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

Hong Guan[1], Yuejian Zhu[2], Eric Sinsky[3], Wei Li[3], Xiaqiong Zhou[3], Dingchen Hou[2],

Christopher Melhauser[3], Richard Wobus[3]

Version 2.0

Update: 01/30/2018

In preparation to be submitted to Monthly Weather Review

*Corresponding Author: Dr. Hong Guan, Email: Hong.Guan@noaa.gov, Environmental Modeling Center/NCEP/NOAA, 5830 University Research Court, College Park, MD 20740

Abstract

EMC/NCEP generated an 18-year (1999-2016) subseasonal (weeks 3&4) reforecast to support the CPC's operational mission. The SubX version of the Global Ensemble Forecast System was run weekly initialized at 0000 UTC with 11 members. The Climate Forecast System Reanalysis (CFSR) and Global Data Assimilation System (GDAS) were served as an initial analysis for 1999-2010 and 2011-2016. The analysis of 2-m temperature error characteristics demonstrated that the model has a strong warm bias in the Northern Hemisphere (NH) and North America (NA) warm season. During the winter, the 2-m temperature errors in NA exhibit a large inter-annual and intra-seasonal variability. For NA and the NH, weeks 3&4 errors are mostly saturated with a negligible impact of initial condition to forecast and week 2 errors (day-11) also reach ~88.6% and 86.6% of their saturated levels.

In this work, 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, Southern Hemisphere and Tropics with a maximum benefit for the NA 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 skills. The calibration using week-2 bias gives a very similar skill to using weeks 3&4 bias, promising the feasibility of using week-2 bias to calibrate weeks 3&4's forecast. Our results also demonstrate 10-yr reforecasts are an optimal training period. This is particularly beneficial considering limited computation resource.

1. Introduction

To provide a seamless numerical guidance to a broad range of users and partners, NOAA is extending the service from a weather forecast (week 1) and extended forecast (week 2) to sub-seasonal (weeks 3&4) forecast through the Next Generation Global Prediction System (NGGPS) project. The lack of memory of the atmospheric initial analysis as well as the effects of the atmosphere-land and ocean-sea-ice interactions, which benefits weather forecast and seasonal and longer timescale forecast, respectively, arises a particular challenge to the sub-seasonal forecasts (Johnson et al. 2014 and Li et al. 2018). On the sub-seasonal timescale, the numerical model is a major driver for forecast error and skill. Thus, the improvement in the dynamical forecast system is a critical aspect of advancing the sub-seasonal forecast skill. In addition to it, the statistical method (i.e. post-processing technique) is another important aspect as it improves the forecast quality after calibration thus could improve the forecast skill. The post-processing is especially important for sub-seasonal time scale due to larger forecast error existed in this time scale.

Regarding the potential improvement of forecast skill in dynamical forecast system, recent studies demonstrate that sea surface temperature (SST) forcing (Zhu et al. 2017), the updated convection parameterization scheme (Vitart 2009; Zhu et al. 2018; and Li et al. 2018) and new stochastic physics (Zhu et al. 2018; and Li et al. 2018) significantly improve Madden-Julian Oscillation (MJO) forecast skill and 500 hPa geopotential height. Although the forecast skill for the source of predictability of the sub-seasonal scale in tropics and the large-scale circulation raw forecast is promising, the weeks 3&4 forecast of near-surface variables is still challenging. For example, the improvement in forecast skill for 2-m temperature and accumulated precipitation raw forecast is only marginal (Zhu et al. 2018). This suggests

developing a suitable post-processing technique to calibrate the raw forecast and further improve the forecast skill of near-surface variables is especially important on a SubX (Subseasonal Experiments) timescale. Previous studies (Hamill et al 2004; 2008; Cui et al. 2012; Guan et al. 2015; Guan and Zhu 2017; and Ou et al. 2016) reveal the importance of a hindcast (or reforecast) in extreme weather forecasts or bias correction on week 1 or week 2 timescales. Thus, the hybrid decaying and reforecast bias-correction method (Guan et al. 2015) is being operationally applied into the North American Ensemble Forecast System (NAEFS) (Candille 2009) in order to improve 1 to 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 analysis results is crucial to further develop an inexpensive reforecast configuration and an effective bias-correction method in operations. It is known that creating a multi-year reanalysis and reforecast dataset requires considerable computational and human resources. It is desirable to produce a high-quality of forecast but using less resource. To reach this goal, we determine the time scale when 2-m temperature error reaches a saturated level and then address whether the week 2 2-m temperature bias can be used to calibrate weeks 3&4 forecasts. We also explore the impact of an inconsistent initial analysis on weeks 3&4 forecast and find out a backup solution (or analysis adjustment) when a consistent reanalysis dataset is not available.

We first describe the forecast system and datasets in section 2. Then, we explore the temporal and spatial distributions of 2-m temperature bias and error saturation in section 3. In

section 4, we develop weeks 3&4 bias correction methods, including analysis adjustment and calibration sensitivity tests. Summary and conclusion are given in section 5.

2. Forecast system and data

In May 2017, the National Centers for Environmental Prediction (NCEP) Environmental Modeling Center (EMC) generated an 18-year (1999 - 2016) reforecast dataset to support the NCEP Climate Prediction Center (CPC)'s operational mission. With the exception of having a smaller ensemble size (1 control member and 10 perturbed members for reforecasts vs 1 control member and 20 members for real-time), the Global Ensemble Forecast System (GEFS) is essentially 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 having a new set of perturbed physics schemes, an updated scale-aware convection scheme (Han et al. 2017), and bias-corrected CFSv2 forecast sea surface temperature (SST). 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 rest of the lead days. The dataset used here was bilinear interpolated onto 1°x1° latitude and longitude grids from the model native resolution. Similar to Zhu et al. (2018), the forecast skills are defined relative to NCEP/NCAR 40 year reanalysis (Kalnay et al. 1996) climatology.

Ideally, creating a full set of consistent reanalysis data, including observations and model, is an important part of reforecast process. A reforecast with an initial condition from a different analysis system would bring a different bias to the forecast. However, the frequent updating of the model, satellite data, or analysis system makes running a reanalysis impractical in operations because generating a multi-year reanalysis is computationally expensive. As illustrated in Fig.1,

here we use the two major sets of existing analysis data because there is not a consistent 18 year reanalysis available. The Climate Forecast System Reanalysis (CFSR, Saha et al. 2010) and NCEP operational Global Data Assimilation System (GDAS) (varied generations of hybrid GSI (Gridded Statistical Interpolation)/EnKF (Ensemble Kalman Filter)) analyses were used as model initial conditions for the time period of Jan.1999 – 2010 and 2011 -2016, respectively. The analyses data are consistent prior to 2011 and then varied with the GFS/GSI/EnKF upgrades after merging to the GDAS period. It should also be mentioned that using a new surface roughness formulation in the Global Forecast System (GFS) upgrade of May 11, 2011 (Zheng et al. 2012) lead to a significant change in 2-meter temperature analysis and forecasts for the arid areas or dust areas. It is expected that the impact of initial conditions on the forecast becomes less important at longer lead times. The current study also provides an opportunity to assess the impact of using initial conditions from different analysis systems on weeks 3&4 forecast.

The Breeding Vector and ensemble transform with rescaling (BV-ETR) technique (Wei et al., 2008) and hybrid 3D-Var EnKF DA system were used to produce initial perturbations for the period of Jan. 1996 - Dec. 2, 2015 and afterwards, respectively. The studies in Zhou et al. (2016; 2017) show that the initial perturbation could impact the ensemble spread significantly, but have less impact on the ensemble mean errors and skills. Furthermore, the impact on the spread is only limited in the shorter forecast lead times (week 1, Zhou et al, 2017). Therefore, inconsistent perturbation schemes may have negligible impact of the weeks 3&4 forecasts due to short memory of the atmosphere (Zhu et al, 2005; Song and Mapes 2012).

3. Bias analysis

To calculate bias, the analysis fields of CFSR (Jan.1 1999 to May 11 2011) and GDAS (May 12 2011 to Dec. 31 2016) were used as an approximate truth. Reforecast bias is climatological mean forecast error. The forecast error is defined as the difference of the 11-member ensemble mean forecast and analysis at the same valid time. In calculating the bias climatology from 17-year weekly sampling reforecast dataset, we use a time window of 31 days centered on the day being considered, leading to a total sample size of 17 yr x 4-5 samples/yr = 68 – 85 samples for each grid point and each forecast day.

3.1 Bias distribution

The land-only 2-m temperature errors (or bias) over the Northern Hemisphere (NH) and North America (NA) display a strong seasonal dependence (Fig. 2). Warm bias is prevalent for warm season (April–September) for these large and small domains. It is also evident in NA that the inter-annual and intra-seasonal variability of the bias is larger during boreal winter than other three seasons, hinting a poor predictability of winter-related physical processes. In 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 inter-annual and intra-seasonal variability in snow cover and depth (Robinson 1996; Frei and Robinson 1999; Robinson and Frei 2000; Klingaman et al. 2007). Therefore, it is probable that the large variability of 2-m temperature bias over NA winter was directly associated with the variability of snow characteristics. Of course, this statement needs to be confirmed in the future.

There is also a clear tendency to have a larger (slightly larger) warm (cold) bias for the CFSR period (1999-2010) than the GDAS (2011-2016) period during the summer season (winter season) over the NH. To find out where the systematic difference between the two analysis system periods comes from, we compare the spatial distributions of global 2-m temperature errors between the 2006-2010 and 2011-2015 for July and January 2016 (monthly average) in Fig. 3. In summer month (July), the large difference between the two 5-year periods occurs mainly near Sahara and Middle-East desert (or arid) areas. This may be largely attributed to the modification of surface roughness length formula in the 2011 GFS upgrade (Zheng et al 2012) that lead to a larger change in 2-m temperature analysis and forecast over arid and desert regions. In winter, the difference between the CFSR to GDAS periods is relatively small. But we do find opposite bias characteristic in Kazakhstan and nearby with a positive bias for the GDAS period and negative bias for the CFSR period.

3.2 Saturation analysis of 2-m temperature errors

It is well known that the forecast error grows with lead time, until at some asymptotically long lead; the error reaches a saturated status. Error growth of 2-m temperature with lead-time for the full reforecast period (1999-2016) over NA and the NH land-only domains are depicted in Fig.4. For both domains, errors quickly grow within the first 10 days and gradually saturate afterwards through the weeks 3&4 time scale. The absolute errors (ABSE; dotted curve) for NA and the NH domains are ~79% (4.13/5.25) and 77% (3.82/4.97) of root-mean-square-error (RMSE; solid curve), respectively, if they are at a saturated level (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 error variance is zero or error is uniformly distributed. In our cases, the contribution of

error variance to RMSE is less than ~21% (~23%) for NA (the NH). In general, the errors in NA are slightly larger than those in the NH. On average, the errors of day-11 (mid-day of week 2) forecast for NA and the NH are about 88.6% (3.57/4.13) and 86.6% (3.31/3.82) of their saturation values. It is understood that the timescale of error saturation is strongly dependent on the geographical area. For example, the timescale for the land error saturation (weeks) is shorter than that for the ocean (months) (Song and Mapse 2012). Our preliminary analysis shows the error saturation time is shorter over the Southern CONUS than the Northern CONUS for both summer and winter (not shown). The detailed diagnosis for the reasons causing this difference is needed in the future work.

In order to better understand the error saturation, global 2-m temperature is compared between lead week 2 (7-day mean) and weeks 3&4 (14-day mean). It is evident that the error patterns have nearly fixed geographical structure with lead time in both the summer month (July; Fig. 5a; 5b) and winter month (January; Fig. 5c; 5d). The value of error is also very close to each other except for the north parts of NA and Europe, where the errors at weeks 3&4 are noticeably larger than that at week 2 during the winter season. Longer saturation times (scale) for the high-latitude winter was linked to some larger system with more thermal or mechanical inertia (Song and Mapes, 2012).

To assess the impact of initial conditions on 2-m temperature forecast, we examine time series of year-by-year evolutions for 24-hr, 120-hr, and 480-hr forecasts for NH land only. In the beginning of the model integration (24-hr; Fig. 6a), an impact from using different initial analysis systems can be noted. For example, the 2-m temperature forecast for the GDAS period (green curves) is systematically higher than for the CFSR period (red curves) between July and October. The impact from initial conditions is apparently getting less at 120-hr forecast (Fig. 6b)

and eventually minimal on weeks 3&4 time scale (480-hr; Fig. 6c). This also implies that the observed difference of weeks 3&4 bias (Fig. 2a) between the two analysis periods must come from an inconsistent reference (analysis).

4. Bias correction for weeks 3&4

4.1 Methodology and analysis adjustment

In this study, we use 17 years (1999 - 2015) average reforecast bias to calibrate the 2016 GEFS forecast since we intend to do the calibration for recent forecast using historical information. Therefore, the forecasts being verified are independent from the training data. The bias-corrected forecast F for each grid point i,j is obtained by simply subtracting weeks 3&4 bias $b_{i,j}(t_{w34})$ from raw forecast f,

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

We also apply week 2's bias $b_{i,j}(t_{w2})$ to calibrate weeks 3&4 forecast. Here the biases of week 2 and weeks 3&4 are two 7-day (days 8-14) and one 14-day (days 15-28) averaged forecast errors to match a validated forecast period, respectively. To test the sensitivity of the forecast skill to the number of training years, we also compare the calibrated forecast by using the bias from the most recent 5- (2011–2015), 10- (2006–2015), 17-yr (1999–2015) of training data, and evaluate the 2016 forecasts.

The calibration of the ensemble forecast system is evaluated via the root-mean-square error (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 2016; Melhauser et al., 2016; Zhu et al. 2017). The score measures the improvement of a multi-

category forecast to a reference. The higher the RPSS, the better the probabilistic system performs.

As noted in Fig.2, there is a systematic difference in 2-m temperature bias between the CFSR and GDAS period for the NH domain, which most likely arose from inconsistent references (analyses). To confirm this, we show a land-only year-by-year analysis for the four geographic domains in Fig. 7. As expected, the analysis difference between the two assimilation periods is evident for the NH and tropics (TR) domains with a maximum difference of more than 1º (Fig. 7a and Fig. 7c). The solid curves represent the averages for each analysis period. Note both domains encompass the desert or arid regions in North Africa and Middle-East, the most affected regions by the 2011 GFS upgrade. For the most of the time, the analysis is systematically higher in the GDAS than the CFSR period for the NH and TR. A higher reference analysis in the GDAS period (Fig.7a) induces a smaller warm bias (Fig. 2a) assuming that the forecast is less dependent on the initial analysis for weeks 3&4 time scale.

To make a consistent reference, it is necessary to make some adjustment for the early CFSR analysis. We first calculate 12-year (a^{12y}) (1999-2010) and 5-year (a^{5y}) (2011-2015) averaged analysis for each grid *i,j* and then apply the difference a' to the first 12-year analysis as follow:

$$a'_{i,j} = a^{12y}_{i,j} - a^{5y}_{i,j} \qquad (2)$$

$$a^{adj}_{i,j} = a_{i,j} - a'_{i,j} \qquad (3)$$

Please note that an "analysis adjustment" is based on our early assumption (Fig. 6c; Zhu 2005) which states that the weeks 3&4 forecast errors (or longer lead forecast) have less (or no) impact from the initial condition. Climate trend cannot be well estimated in this study because a full set of the CFSR analysis or GDAS analysis for the studied period is not available. However,

our comparison for the North Africa and Middle East regions does illustrate that the large difference (~3.4°) between the two analysis periods is mainly caused from the inconsistent analysis as indicted by a sharp increase in 2-m temperature in 2011 (Fig.8). In contrast, the natural changing trend during either the earlier 12 years or later 5 years is relatively minor.

To demonstrate the consistency of forecast errors, we have presented domain average errors (or bias) without and with analysis adjustment in Fig. 9. Analysis adjustment merges the two separate groups together and mitigates the inconsistency of 2-m temperature bias for the both regions (Fig.9a vs Fig.9b; Fig.9c vs Fig.9d). In next section, we will examine the bias correction (or calibration) from the different biases (with/without analysis adjustment).

4.2 Calibrating the 2016 forecasts using the 17-yr training dataset

We present here a comparison of the verifications of the raw and the four calibrated weeks 3&4 forecasts over the four geographic domains (NH, NA, the SH, and TR). The week 2 and weeks 3&4 bias with and without analysis adjustment were used to calibrate weeks 3&4 forecasts. The forecast skills for both RMS errors (Fig. 10a) and RPSS (Fig. 10b) get improved after bias correction and analysis adjustment for all the four domains. Analysis adjustment does an excellent job for the NH and TR, but not for NA. Errors for NA are reduced by nearly 20% through the bias correction. It is also evident that forecast skills are very similar whether week 2 or weeks 3&4 bias is used to do calibration. Therefore, this indicates that we could use week 2's bias to calibrate weeks 3&4 forecasts which would optimize the use of computer resources without sacrificing the effectiveness of the calibration. Although its skill improvement is the most substantial, NA still has the lowest RPSS even though it has a similar RMS error to NH.

This could be due to its large bias variance and therefore less predictability compared to the other domains.

To find out the seasonal dependence of the bias corrections, we show the time series of RMSE and RPSS for the raw, bias-corrected weeks 3&4 forecasts with and without analysis adjustment (Fig.11 and 12). The largest improvement occurs over NA (Fig. 11b; Fig. 12b) for the warm season from bias correction mainly. The RPSS increases from a near-zero value to ~ 0.4, while RMSE gets substantially reduced with a maximum reduction up to ~50% in July. A large skill improvement due to the analysis adjustment is for the tropical area (land only; Fig. 11d and Fig. 12d) throughout most of the year.

Figure 13 depicts the distribution of RPSS for the raw (a) and calibrated (b) weeks 3&4 forecasts in 2016. There is a negative skill relative to the climatology for the raw forecast over a considerable region of the Continental United States (CONUS). The 2-m temperature prediction is extremely challenging in the Great Plains, consistent with the findings in Klein et al. (2006). The bias-corrected forecast produces much higher forecast skill throughout the entire CONUS domain. Substantial improvements are detected over the Great Plains where the maximum increase in skill reaches ~0.6 (from ~ -0.45 for the raw forecast to ~0.15 for the bias-corrected forecast) near South Dakota.

4.3 Skill sensitivity to number of training years

The sensitivity of forecast skills to the number of training years has been studied by Hamill (2004); Guan et al. (2015); and Ou et al. (2016). Using the first-generation GEFS reforecast data, 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 once 10-12 years were reached, the incremental increase was much smaller. The sensitivity experiments (Guan et al.

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

To test the sensitivity of the weeks 3&4 forecast skill to the number of training years, here we calibrate the 2016 forecast using the 5-yr (2011-2015), 10-year (2006-2015), and 17-yr (1999-2015) training data. Our results (Fig. 14) show that increasing the number of training years from 5-yr to 10-yr leads a skill gain of 0.016 (or ~5%), while further increasing to 17-yr does not exhibit an important difference from the result using a 10-yr sample. This indicates a 10-yr sample should be an optimal requirement for the weeks 3&4 2-m temperature calibration of the NCEP GEFS SubX version. Our optimal sampling year (10 yrs) for weeks 3&4 forecast is similar to the year (10-12 yrs) for week 2 forecast estimated by Hamill (2004) but less than that (18 yrs) in Ou et al (2016). The difference could be partially attributed to the difference in forecast lead-time, model version, and verification period as pointed out in Ou et al. (2016).

5. Summary and Conclusion

The NCEP/EMC generated an 18-year sub-seasonal reforecast dataset to support the CPC's operational mission. The GEFS-SubX version was run every 7 days initialized at 0000 UTC (every Wednesday) with 11 members, and with inconsistent initial analyses and initial perturbations. Using the dataset, we explore the analyses difference and adjustment, the bias characteristics of weeks 3&4 2-m temperature for the reforecast period and apply the 17-yr (1999-2015) bias to calibrate weeks 3&4 forecasts of 2016. The works have led to a number of conclusions as following:

(1) The forecast of 2-m temperature is strongly biased in North America and the Northern Hemisphere with a warm bias in the summertime. In winter, there is a large inter-annual and intra-seasonal variability of 2-m temperature bias for NA region, likely related to its high variability in snow characteristics. Therefore, it is a challenge to find out corresponding model systematic errors for 2-meter temperature.

(2) The model errors quickly grow within the first 10 days and afterward gradually saturate until the weeks 3&4 time scale. The error of day-11 (or middle of week 2) forecast for NA (the NH) reaches about 88.6% (86.6%) of its saturated (day-28) value. The impact of the initial conditions is almost completely gone at weeks 3&4 time scale. We have noticed that there is a geographic difference for the error saturation. The Southern CONUS tends to have a shorter saturation timescale than that over the Northern CONUS. Further diagnosis is needed to address the reason causing this difference.

(3) A consistent analysis is very important to generate reforecasts and real-time forecasts. Analysis adjustment is an alternative way to make bias characteristics more consistent between the CFSR and GDAS periods. Adjusted analysis can be considered as a backup solution when a full set of reanalysis (or reference) is not available.

(4) Bias correction is very important to reduce systematic error and increase forecast skill for all four domains with a maximum benefit for North America during warm season. The calibration using week-2 bias gives a very similar skill to that using week 3&4 bias, suggesting that week

2's bias could be used to correct weeks 3&4's forecast. This could help save huge computation resources and storage for other applications.

(5) The weeks 3&4 2-m temperature calibrations using 5-yr, 10-yr, and 17-yr sample data have been performed, aimed to determine an optimal sample year. Our results demonstrated a 10-yr training period is close enough to obtain a more skillful forecast in 2016 if analyses (or reference) are consistent.

The current study demonstrates the important value of using reforecast information to improve weeks 3&4 forecast skill for 2-m temperature through fully evaluating the analysis difference as well as temporal and spatial distributions of forecast errors. Analysis of bias characteristics of weeks 3&4 precipitation forecasts and its calibration are being performed. Since July 1 2017, the NCEP GEFS SubX version has generated 35-day forecast in real-time, once per week (every Wednesday 0000 UTC). In the future, we will continue generating the calibration statistics with incoming real-time SubX forecast and further examine the effectiveness and robustness of the calibration method with more data.

Acknowledgements:

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

References:

Candille, G., 2009: The multiensemble approach: The NAEFS example. Mon. Wea. Rev., 137, 1655-1665, doi: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., Vol. 7, 1247-1250.

Cui, B., Z. Toth, Y. Zhu and D. Hou, 2012: Bias Correction For Global Ensemble Forecast, Weather and Forecasting, Vol. 27, 396-410

Frei, A. and D. A. Robinson, 1999: Northern Hemisphere snow extent: Regional variability 1972–1994. Int. J. Climatol., 19, 1535–1560.

Guan, H., B. Cui, Y. Zhu, 2015: Improvement of Statistical Postprocessing Using GEFS Reforecast Information, Weather and Forecasting, Vol. 30, 841-854.

——, 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.

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, doi:10.1175/ 1520-0493(2004)132,1434:ERIMFS.2.0.CO;2.

——, J. S. Whitaker, 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 data set. Bull. Amer. Meteor. Soc., 94, 1553–1565, doi:10.1175/ BAMS-D-12-00014.1.

——, 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.

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. and Forecasting, Vol 32, 2005-2017.

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. and Forecasting, 29, 23–38, doi:10.1175/ WAF-D-13-00102.1.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Wollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, D. Joseph, 1996: The NCEP/NCAR 40-Year Reanalysis Project, Bulletin of the American Meteorological Society, Vol. 77, 437-471.

Kazakova, E. and I. Rozinkina, 2011: Testing of Snow Parameterization Schemes in COSMO-Ru: Analysis and Results, COSMO Newsletter, No.11, 41-51.

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, doi: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, doi: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, doi:10.1175/MWR-D-11-00354.1.

Li. W., Y. Zhu, X. Zhou, D. Hou, E. Sinsky, C. Melhauser, P. Malaquias, H. Guan and R. Wobus, 2018: "Evaluating the MJO Forecast Skill from Different Configurations of NCEP GEFS Extended Forecast". J. of Climate (under review),

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, STI Climate Bulletin: http://www.nws.noaa.gov/ost/climate/STIP/41cdpw_digest.html, p30-34.

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, Weather and Forecasting, Vol 31 1093-1107

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

——, and A. Frei, 2000: Seasonal variability of Northern Hemisphere snow extent using visible satellite data. Prof. Geogr., 52, 307–315.

Saha, S. and Coauthors, 2014: The NCEP Climate Forecast System Version 2. J. Climate, 27, 2185–2208, doi: 10.1175/JCLI-D-12-00823.1.

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, doi:10.1029/2011MS000094.

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, doi:10.1029/2009GL039089.

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, doi:10.1111/j.1600-0870.2007.00273.x.

Wilks, D. S., 2011: Statistical Methods in the Atmospheric Sciences. Academic Press, 676 pp.

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, doi: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. Weather and Forecasting, Vol. 31, 2058-2074.

——, ——, ——, Y. Luo, J. Peng and D. Wobus, 2017: The NCEP Global Ensemble Forecast System with the EnKF Initialization. Weather and Forecasting. Vol. 32, 1989-2004.

Zhu, Y., 2005: Ensemble Forecast: A New Approach to Uncertainty and Predictability, Advance in Atmospheric Sciences, Vol. 22, No. 6, 781-788

——, X. Zhou, M. Pena, W. Li, C. Melhauser and D. Hou, 2017: Impact of Sea Surface Temperature Forcing on Weeks 3&4 Forecast Skill in the NCEP Global Ensemble Forecast System, Wea. and Forecasting, Vol. 32, 2159-2173

——, X. Zhou, L. Wei, D. Hou, C. Melhauser, E. Sinsky, M. Pena, B. Fu, H. Guan, W. Kolczynsk, R. Wobus and V. Tallapragada, 2018: An Assessment of Subseasonal Forecast Skill Using an Extended Global Ensemble Forecast System (GEFS), Submit to Journal of Climate (in review).

Figure captions:

Figure 1. Evolution of initial analyses and perturbations during the 18-year GEFS reforecast period (Jan.1999 - Dec. 2016). Since using GDAS analyses from Jan. 1 2010, there were four GFS/GDAS upgrades those are May 9 2011, May 22 2012, January 14 2015, and May 11 2016, respectively.

Figure 2. Time series in 2-m temperature forecast error (or bias) for weeks 3&4 of the NH (left panel) and NA (right panel) domains. Each curve represents one particular year. Red and green curves indicate the errors for 1999–2010 and 2011–2015, respectively. Thick black curves are the errors for 2016.

Figure 3. Spatial distributions of weeks 3&4 2-meter temperature bias (30 days running mean) for summer month (July) of 5 years (a. year 2006-2010; b: year 2011-2015) and winter month (January) of 5 years (c. year 2006-2010; d. year 2011-2015).

Figure 4. 18-year domain averages (land only) of 2-meter temperature RMSE and absolute error for the NA (left panel) and NH (right panel) out to 35 days.

Figure 5. Spatial distribution comparisons of week 2, weeks 3&4 2-meter temperature mean error (or bias) from 30 days running mean) for 18 years (1999-2016) of summer month - July (a. week 2; b. weeks 3&4) and winter month - January (c. week 2; d. week 3&4).

Figure 6. Time series in 2-m temperature forecast over the NH region (land only) for the 24-hr

(a), 120-hr (b), and 480-hr (c) lead-time. Each curve represents one particular year. Red curves are for 1999–2010 and green curves are for 2011–2015.

Figure 7. The time series of year-by-year 2-m temperature analyses for the NH (a), NA (b), SH (c), and TR (d) (land only). Red curves are for 1999–2010 and green curves are for 2011–2015. Black solid curves are the averages for the CFSR period and black dash lines are the averages for the GDAS period.

Figure 8. Domain averaged 2-m temperature analyses for North Africa and Middle East regions during years of 1999-2015. Red plus (+) is for the CFSR period and blue cross (x) is for the GDAS period. Red (blue) solid line represents the line of best fit for the CFSR (GDAS) period. Black lines are the averaged values for the corresponding two periods.

Figure 9. The time series of weeks 3&4 2-m temperature forecast errors (or biases) for the NH (top panels) and TR (bottom panels) domains without (left panels) and with (right panels) analysis adjustments. Red lines indicate the errors for 1999–2010 and green lines indicate the errors for 2011–2015. Black lines indicate error for 2016.

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

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

Figure 12. The same as Fig. 11 except for RPSS (Ranked Probabilistic Skill Score).

Figure 13. RPSS of weeks 3&4 2-m temperature forecasts for the CONUS in 2016 for the raw a) and bias-corrected (BC_BIASwk34adj) forecasts (b).

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

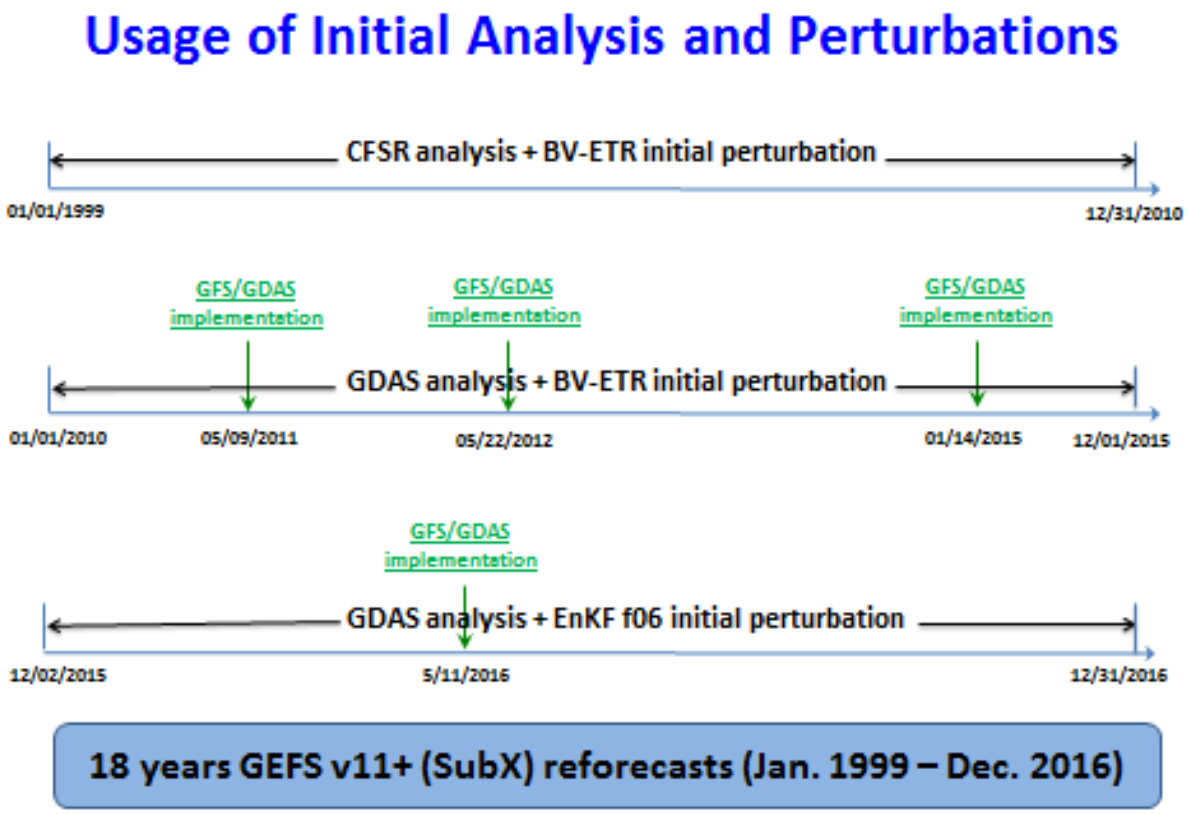

Figure 1. Evolution of initial analyses and perturbations during the 18-year GEFS reforecast period (Jan.1999 - Dec. 2016). Since using GDAS analyses from Jan. 1 2010, there were four GFS/GDAS upgrades those are May 9 2011, May 22 2012, January 14 2015 and May 11 2016 respectively.

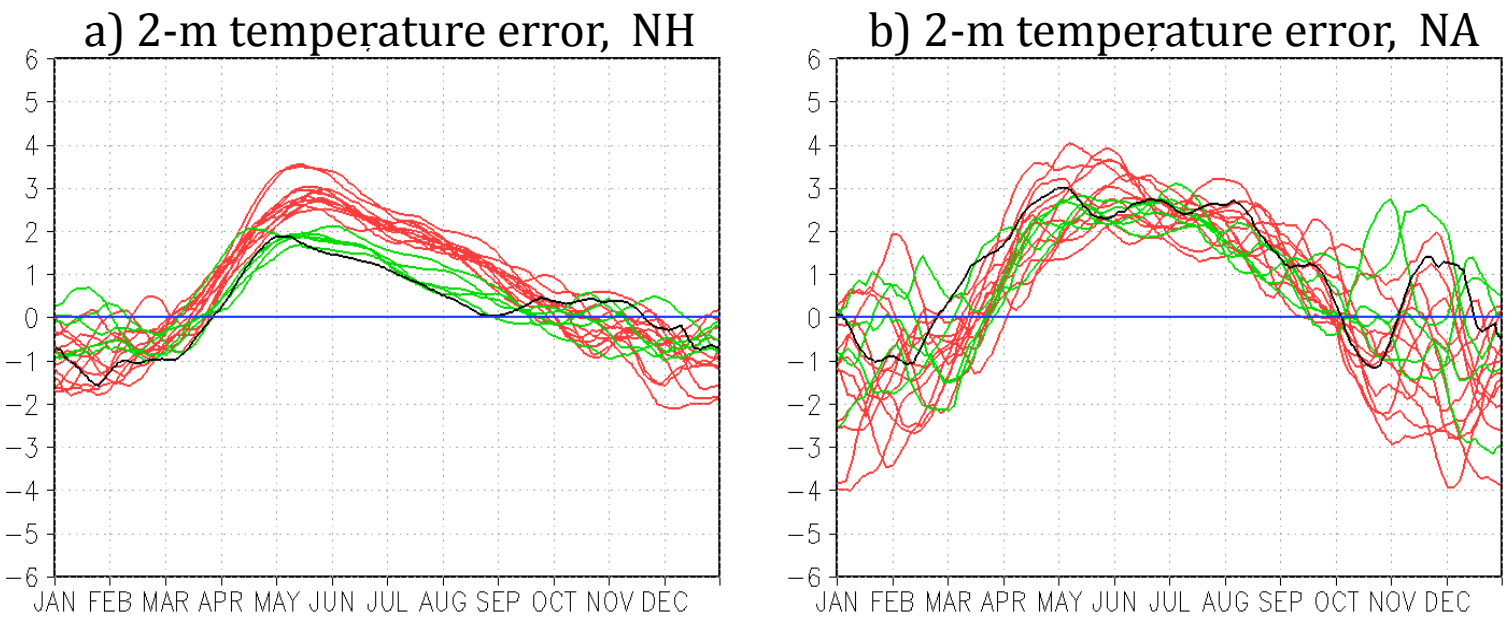

Figure 2. Time series in 2-m temperature forecast error (or bias) for weeks 3&4 of the NH (left panel) and NA (right panel) domains. Each curve represents one particular year. Red and green curves indicate the errors for 1999–2010 and 2011–2015, respectively. Thick black curves indicate the errors for 2016.

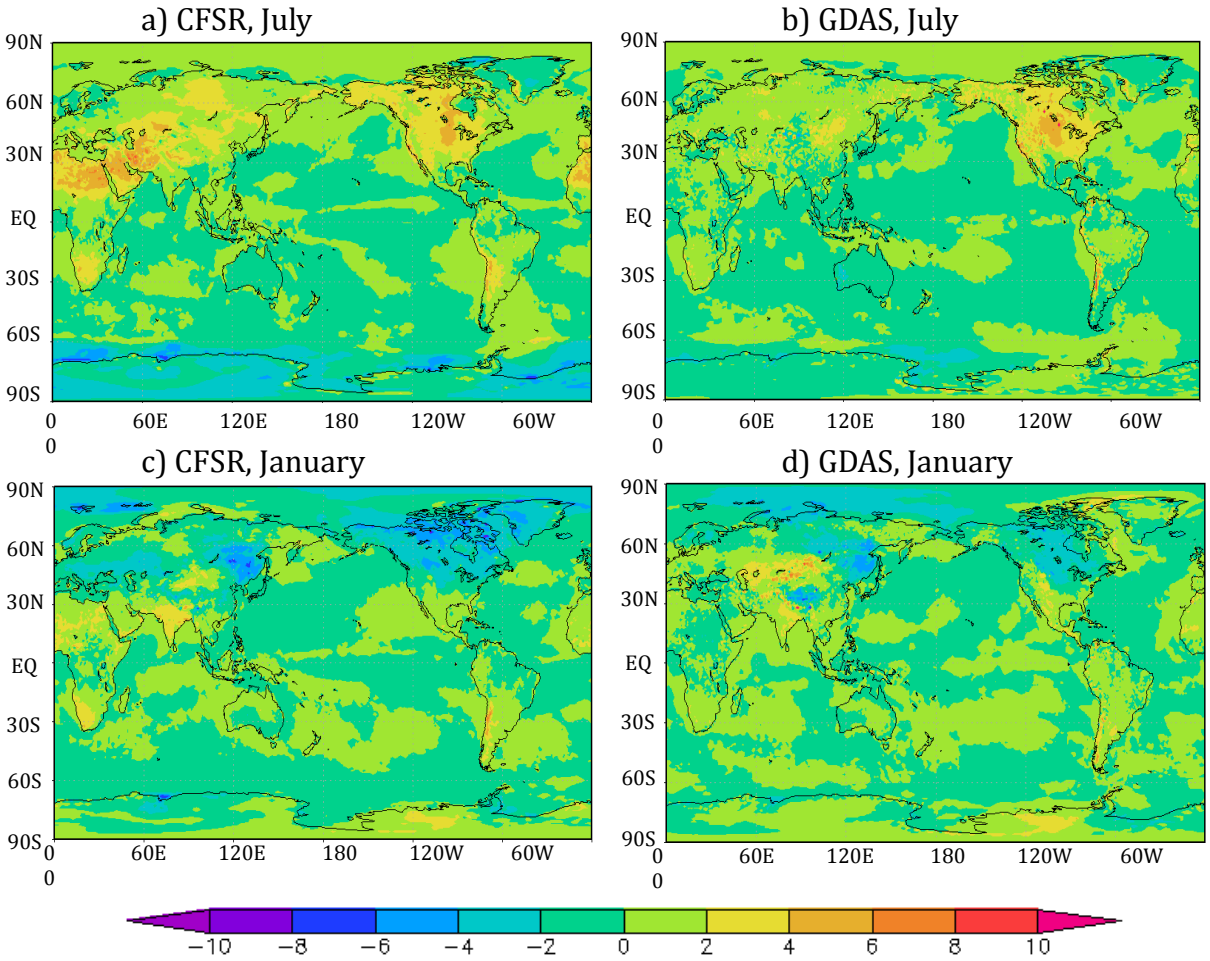

Figure 3. Spatial distributions of weeks 3&4 2-meter temperature bias (30 days running mean) for summer month (July) of 5 years (a. year 2006-2010; b: year 2011-2015) and winter month (January) of 5 years (c. year 2006-2010; d. year 2011-2015).

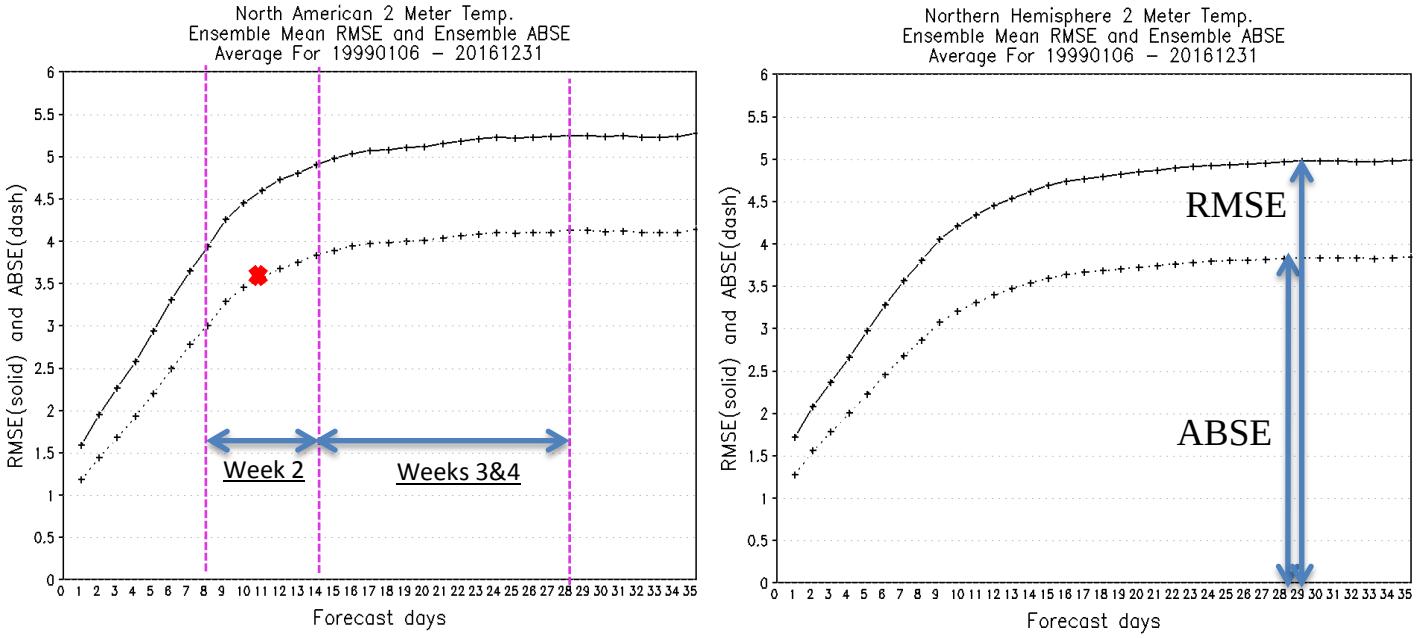

Figure 4. 18-year domain averages (land only) of 2-meter temperature RMSE and absolute error (ABSE) for the NA (left panel) and NH (right panel) out to 35 days.

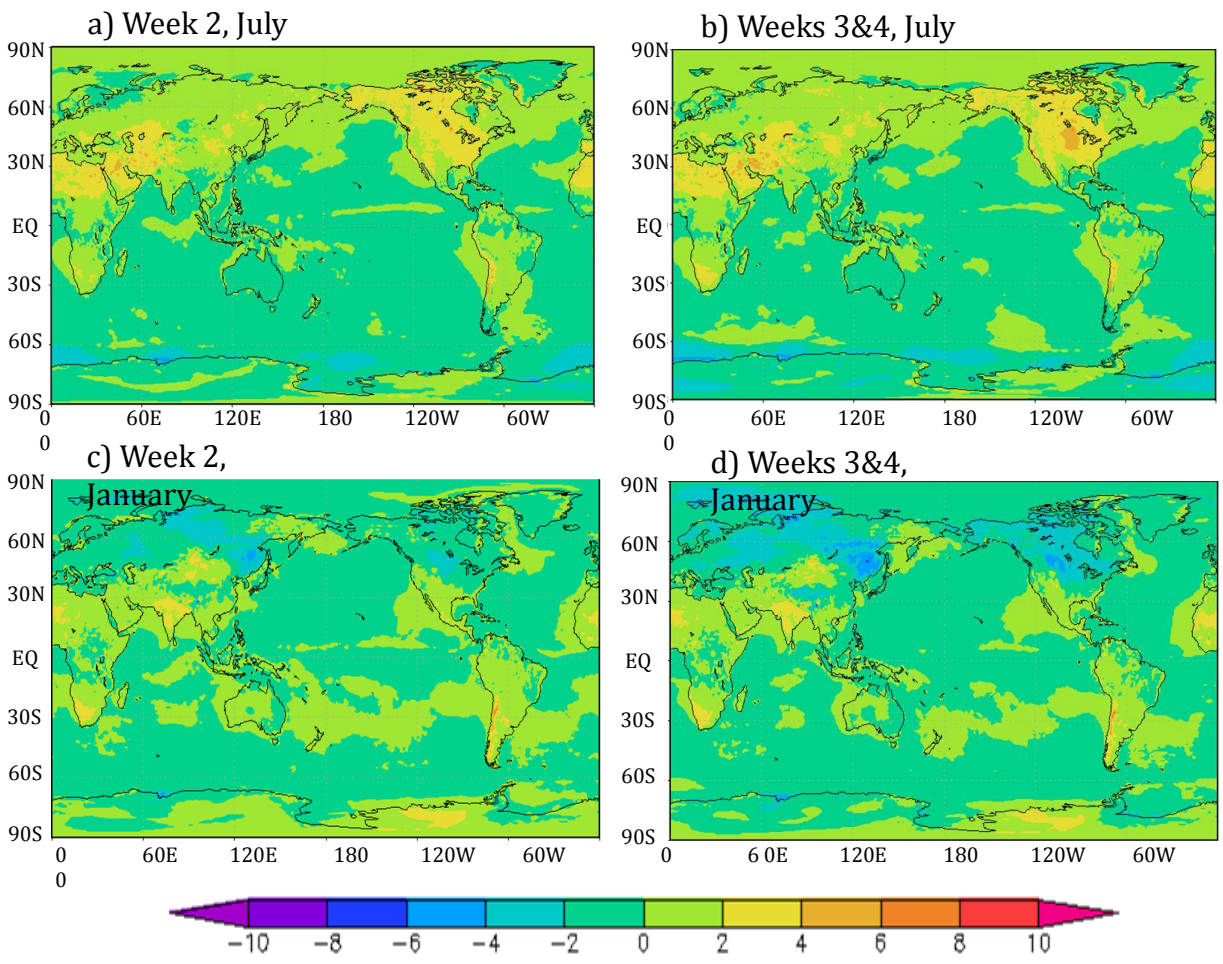

Figure 5. Spatial distribution comparison of week 2, weeks 3&4 2-meter temperature mean error (or bias) from 30 days running mean for 18 years (1999-2016) of summer month - July (a. week 2; b. weeks 3&4) and winter month - January (c. week 2; d. weeks 3&4).

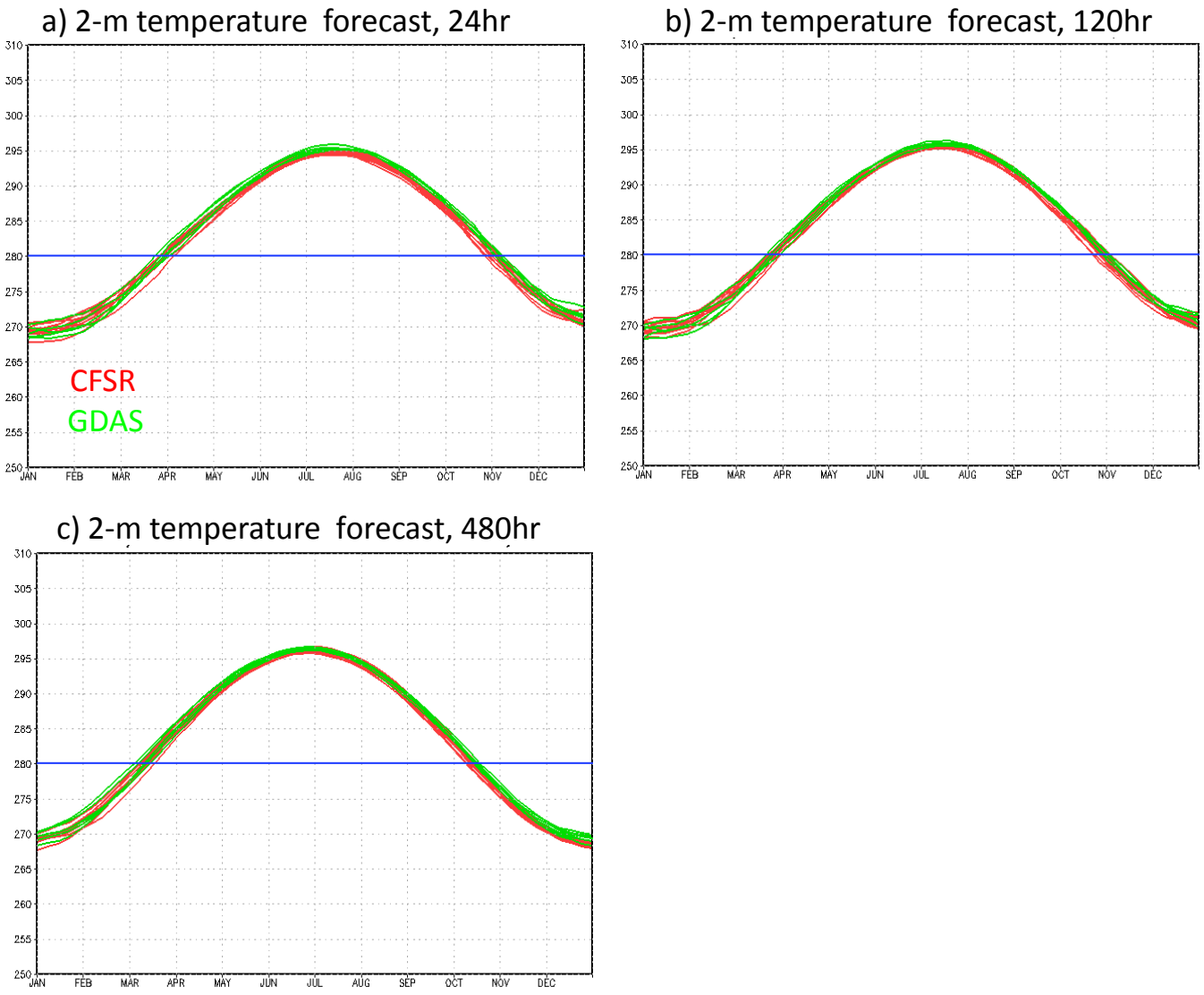

Figure 6. Time series in 2-m temperature forecast over the NH region (land only) for the 24-hr (a), 120-hr (b), and 480-hr (c) lead-time. Each curve represents one particular year. Red curves are for 1999–2010, green curves are for 2011–2015.

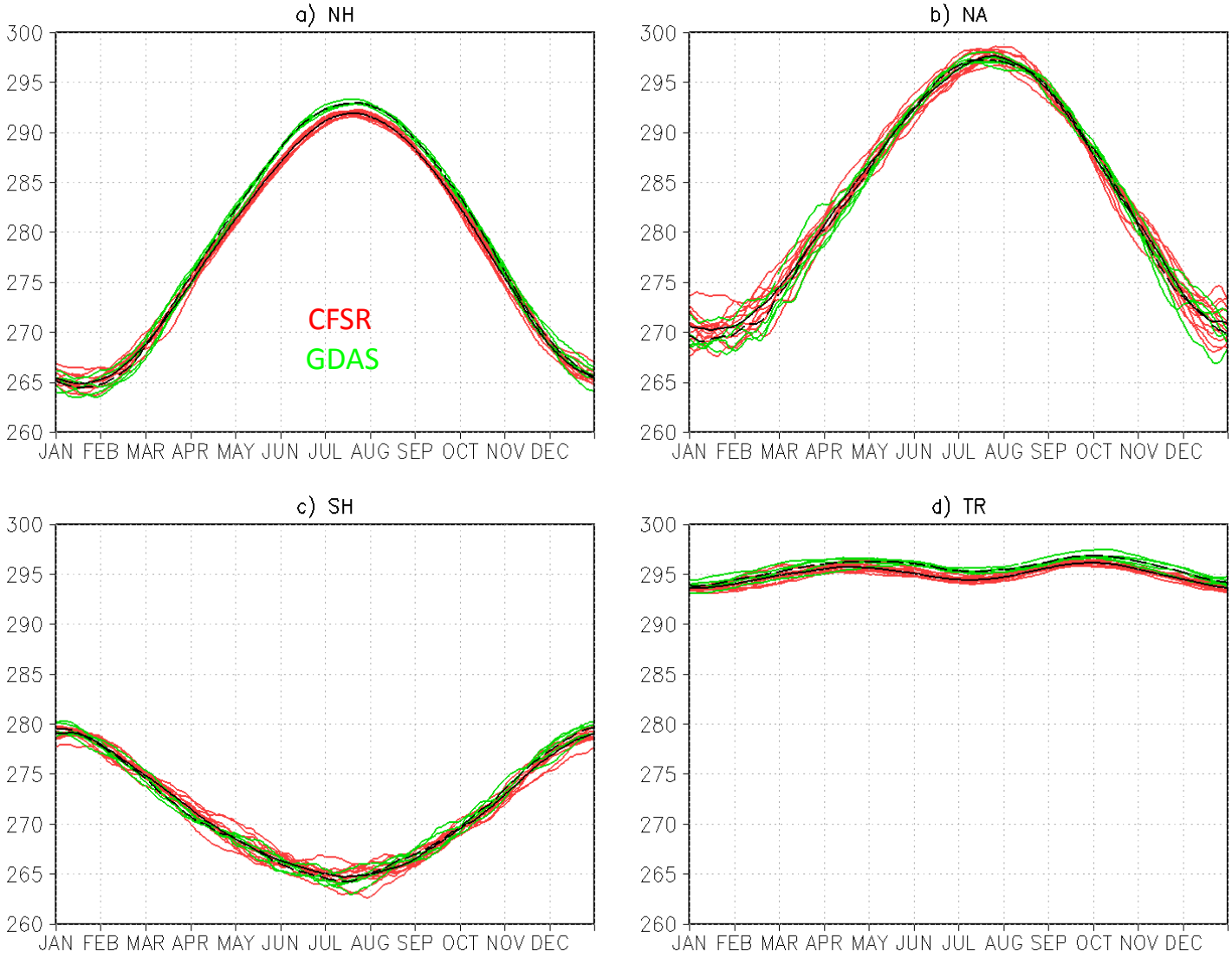

Figure 7. The time series of year-by-year 2-meter temperature analyses for the NH (a), NA (b), SH (c), and TR (d) (land only). Each curve represents one particular year. Red curves are for 1999–2010 and green curves are for 2011–2015. Black solid curves are the averages for the CFSR period and black dash lines are the averages for the GDAS period.

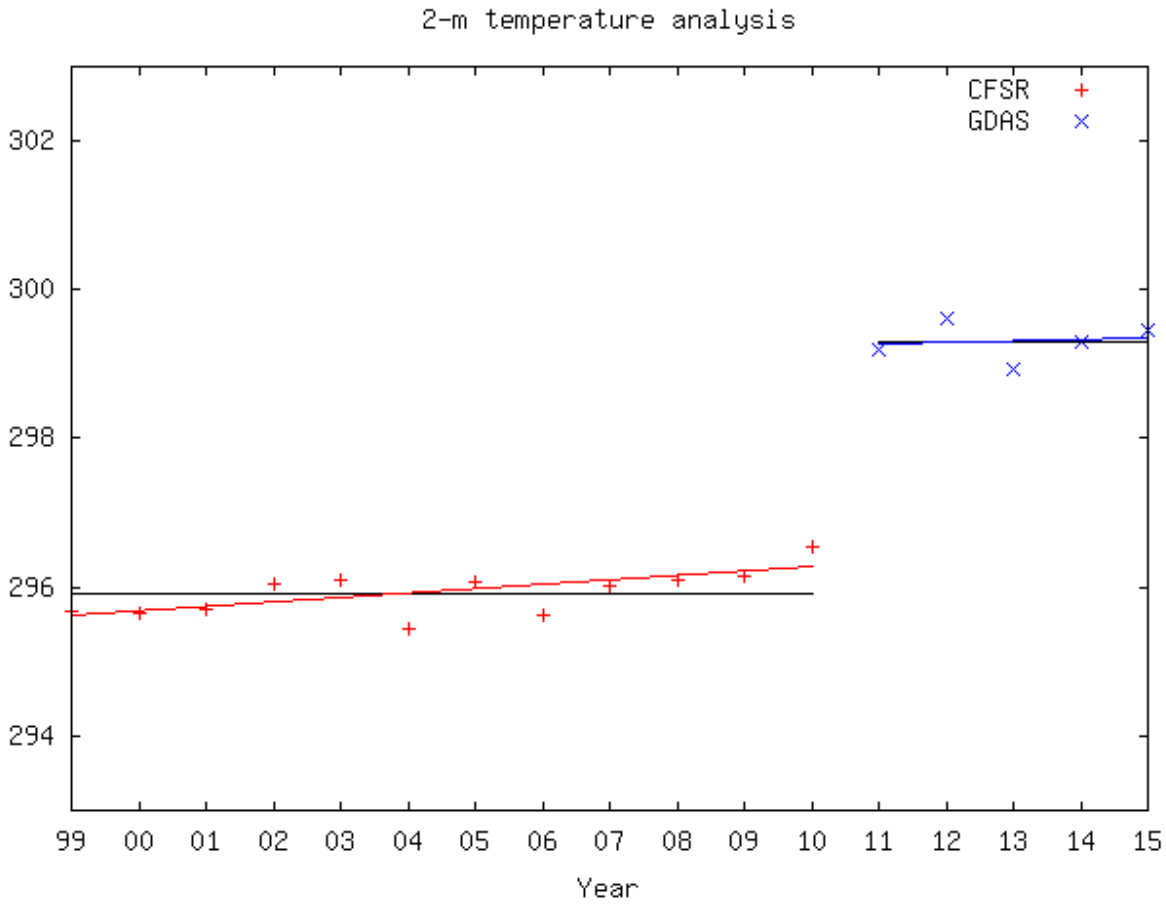

Figure 8. Domain averaged 2-m temperature analyses for North Africa and Middle East regions during years of 1999-2015. Red plus (+) is for the CFSR period and blue cross (x) is for the GDAS period. Red (blue) solid line represents the line of best fit for the CFSR (GDAS) period. Black lines are the averaged values for the corresponding two periods.

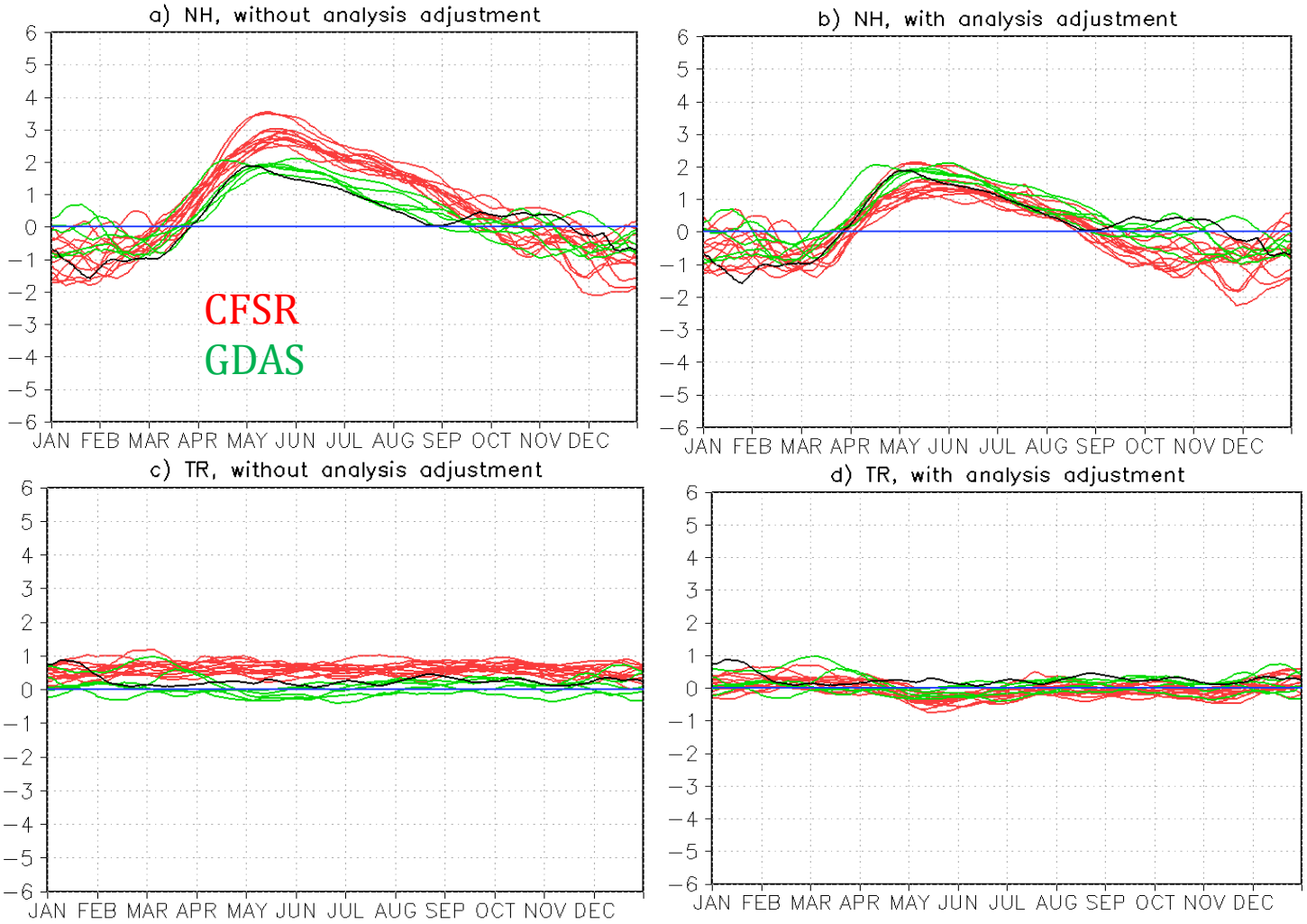

Figure 9. The time series of weeks 3&4 2-m temperature forecast errors (or biases) for the NH (top panels) and TR (bottom panels) domains without (left panels) and with (right panels) analysis adjustments. Each curve represents one particular year. Red curves indicate the errors for 1999–2010 and green curves indicate the errors for 2011–2015. Black lines indicate errors for 2016.

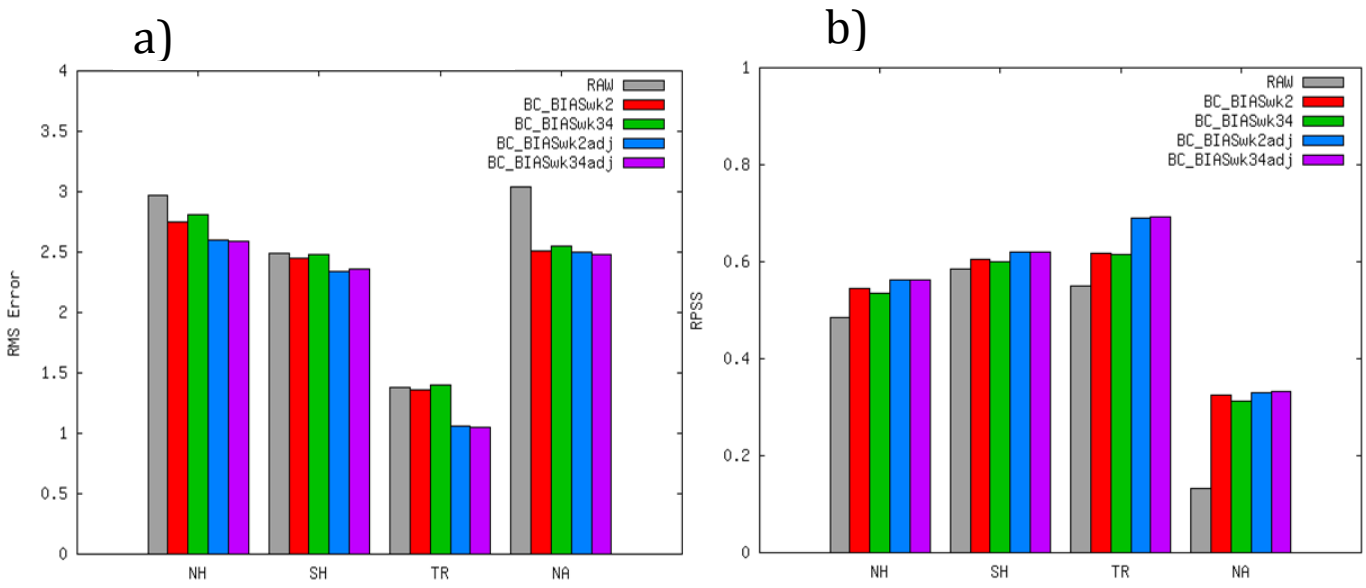

Figure 10. RMSE (a) and RPSS (b) of weeks 3&4 land-only 2-m temperature forecasts in 2016, averaged over the NH, NA, SH, TR for the raw (grey bar) and four bias-corrected (other color bars) forecasts. The BC_BIASwk2 (red) and BC_BIASwk34 (green) (BC_BIASwk2adj (blue) and BC_BIASwk34adj (purple)) denote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

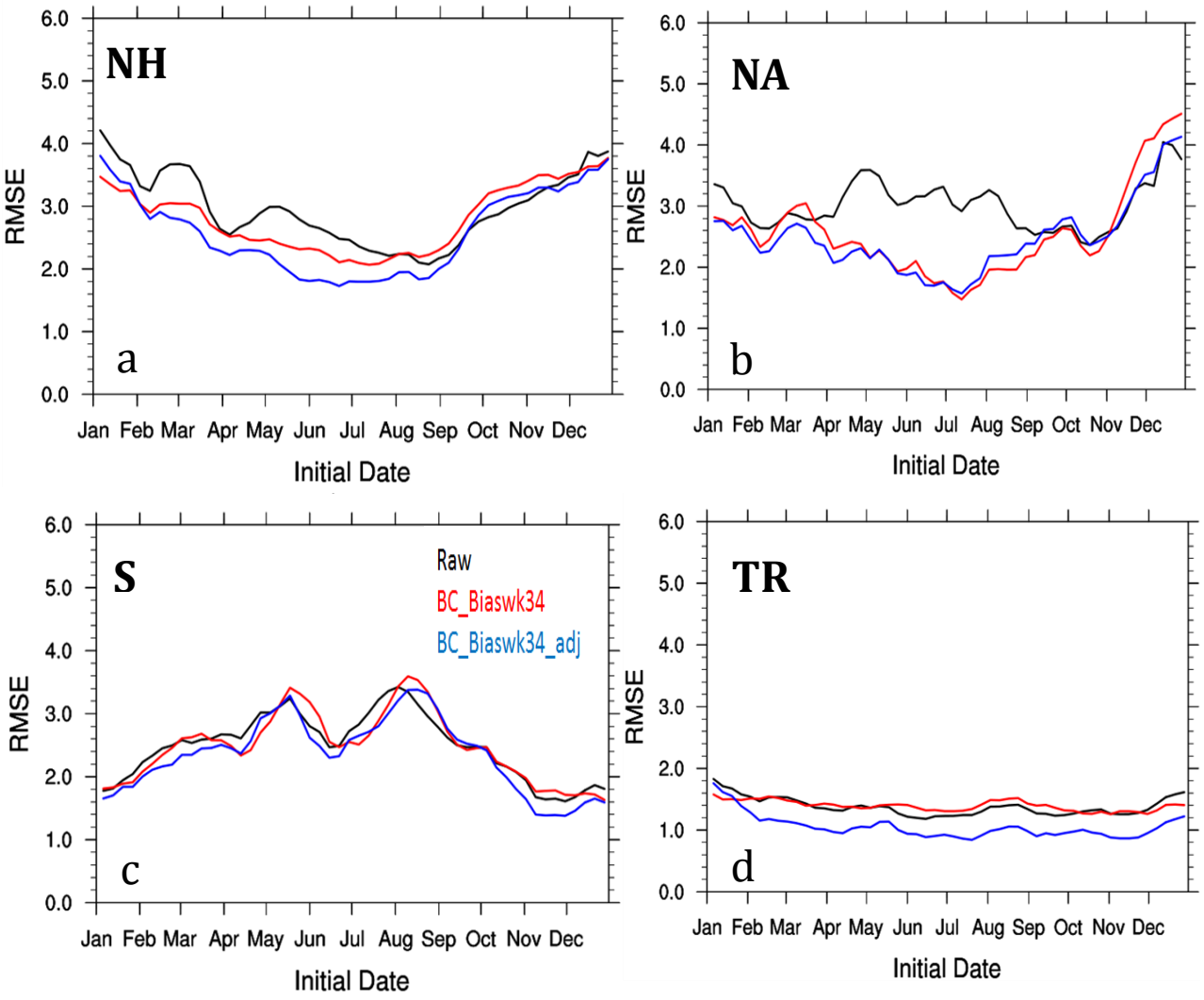

Figure 11. RMSE of weeks 3&4 land-only 2-m temperature forecasts in 2016 averaged over the NH (a), NA (b), SH (c), and TR (d) for the raw (black) and two bias-corrected forecasts. The BC_BIASwk34 and BC_BIASwk34adj denote the bias-corrected forecasts without (red) and with (blue) analysis adjustment,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

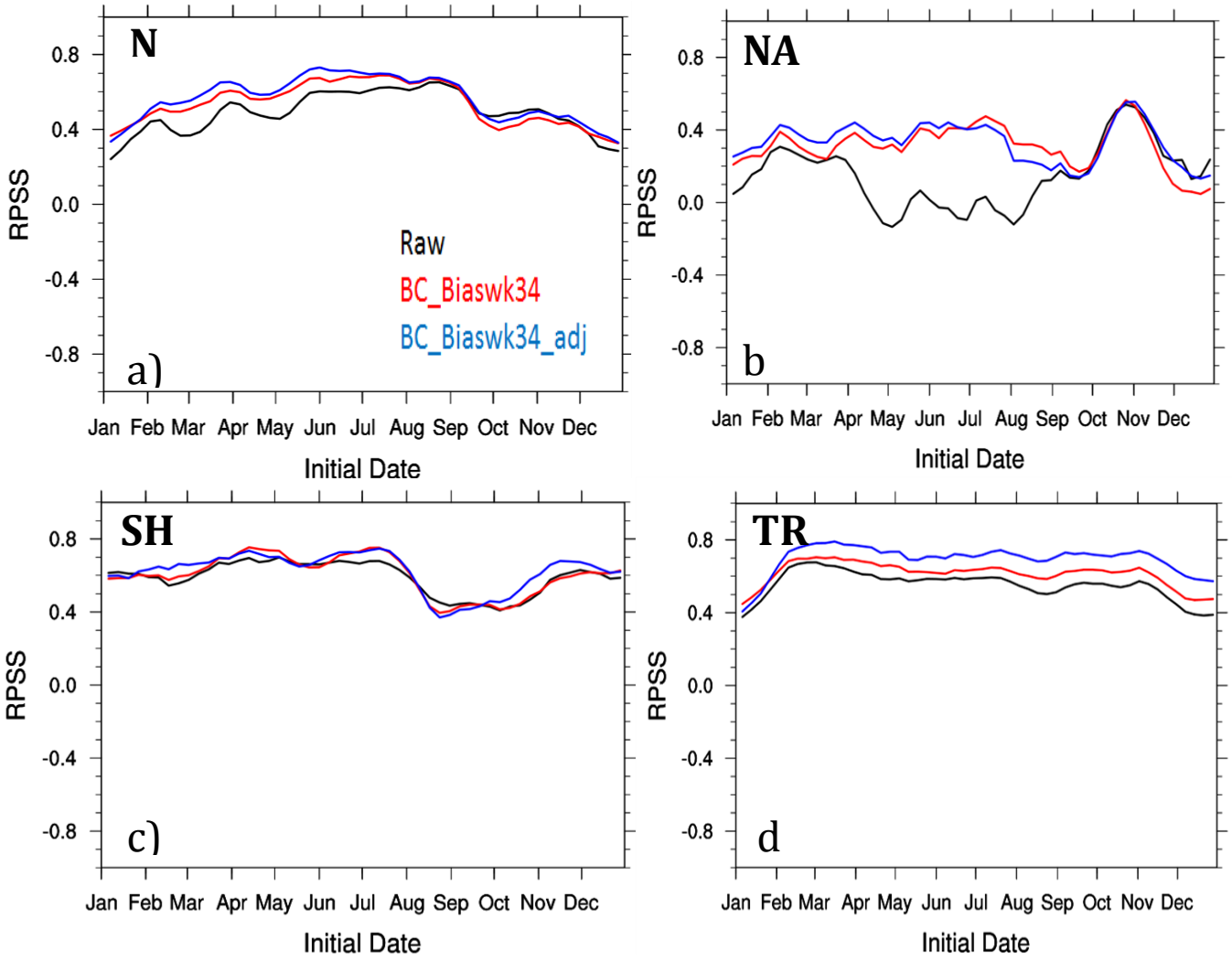

Figure 12. The same as Fig.11 except for RPSS (Ranked Probabilistic Skill Score).

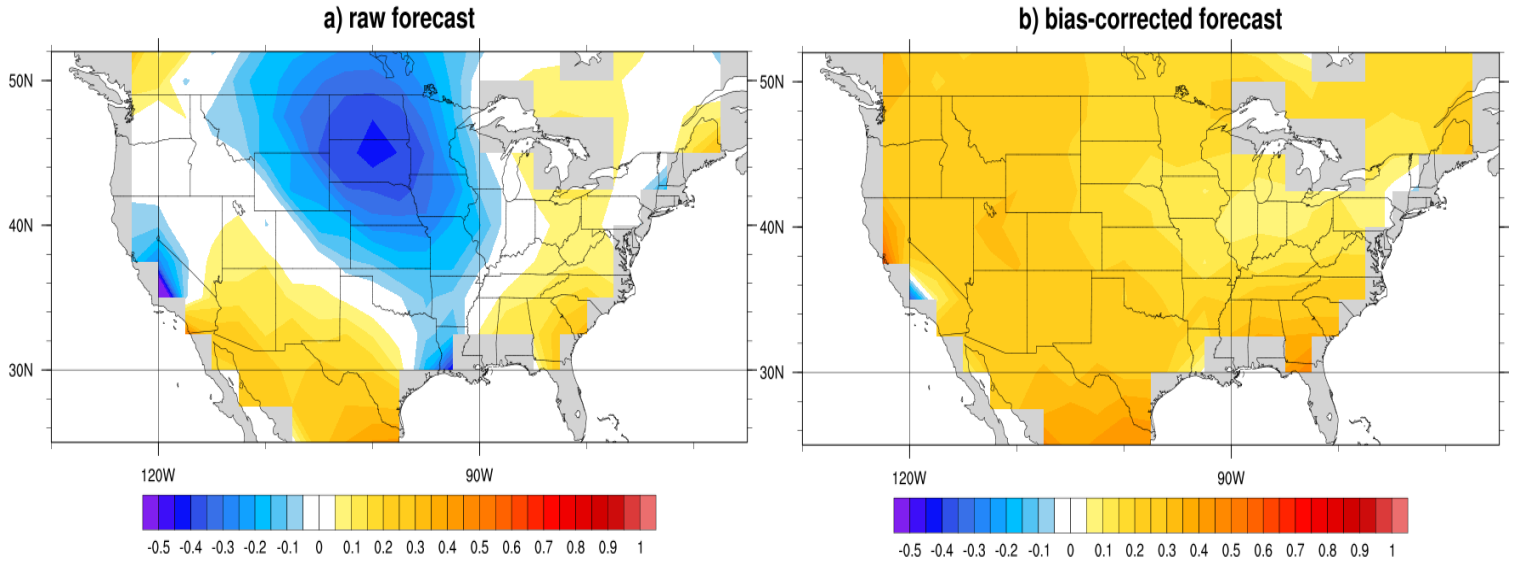

Figure 13. RPSS of weeks 3&4 2-m temperature forecasts for the CONUS in 2016 for the raw a) and bias-corrected (BC_BIASwk34adj) b) forecasts.

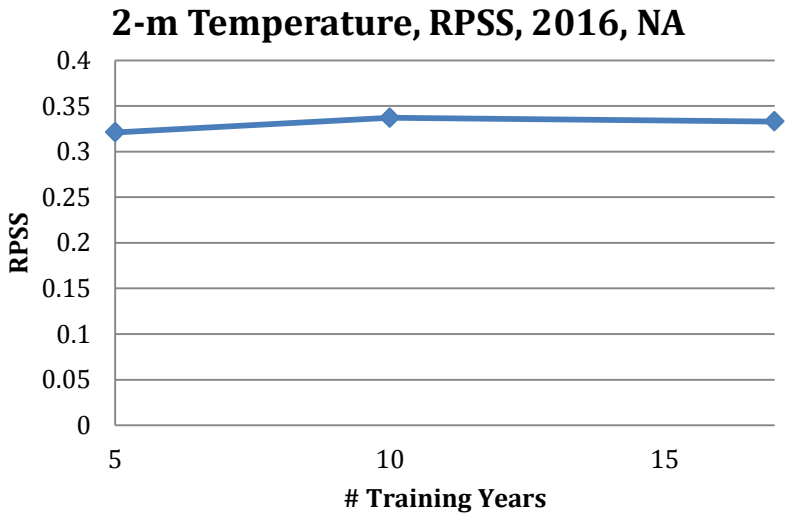

Figure 14. RPSS of weeks 3&4 2-m temperature forecasts as a function of the number of training years. Skill scores represent the average score across the CONUS in 2016.