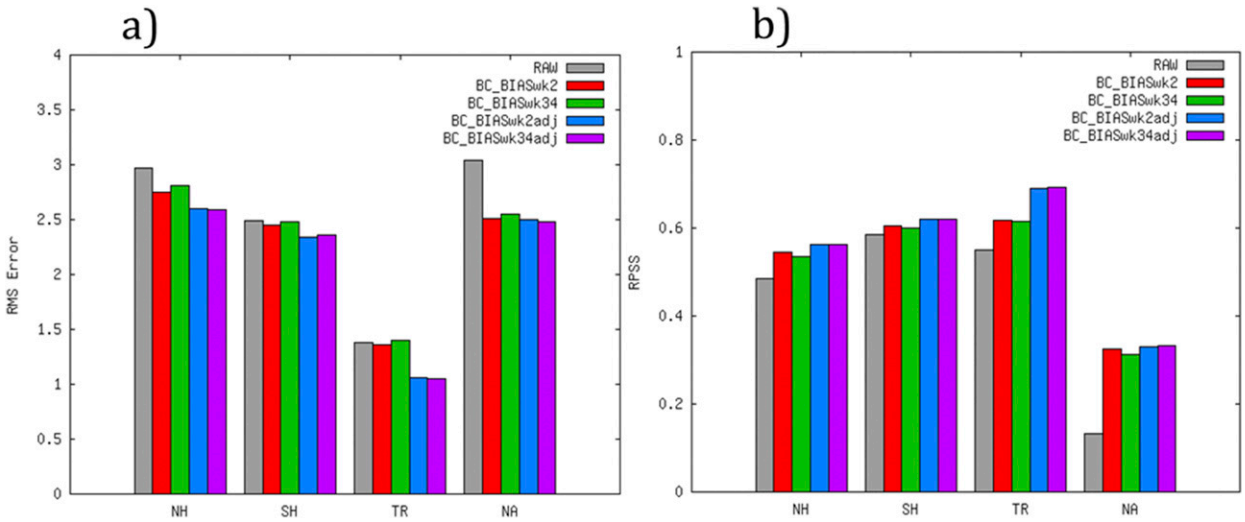

FIG. 10. (a) RMSE and (b) RPSS of 2-m temperature forecasts during weeks 3 and 4 in 2016 (land only), averaged over NA and the NH, SH, and TR for the raw (gray bar) and four bias-corrected forecasts: BC_BIASwk2 (red), BC_BIASwk34 (green), BC_BIASwk2adj (blue), and BC_BIASwk34adj (purple). The BC_BIASwk2 (red) and BC_BIASwk34 (green) forecasts denote the calibration using week 2 and weeks 3 and 4 biases without analysis adjustment, whereas BC_BIASwk2adj (blue) and BC_BIASwk34adj (purple) denote the calibration using the week 2 and weeks 3 and 4 biases with analysis adjustment.

are improved after bias correction and analysis adjustment for all four domains. Analysis adjustment does an excellent job for the TR and NH (Fig. 10a), reducing RMSEs relative to bias correction alone by ~0.3°C (or 20%) and ~0.25°C (or 7%), respectively. RMSE over NA is reduced by up to ~0.6°C (or 20%) through the bias correction alone, with a slight additional improvement following analysis adjustment. Figure 10 also reveals that forecast skill is very similar whether the week 2 or weeks 3 and 4 bias is used for the calibration. This finding would suggest for 2-m temperature that we could use the bias from week 2 to calibrate the weeks 3 and 4 forecasts, which would optimize the use of computer resources without sacrificing the effectiveness of the calibration. Although bias correction produces the most substantial improvement for NA, it still has the lowest RPSS (Fig. 10b) even though it has a similar RMSE to the NH (Fig. 10a). This could be partially due to its large bias variance (Fig. 2), which makes 2-m temperature forecasts less predictable compared to the other domains.

To find out the seasonal dependence of the bias corrections on the 2-m temperature forecast, we show the time series of RMSE (Fig. 11) and RPSS (Fig. 12) for the raw and bias-corrected weeks 3 and 4 forecasts without and with analysis adjustment. The largest improvement in RMSE occurs over NA during the warm season (Fig. 11b) primarily from the bias correction. The RPSS increases from a near-zero value to ~0.4 over NA during this period (Fig. 12b), while RMSE is substantially reduced with a maximum reduction of up to ~50% in July (Fig. 11b). A considerable skill improvement due to the analysis adjustment is shown for the TR (Fig. 11d and Fig. 12d) throughout most of the year.

The distributions of RPSS for the land-only raw (Fig. 13a) and bias-corrected and analysis-adjusted (Fig. 13b) weeks 3 and 4 forecasts during 2016 are presented. There is negative skill relative to the climatology for the raw forecast over a considerable portion of the CONUS (Fig. 13a). Prediction of 2-m temperatures is shown to be extremely challenging over the Great Plains, consistent with the findings in Klein et al. (2006). Both our study and Klein et al. (2006) reveal the large warm bias over the Great Plains for the warm season, which may partly account for the corresponding low RPSS. The calibrated forecast (Fig. 13b) produces much higher forecast skill over the entire CONUS domain. Substantial improvements are detected over the Great Plains where the maximum increase in RPSS reaches ~0.6 [from ~ −0.45 for the raw forecast (Fig. 13a) to ~0.15 for the calibrated forecast (Fig. 13b)] near South Dakota.

c. *Skill sensitivity to number of training years*

The sensitivity of forecast skill to the number of training years has been studied by Hamill et al. (2004), Guan et al. (2015), and Ou et al. (2016). Using the first-generation GEFS reforecast dataset, Hamill et al. (2004) demonstrated that there was a significant increase in skill from 2 to 5 years of training data for week 2 surface temperature, but only showed small incremental increases by 10–12 years. The sensitivity experiments of

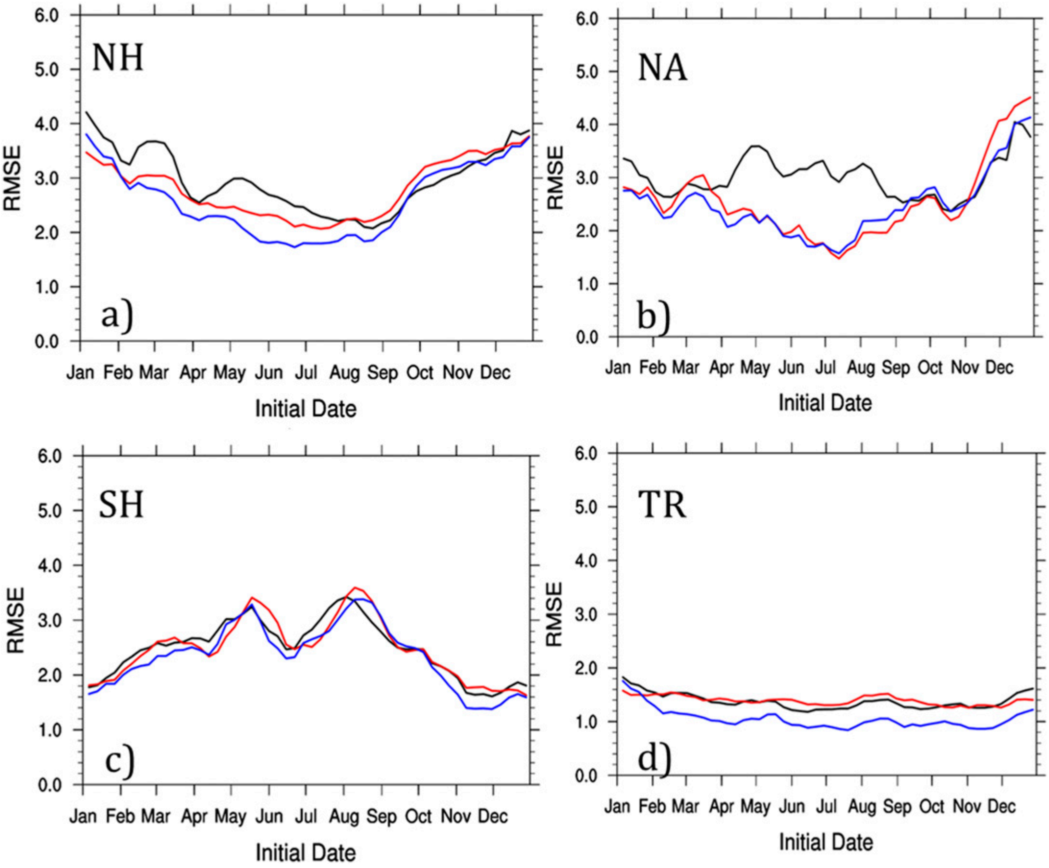

FIG. 11. RMSE of 2-m temperature forecasts during weeks 3 and 4 in 2016 (land only) averaged over (a) the NH, (b) NA, (c) the SH, and (d) the TR for the raw forecast (black) and two bias-corrected forecasts: BC_BIASwk34 (red) and BC_BIASwk34adj (blue). The BC_BIASwk34 (red) and BC_BIASwk34adj (blue) forecasts denote the bias-corrected forecasts without and with analysis adjustment, respectively.

Guan et al. (2015) with more skillful GEFSv10 reforecast data (Hamill et al. 2013) reveal that the improvement from using a 5-yr training period is almost equivalent to the improvement seen when using a 10- or 25-yr training period for lead times up to 16 days. Using the same dataset, Ou et al. (2016) showed an 18-yr training period is desirable for high quality week 2 calibration over the CONUS.

To test the sensitivity of the weeks 3 and 4 forecast skill to the number of training years, we calibrate the 2016 forecast using the 5-yr (2011–15), 10-yr (2006–15), and 17-yr (1999–2015) training datasets. These specific training years were chosen because the forecast skill in predicting 2-m temperature displays a steady increase from the 5- to 10-yr training periods and then nears saturation (Hamill et al. 2004; Ou et al. 2016; Guan et al. 2015). Figure 14 shows that increasing the number of training years from 5 to 10 years leads to a gain in skill of 0.016 (or ~5%), whereas further increasing to 17 years does not yield additional improvement. This result indicates that a 10-yr training period should be an optimal requirement for the 2-m temperature calibration during weeks 3 and 4 of the NCEP GEFS SubX version. Our optimal training period (10 years) for the weeks 3 and 4 forecast is similar to the 10–12-yr training period for the week 2 forecast estimated by Hamill et al. (2004), but less than the 18-yr training period in Ou et al. (2016). This difference could be partially attributed to the differences in forecast lead time, model version, and verification period, as pointed out in Ou et al. (2016).

5. Summary and conclusions

NCEP EMC generated an 18-yr subseasonal reforecast dataset to support the CPC's operational mission. The GEFS-SubX version was run weekly, initialized at 0000 UTC every Wednesday, with 11 members using the CFSR-12 and GDAS-5 models as the initial analyses. Using this dataset, we explore the initial

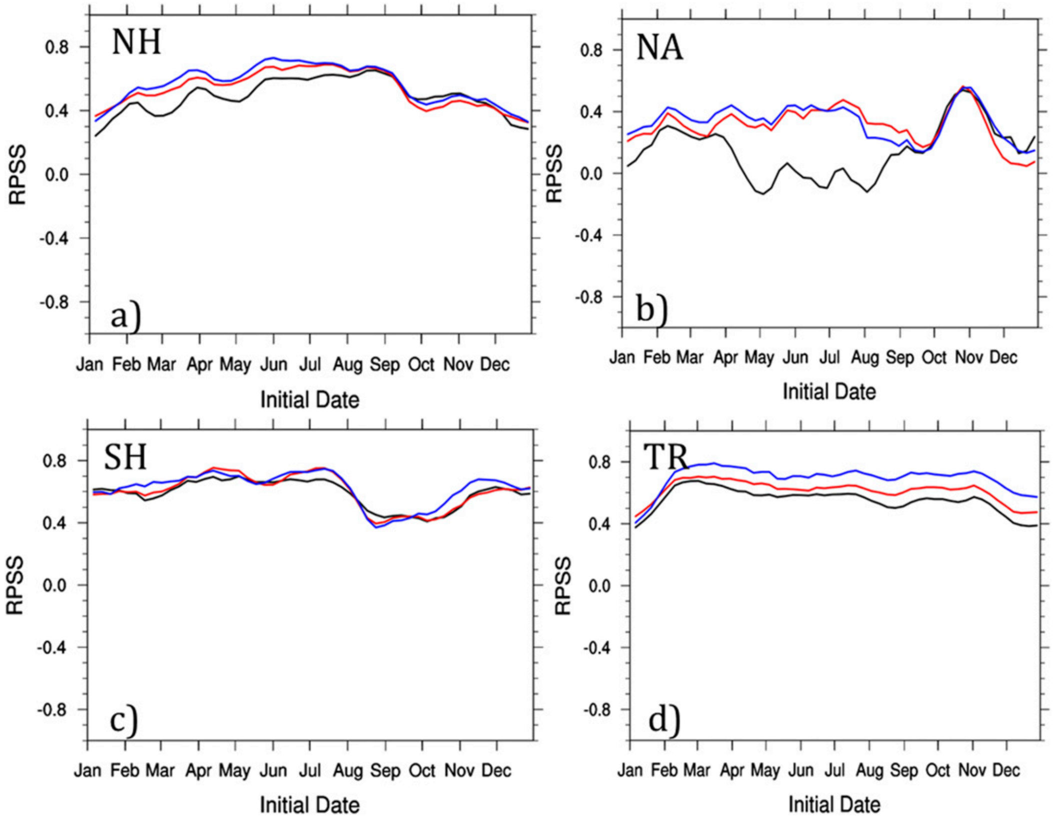

FIG. 12. As in Fig. 11, but for RPSS.

analyses inconsistencies, analyses adjustments, and bias characteristics of 2-m temperature during weeks 3 and 4 of the reforecast period. We subsequently apply the 17-yr (1999–2015) bias to calibrate the weeks 3 and 4 forecasts of 2016.

The main conclusions of the study are as follows:

1) The forecast of 2-m temperature is strongly biased over NA and the NH with a warm bias during the warm season. In boreal winter, there is large interannual variability in the 2-m temperature bias over NA. Therefore, it is a challenge to find out the corresponding model systematic errors for 2-m temperature.

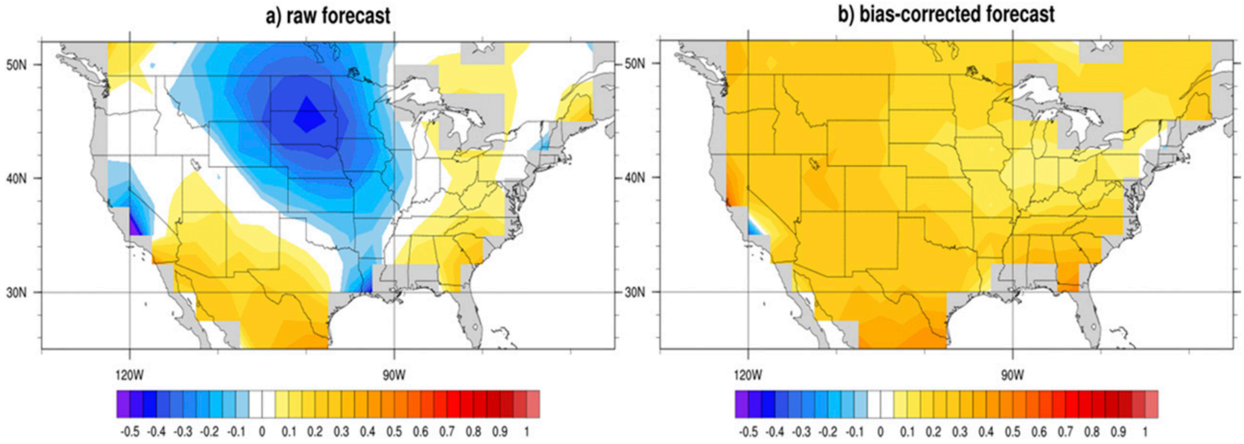

FIG. 13. RPSS of 2-m temperature forecasts during weeks 3 and 4 over the CONUS in 2016 for the (a) raw and (b) bias-corrected and analysis-adjusted forecasts.

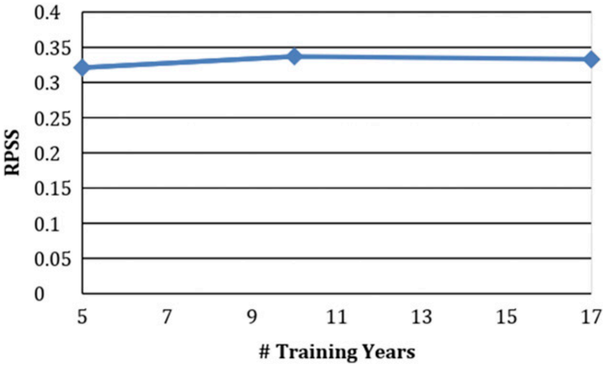

FIG. 14. RPSS of 2-m temperature forecasts during weeks 3 and 4 as a function of the number of training years. Skill scores represent the average score across the CONUS in 2016.

2) Forecast errors quickly grow within the first 10 days of the forecast and gradually saturate by weeks 3 and 4. The error of the day-11 forecast (or the middle of week 2) for NA (the NH) reaches about 88.6% (86.6%) of its saturated value. The impact of the initial conditions on forecast skill is negligible by weeks 3 and 4.
3) A consistent analysis is important for generating reforecasts and real-time forecasts. Analysis adjustment is an alternative method to making the bias characteristics more consistent between the CFSR-12 and GDAS-5 periods. An adjusted analysis can be considered as a backup solution when a reanalysis (or reference) spanning the entire period of interest is not available.
4) Bias correction is important in reducing systematic error. An increase in forecast skill following bias correction is observed for all four domains (NH, SH, TR, and NA). Maximum benefit was found for NA during the warm season. Calibration using the week 2 bias gives very similar skill to using the weeks 3 and 4 bias, suggesting that the week 2 bias could be used to correct the weeks 3 and 4 forecast. This practice could help save computational resources and storage in operations.
5) The 2-m temperature calibrations during weeks 3 and 4 have been performed using 5-, 10-, and 17-yr sample datasets with the aim of determining an optimal training period. Our results demonstrate a 10-yr training period is sufficient to obtain a more skillful forecast of 2-m temperatures during 2016, if the reference analyses are consistent.

The current study demonstrates the importance of using reforecast information to improve the weeks 3 and 4 forecast skill for 2-m temperatures by evaluating the analysis difference, as well as the temporal and spatial distributions of the forecast errors. Analysis of the bias characteristics of weeks 3 and 4 precipitation forecasts and its calibration are currently being performed. Since 1 July 2017, the NCEP GEFS SubX version has generated 35-day forecasts in real time, once per week (every Wednesday at 0000 UTC). In the future, we will continue generating the calibration statistics with incoming real-time SubX forecasts and further examine the effectiveness and robustness of the proposed calibration method with additional data. It is also noted that the current bias correction method is less effectual in NA for the winter season than the summer season. This is due to large interannual variability of 2-m temperature bias, likely associated with large variability of snow cover. It is expected that 2-m temperature bias characteristics are quite different with and without snow cover. Therefore, there is a possibility of improving the calibration method in boreal winter by generating bias climatology with and without snow cover and then performing a bias correction based on snow existence.

Acknowledgments. The authors thank Dr. Weizhong Zheng for providing valuable discussions on the GFS land model. Dr. Glenn White and Alicia Bentley are thanked for their careful reviews of the manuscript. This study is partially supported through NWS OSTI and NOAA's Climate Program Office (CPO) Modeling, Analysis, Predictions, and Projections (MAPP) program.

REFERENCES

Bishop, C. H., and K. T. Shanley, 2008: Bayesian model averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon. Wea. Rev.*, **136**, 4641–4652, https://doi.org/10.1175/2008MWR2565.1.

Buizza, R., M. Milleer, and T. Palmer, 1999: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887–2908, https://doi.org/10.1002/qj.49712556006.

Candille, G., 2009: The multiensemble approach: The NAEFS example. *Mon. Wea. Rev.*, **137**, 1655–1665, https://doi.org/10.1175/2008MWR2682.1.

Chai, T., and R. R. Draxler, 2014: Root mean square error (RMSE) or mean absolute error (MAE)?—Argument against avoiding RMSE in the literature. *Geosci. Model Dev.*, **7**, 1247–1250, https://doi.org/10.5194/gmd-7-1247-2014.

Cheng, W. Y. Y., and W. J. Steenburgh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Wea. Forecasting*, **22**, 1304–1318, https://doi.org/10.1175/2007WAF2006084.1.

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.

Frei, A., and D. A. Robinson, 1999: Northern Hemisphere snow extent: Regional variability 1972–1994. *Int. J. Climatol.*, **19**, 1535–1560, https://doi.org/10.1002/(SICI)1097-0088(19991130)19:14<1535::AID-JOC438>3.0.CO;2-J.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Guan, H., and Y. Zhu, 2017: Development of verification methodology for extreme weather forecasts. *Wea. Forecasting*, **32**, 479–491, https://doi.org/10.1175/WAF-D-16-0123.1.

——, B. Cui, and Y. Zhu, 2015: Improvement of statistical postprocessing using GEFS reforecast information. *Wea. Forecasting*, **30**, 841–854, https://doi.org/10.1175/WAF-D-14-00126.1.

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

——, 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, https://doi.org/10.1175/2007MWR2411.1.

——, and Coauthors, 2013: NOAA's second-generation global medium-range ensemble reforecast dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, https://doi.org/10.1175/BAMS-D-12-00014.1.

Han, J., W. Wang, Y. C. Kwon, S.-Y. Hong, V. Tallapragada, and F. Yang, 2017: Updates in the NCEP GFS cumulus convection schemes with scale and aerosol awareness. *Wea. Forecasting*, **32**, 2005–2017, https://doi.org/10.1175/WAF-D-17-0046.1.

Johnson, N. C., D. C. Collins, S. B. Feldstein, M. L. L'Heureux, and E. E. Riddle, 2014: Skillful wintertime North American temperature forecasts out to 4 weeks based on the state of ENSO and the MJO. *Wea. Forecasting*, **29**, 23–38, https://doi.org/10.1175/WAF-D-13-00102.1.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Kazakova, E. and I. Rozinkina, 2011: Testing of snow parameterization schemes in COSMO-Ru: Analysis and results. *COSMO Newsletter*, No. 11, 41–51, http://www.cosmo-model.org/content/model/documentation/newsLetters/newsLetter11/2_kazakova.pdf.

Klein, S. A., X. Jiang, J. Boyle, S. Malyshev, and S. Xie, 2006: Diagnosis of the summertime warm and dry bias over the U.S. southern Great Plains in the GFDL Climate Model using a weather forecasting approach. *Geophys. Res. Lett.*, **33**, L18805, https://doi.org/10.1029/2006GL027567.

Klingaman, N. P., B. Hanson, and D. J. Leathers, 2008: A teleconnection between forced Great Plains snow cover and European winter climate. *J. Climate*, **21**, 2466–2483, https://doi.org/10.1175/2007JCLI1672.1.

Lavaysse, C., M. Carrera, S. Bélair, N. Gagnon, R. Frenette, M. Charron, and M. K. Yau, 2013: Impact of surface parameter uncertainties with the Canadian Regional Ensemble Prediction System. *Mon. Wea. Rev.*, **141**, 1506–1526, https://doi.org/10.1175/MWR-D-11-00354.1.

Li, W., and Coauthors, 2018: Evaluating the MJO forecast skill from different configurations of NCEP GEFS extended forecast. *Climate Dyn.*, https://doi.org/10.1007/s00382-018-4423-9, in press.

Melhauser, C., W. Li, Y. Zhu, X. Zhou, M. Pena, and D. Hou, 2016: Exploring the impact of SST on the extended range NCEP Global Ensemble Forecast System. *41st Annual Climate Diagnostics and Prediction Workshop*, Orono, ME, NOAA/NWS, 30–34, http://www.nws.noaa.gov/ost/climate/STIP/41CDPW/41cdpw-CMelhauser.pdf.

NOAA, 2017: Service Change Notice 1767 update. NOAA/NWS, http://www.nws.noaa.gov/os/notification/scn17-67gfsupgradeaab.htm.

Ou, M., M. Charles, and D. Collins, 2016: Sensitivity of calibrated week-2 probabilistic forecast skill to reforecast sampling of the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 1093–1107, https://doi.org/10.1175/WAF-D-15-0166.1.

Palmer, T. N., R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G. Shutts, M. Steinheimer, and A. Weisheimer, 2009: Stochastic parametrization and model uncertainty. ECMWF Tech. Memo. 598, 42 pp., http://www.ecmwf.int/publications/.

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, https://doi.org/10.1175/MWR2906.1.

Robinson, D. A., 1996: Evaluating snow cover over Northern Hemisphere lands using satellite and in situ observations. *Proc. 53rd Eastern Snow Conf.*, Williamsburg, VA, Eastern Snow Conference, 13–19.

——, and A. Frei, 2000: Seasonal variability of Northern Hemisphere snow extent using visible satellite data. *Prof. Geogr.*, **52**, 307–315, https://doi.org/10.1111/0033-0124.00226.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus*, **55A**, 16–30, https://doi.org/10.1034/j.1600-0870.2003.201378.x.

Saha, S., and Coauthors, 2010: The NCEP Climate Forecast System Reanalysis. *Bull. Amer. Meteor. Soc.*, **91**, 1015–1057, https://doi.org/10.1175/2010BAMS3001.1.

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quart. J. Roy. Meteor. Soc.*, **131**, 3079–3102, https://doi.org/10.1256/qj.04.106.

——, and T. N. Palmer, 2004: The use of high-resolution numerical simulations of tropical circulation to calibrate stochastic physics schemes. *Proc. Workshop on Simulation and Prediction of Intra-Seasonal Variability with Emphasis on the MJO*, Reading, United Kingdom, ECMWF, 83–102, https://www.ecmwf.int/en/learning/workshops-and-seminars/past-workshops/2003-simulation-prediction-intra-seasonal-variability.

Song, S. W., and B. Mapes, 2012: Interpretations of systematic errors in the NCEP Climate Forecast System at lead times of 2, 4, 8, ..., 256 days. *J. Adv. Model. Earth Syst.*, **4**, M09002, https://doi.org/10.1029/2011MS000094.

Tompkins, A. M., and J. Berner, 2008: A stochastic convective approach to account for model uncertainty due to unresolved humidity variability. *J. Geophys. Res.*, **113**, D18101, https://doi.org/10.1029/2007JD009284.

Vitart, F., 2009: Impact of the Madden–Julian oscillation on tropical storms and risk of landfall in the ECMWF forecast system. *Geophys. Res. Lett.*, **36**, L15802, https://doi.org/10.1029/2009GL039089.

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.*, **131**, 965–986, https://doi.org/10.1256/qj.04.120.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, https://doi.org/10.1111/j.1600-0870.2007.00273.x.

Wilks, D. S., 2011: *Statistical Methods in the Atmospheric Sciences*. 3rd ed. Elsevier, 676 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, **135**, 2379–2390, https://doi.org/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, https://doi.org/10.1175/MWR3347.1.

Yuan, H., X. Gao, S. L. Mullen, S. Sorooshian, J. Du, and H. H. Juang, 2007: Calibration of probabilistic quantitative precipitation forecasts with an artificial neural network. *Wea. Forecasting*, **22**, 1287–1303, https://doi.org/10.1175/2007WAF2006114.1.

Zheng, W., H. Wei, Z. Wang, X. Zeng, J. Meng, M. Ek, K. Mitchell, and J. Derber, 2012: Improvement of daytime land surface skin temperature over arid regions in the NCEP GFS model and its impact on satellite data assimilation. *J. Geophys. Res.*, **117**, D06117, https://doi.org/10.1029/2011JD015901.

Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: Comparison of the ensemble transform and the ensemble Kalman filter in the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 2058–2074.

——, ——, ——, Y. Luo, J. Peng, and D. Wobus, 2017: The NCEP Global Ensemble Forecast System with the EnKF initialization. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.

Zhu, Y., 2005: Ensemble forecast: A new approach to uncertainty and predictability. *Adv. Atmos. Sci.*, **22**, 781–788, https://doi.org/10.1007/BF02918678.

——, and Z. Toth, 2008: Ensemble based probabilistic forecast verification. *19th Conf. on Probability and Statistics*, New Orleans, LA, Amer. Meteor. Soc., 2.2, https://ams.confex.com/ams/88Annual/webprogram/Paper131645.html.

——, X. Zhou, M. Pena, W. Li, C. Melhauser, and D. Hou, 2017: Impact of sea surface temperature forcing on weeks 3 and 4 forecast skill in the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **32**, 2159–2173, https://doi.org/10.1175/WAF-D-17-0093.1.

——, and Coauthors, 2018: Toward the improvement of subseasonal prediction in the NCEP Global Ensemble Forecast System (GEFS). *J. Geophys. Res. Atmos.*, **123**, 6732–6745, https://doi.org/10.1029/2018JD028506.