Development of verification methodology for extreme weather forecasts

Hong Guan[1,2] and Yuejian Zhu[1]

[1]Environmental Modeling Center/NCEP/NWS/NOAA, College Park, MD

[2]System Research Group Inc., Colorado Springs, CO

Submitted to *Weather and Forecasting*

* *Corresponding author address:*

Dr. Hong Guan,

Environmental Modeling Center/NCEP/NWS/NOAA

5830 University Research Court

College Park, MD 20740

E-mail: Hong.Guan@noaa.gov

Abstract

In 2006, the statistical post-processing of the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) and North American Ensemble Forecast System (NAEFS) was implemented to enhance probabilistic guidance. Anomaly Forecast (ANF) is one of the NAEFS products, generated from bias-corrected ensemble forecasts and reanalysis climatology. The Extreme Forecast Index (EFI), based on a raw ensemble forecast and model-based climatology, is another way to build an extreme weather forecast.

In this work, the ANF and EFI algorithms are applied to extreme cold temperature and extreme precipitation forecasts during the winter of 2013-2014. A highly-correlated relationship between the ANF and EFI allows the determination of two sets of thresholds to identify extreme cold and extreme precipitation events for the two algorithms. An EFI of -0.78 (0.687) is approximately equivalent to a -2σ (0.95) ANF for the extreme cold event (extreme precipitation) forecast.

The performances of the two algorithms in forecasting extreme cold events are verified against analysis for different model versions, reference climatology, and forecasts. The verification results during the winter of 2013-2014 indicate the ANF forecasts more extreme cold events with a slightly higher skill than the EFI. The bias-corrected forecast performs much better than the raw forecast. The current upgrade of the GEFS has a beneficial effect on the extreme cold weather forecast. Using the NCEP Climate Forecast System Reanalysis and Reforecast (CFSRR) as a climate reference gives a slightly better score than the 40-year reanalysis. The verification methodology is also extended to an extreme precipitation case, showing a broad potential use in the future.

1. Introduction

An extreme weather event is unusual, unexpected, or rare weather. It could be defined from either a climatological base, forecast base, or a user specification. In general, it results in the loss of lives, property, equipment, etc. For example, the special report of the Intergovernmental Panel on Climate Change (IPCC) (2011) shows the annual losses from weather- and climate-related disasters since 1980 has ranged from a few US$ billion to more than 200 billion. Therefore, developing accurate forecast guidance and products to warn users about weather related risks has an important impact on the social economy. A good guidance product would allow users make early decisions and improve protection.

A number of forecast methods have been developed and applied to identifying extreme weather events at various world forecast centers (Zhu and Cui 2007; Lalaurette 2003; Zsótér 2006; Dutra et al. 2013; and Hamill et al. 2013). The concept of the Extreme Forecast Index (EFI), originally introduced by Lalaurette, (2003), is a measure of the difference between a forecast probabilistic distribution and model climate distribution. To increase the sensitivity of forecasts of extreme events, this index was further adapted in 2006 (Zsótér) by adding more weight to the tails of probability distributions. This index has been applied to extreme temperature, wind, and precipitation forecasts at the European Centre for Medium-Range Weather Forecasts (ECMWF), Canadian Meteorological Center (CMC), and the Earth System Research Laboratory (ESRL) of the National and Oceanic and Atmospheric administration (NOAA).

Anomaly Forecast (ANF) is a more natural method to forecast extreme weather events. It measures forecast distribution departure from the climatological distribution. The method has been widely applied to forecasts of extreme heat waves, winter storms, etc. (Grumm 2001;

Graham and Grumm 2010). ANF was implemented as a forecast product at NOAA's National Weather Service (NWS) in December 2007 (Zhu and Cui 2007). Based on the NCEP/NCAR 40-year reanalysis, a daily climatological distribution (probability distribution function or PDF) has been created for 19 atmospheric variables such as height, temperature, winds, etc. ANF products have been generated from a bias corrected ensemble forecast (or probabilistic forecast). The products provide 1) ensemble mean as a percentile of the climatological distribution and 2) each ensemble member as a percentile of the climatological distribution. Based on these products, users could build various ANFs, such as greater than 1-sigma, 2-sigma, and 3-sigma standard deviations ANFs for various meteorological elements. Furthermore, by comparing the forecast PDF to climatological PDF, the users could easily identify an extreme weather event.

In this paper, we develop a verification methodology to compare and evaluate the extreme weather forecast products from the ANF and EFI. After explaining the verification metrics, we evaluate products from different model versions (or model upgrade), different references, and products based on a raw forecast and bias corrected forecast. We first introduce the model and dataset in Section 2 and then highlight the two extreme weather forecast methods in Section 3. We also develop and apply a verification methodology to evaluate extreme cold weather forecasts and extreme precipitation forecasts in Section 4. The summary will be given in Section 5.

2. Model and data sets

In this study, Global Ensemble Forecast System (GEFS) version 10 (v10) (Zhu et al. 2012) and v11 (Zhou et al. 2016) forecasts are used to calculate the ANF and EFI. The outputs include raw and bias-corrected ensemble forecasts (Cui et al. 2012). The model climatology and analysis (or observation) climatology serve as the reference climatology for the raw and bias-corrected forecasts, respectively. For the raw forecast, GEFS v11 is tested. The model climatology is calculated using an 18-year control only reforecast dataset.

The GEFS v10 was implemented on February 14, 2012 at NCEP. It consists of 21 members (one control member and 20 perturbed members) and is run 4 times daily (0000, 0600, 1200, and 1800 UTC). In this study, we use only the 0000 UTC cycle forecasts. All members use an identical set of physical parameterizations (Zhu et al. 2007). The model is run at a horizontal resolution of T254 (~55 km) for the first 8 days and T190 (~70 km) for the next 8 days, with 42 hybrid vertical levels. The hybrid GSI/EnKF analysis (Kleist and Ide 2015) is used as the initial condition. The initial perturbations are created with the Bred Vector–Ensemble Transform with Rescaling (BV-ETR, Wei et al. 2008) technique. Model uncertainty is estimated using the stochastic total tendency perturbations (STTP) method (Hou et al. 2008). For the bias-corrected dataset, the model bias was removed using a decaying averaging post-processing technique (Cui et al. 2012).

There are three major changes from the v10 to v11. First, in the v11, Euler's integration method is replaced by the Semi-Lagrangian method in order to save computing time (Sela 2010). Second, the Ensemble Kalman Filter (EnKF) 6-h forecast is used as the basis of ensemble initial perturbation instead of BV-ETR generation. The details of the EnKF technique can be found in references by Whitaker and Hamill, 2012; Whitaker et al. 2008; Wang et al. 2013; and Kleist and Ide 2015. Third, the horizontal resolution was increased to 34 km (T574) and 55 km (T384) for the first and next 8 days, respectively. The number of vertical levels was increased to 64 levels.

The 18-year (1995-2012) control-only v11 reforecast was run at the 0000 UTC cycle every other day. The reforecast dataset was interpolated bilinearly to 1°x1° latitude/longitude grids from the native resolutions. The model native resolutions are about 34km and 55 km at mid-latitudes for the first and last 8 days, respectively. From the 1°x1° dataset, the model climatology for each day and each grid point was generated. In calculating the climatology, we also include 8 nearby points and use a time window of 5 days centered on the day being considered, leading to a total sample size of 243 data (9 years x 3 day/year x 9 points) for each gridpoint.

The analysis climatology of 2-m temperature includes NCEP/NCAR 40-year reanalysis data (1959-1998) (Kalnay et al. 1996) and NCEP Climate Forecast System Reanalysis and Reforecast (CFSRR) 30-year reanalysis data (1979-2008) (Saha et al. 2010). The CFSRR climatology has been generated from the latest numerical weather prediction (NWP) model and assimilation system. Therefore, its quality has been much improved through various enhancements, such as improved quality of observations, a state-of-art model and assimilation system, and much higher spatial resolution. It has been pointed out that for the near-surface temperature the CFSRR produces a much finer structure than the NCEP/NCAR reanalysis (Personal communication with Bo Yang).

A climatological distribution could be presented in terms of the climatological mean and standard deviation if a variable has a (quasi-) normal distribution. For the two sets of reanalysis, the first four Fourier modes (higher smoothing) have been used to generate daily climatological means to include annual, semi-annual, and seasonal cycles. Climatological standard deviations are linearly interpolated from monthly to daily means. For the NCEP/NCAR 40-year reanalysis, the best analysis resolution is 2.5° x 2.5° globally. We have to interpolate the data to 1.0° x 1.0° to match the forecast resolution. The original resolution of the CFSRR is 1.0° x 1.0° resolution.

The analysis climatology of precipitation was calculated based on Climatology-Calibrated Precipitation Analysis (CCPA) (Hou et al. 2014) over the CONUS. A gamma distribution was used to fit precipitation distribution for each day of the year and each 1x1 grid point. The distribution parameters were determined via L-moment method (Hosking 1990; Hosking and Wallis 1997). The details on the generation of climatology can be found in the website (http://www.emc.ncep.noaa.gov/gmb/yluo/AMS_CCPA_Climatology%20[Compatibility%20Mode].pdf, updated January 2013).

3. Forecast product generation methodology

a) ANF

ANF is defined as the difference between ensemble forecast ($F_{en}(p)$) and the expected value of climate distribution (C),

$$ANF = F_{en}(p) - C \tag{1}$$

In this work, we specifically calculate the ANF for the ensemble mean and the 50th percentile for 2-m temperature and precipitation, respectively. For 2-meter temperature, we calculate the value of ANF divided by one climatological standard deviation, so called standardized anomaly in Grumm (2001). For 24-hr accumulated precipitation, we find the location (or value) where the 50th percentile (or median) of the ensemble forecast lies on the climatological distribution. The climatological distribution for the 2-m temperature and precipitation are assumed as normal distribution $C=N(x, \mu, \sigma^2)$ and Gamma distribution $C=\Gamma(x, k, \theta)$, respectively. Previous work (Hou et al. 2014) demonstrated that a gamma distribution can well simulate the distribution of preciptation over North Amrica. The x, μ, σ^2, k, and θ represent location, mean, variance, shape factor, and scale parameter for the corresponding distributions, respectively.

b) Extreme Forecast Index (EFI)

For any given variable, the EFI (Lalaurette 2003; Zsoter 2006) may be expressed as

$$EFI = \frac{2}{\pi}\int_0^1 \frac{p - F_f(p)}{\sqrt{p(1-p)}}dp \qquad (2)$$

Where p is the proportion of ranked climate record and $F_f(p)$ is a function denoting the proportion of ensemble members lying below the p quantile of the climate record. The values of EFI are between -1and 1. If the ensemble member probability distribution agrees with the climate probability distribution, then EFI = 0. In special cases where the values of all ensemble member forecasts are above the absolute maximum in the model climate, the EFI = +1; if all forecast values are below the absolute minimum in the model climate, the EFI = -1. The equation is solved numerically with an increment of p equal to 0.01.

4. Verification

4.1 Methodology

Although various products of extreme weather forecasts have been generated in real time and the applications are widely used in many areas, the verification of these products has been a challenge. To our knowledge, the verification methodology is mainly based on scatter-plots of analysis anomalies and EFI, hit rates, false alarm rate, and ROC (relative operational characteristics) area (Toth et. al. 2003; Petroliagis and Pinson 2012; Matsueda and Takaya 2013). An extreme event is often defined as occurring when verifying analysis is in the tail(s) of the climatological distribution. In this study, we define a threshold of 5th (or -2σ for a normal distribution) and the 95th climatological percentile for extreme cold and extreme precipitation events (high-end only), respectively. The corresponding thresholds are estimated from the 30-year CFSRR climatological data (Saha et al. 2010) and Climatology Calibrated Precipitation Analysis (CCPA) (Hou et al. 2014), respectively.

Similarly, a forecast extreme event is also assessed as a yes if the forecast value is above or below an appropriate threshold value. We use the same threshold as the analysis does to determine an extreme event for the ANF method. The EFI is an integrated measure of the difference between a forecast and its climatology. How to compare these two measures? What EFI value is equivalent (or close) to a specific anomaly? We would like to address this before verification.

Figure 1 shows the comparisons of ensemble mean 96-h ANF and EFI for 2-m temperature on March 01, 2015 over North America. The ANF and EFI were calculated using raw forecasts and model climatology. The corresponding best-fit equation and correlation coefficient are also shown. There is a highly correlated relationship between the two forecasts. We found that a relationship between these two measures could be fitted from the 5th order polynomial function through this sample data set. According to the fitting equation, an EFI value equal to -0.78 is approximately equivalent to a -2σ ANF median (50%) value. This relationship provides an equivalent threshold value for identifying extreme events from the two algorithms and consequently allows corresponding inter-comparisons.

A very similar technique was used to find the two corresponding thresholds for extreme precipitation events. Figure 2 displays a comparison of 72-96 h precipitation ANF and EFI for Jan. 06, 2014 over North America. Similar to the 2-m temperature, ANF and EFI are highly correlated and a 5th order polynomial also best fits the dataset. However, instead of using σ as the ANF unit, here we use percentiles to express precipitation ANF since a normal distribution can not represent the asymmetric character of precipitation. The thresholds for ANF and EFI are taken as 0.95 and 0.685, respectively.

Using these criteria, for each grid point over North America with a coincident model forecast and verifying analysis, one set of yes/no observations for the extreme cold events were assessed. Table 1 incorporates the model and observation into a 2 by 2 contingency table associated with dichotomous forecasts. The quality of the extreme cold event forecast was evaluated based on signal detection theory (Mason, 1982). The statistical scores Hit Rate (HR), False Alarm Rate (FAR), Frequency Bias (FBI), and Equivalent Threat Scores (ETS) (Schaefer 1990) are defined as:

$$HR = A/(A + B) \tag{3}$$

$$FAR = C/(C + D) \tag{4}$$

$$FBI = (A + B)/(A + C) - 1 \tag{5}$$

$$ETS = (A - R(h))/(A + B + C - R(h)) \tag{6}$$

Where,

$$R(h) = (AD - BC)/(A + B + C + D) \tag{7}$$

A perfect forecast is defined by HR=1, FAR=0, FBI=0, and ETS=1. These scores are applied widely in weather forecast evaluations (Swets 1988; Doswell et al. 1990; Zhu 2007).

For ease of interpreting the statistics, Roebber (2009) developed a performance diagram that shows POD (or HR), success ratio (SR), bias, and Critical Success Index (CSI) in a single diagram. Here CSI and SR are defined as:

$$CSI = A / (A + B + C) \tag{8}$$

$$SR = A / (A+C) \tag{9}$$

In Section 4.2, we also use a performance diagram to display the verification results for extreme cold events.

4.2 Verification of extreme cold event forecasts

Using the verification methodology developed in Section 4.1, we compare the performance of the ANF and EFI products in forecasting extreme cold events for different model versions, references, and forecasts. We also examine how using different analysis climatology (NCEP/NCAR 40-year reanalysis vs 30-year CFSRR) impacts the verification.

For 2-m temperature, verification is performed over North America for 11 extreme-cold days (events) that occurred during the winter of 2013-2014. This winter was considered to be colder and snowier than normal as noted in Geert et al., (2015) and the National Weather Service seasonal review (http://www.weather.gov/cle/climate_winter_2013-14_Review). We focus on the two winter cold waves, which occurred for the periods of December 6-10, 2013 and December 29, 2013 – January 7, 2014, respectively. Both cold waves caused extreme cold temperatures and broke daily precipitation and snowfall records across a considerable area of North America.

a. Verification of ANF and EFI products

Figure 3 shows verification of the EFI (Fig.3b) and ANF (Fig.3c) products against observations (Fig.3a) over North America for the GEFS v11 raw forecasts for 0000 UTC on 5 March 2015. The four corresponding statistical scores are also shown at the bottom of the figure. Both EFI and ANF reproduce the observed cold anomaly pattern over the central United States. The HR (0.81) and ETS (0.6) values for the EFI are slightly higher than those for the ANF (HR

(0.8) and ETS (0.58)). The EFI predicts more extreme cold events than the ANF based on the FBI comparison. This may explain why EFI has a slightly higher HR and ETS. There are very similar FAR values (~0.03) for both methods. The very low FAR value mainly results from the combination of a large domain and a small area occupied by the extreme cold event. In addition, the model accurately identifying the extreme cold area is another reason for the low FAR.

b. Verification of raw and bias-corrected forecast products

The verification results for the EFI products from the v11 raw and bias-corrected forecasts are displayed in Fig.4. The forecasts are initiated at 0000 UTC 2 Jan. 2014. Both the raw and bias-corrected forecasts predict extreme cold weather over Canada. However, there is also some difference between the two sets of forecasts. The bias-corrected forecast predicts observed extreme weather over Mexico, which is completely missed by the raw forecast. Based on the verification scores, the bias-corrected forecast performs much better than the raw forecasts for this particular case. The HR and ETS reach 0.76 and 0.6, respectively, for the bias-corrected forecast, which is much higher than in the raw forecast (0.53 and 0.40). The number of extreme cold events from the bias-corrected forecast is very similar to the observed number, which is approximately 20% higher than the raw forecast. The FAR values, again, are very low for both cases. The verification with a larger sample size (11 cases) for both methods is displayed in Fig. 5. It can be seen that increasing the sample size does not change the conclusions. The relative performance of raw and bias-corrected forecasts in the ANF is also very similar to the EFI. Both methods demonstrate much better performance for the bias-corrected than the raw forecast.

Figure 6 is the performance diagram for the above cases. A perfect forecast should have all 4 measures (HR, SR, bias, and CSI) equal to 1. In other words, a good forecast is closer to the upper right corner of the diagram. Obviously, the dots for the bias-corrected forecasts are more concentrated in the upper right than the raw forecasts. Overall, the bias-corrected ANF for entire dataset marked by the green circles is closest to the bias=1 (bias free) line.

One possible explanation of the lower scores for the raw forecast is that the control-only reforecast climatology may not fully represent the model climatology very well. In particular, the produced variance does not completely include model uncertainty. Therefore, the model climatological forecast distribution (or variance) could be incomplete, especially for the tail of a climatological distribution. The impact of ensemble size on the probability forecast has been investigated in Buizza and Palmer (1998) and Ma et al. (2012). An increase in ensemble size is strongly beneficial to the forecast when the size is fewer than 40 members. An effort is being made to create a model climatology using multi-member reforecast runs. This would provide more robust model climatology and improve extreme weather forecasts.

c. Verification of v10 and v11 forecast products

Figure 7 shows the verification for the GEFS v10 and v11 bias-corrected forecasts for 0000 UTC on 2 Jan. 2014 using the EFI method. In general, the both model versions capture the observed major extreme cold regions. But there are also some differences between the two versions. For this particular case, the v11 forecasts have a similar number of extreme cold events as the observations, with a FBI approximately equal to 0, while the v10 underestimates the number of extreme cold events and the FBI value is about -0.26. The v11 version has a higher HR but the ETS is slightly smaller when compared to v10, and has a large negative frequency bias.

The 11-day statistics are shown in Fig.8. Overall, v11 performs better than the v10 version with a higher HR and ETS value. The v11 predicts more extreme events than are observed, while the v10 version underestimates the number of events. The ANF for the new version has the highest ETS and closest match to the observations. The advantage of v11 over v10 can be also demonstrated in the performance diagram (Fig.9). Overall, v11 is closer to the upper right corner. This suggests that the current model upgrade has a more accurate 2-m temperature forecast (Zhu 2015) and a positive impact of extreme cold prediction.

d. Verification of forecast products for different reference climatology

The current NCEP GEFS ANF product uses the 40-year reanalysis as reference climatology. To test the sensitivity of ANF and EFI skills to their reference we make verification comparisons with two different references (30-year CFSRR and 40-year reanalysis) in Fig. 10 and Fig.11. The ANF and EFI calculated relative to the CFSRR climatology have slightly better HR, FBI and ETS than the reanalysis climatology (Fig.10). The relative forecasting performance with the two references can be also identified from the performance diagram (Fig.11). The plotted positions for the CFSRR reference are closer to the upper right corner than for the reanalysis reference, indicating a slightly higher accuracy when a more sophisticated analysis is used. The sensitivity of the verification scores to the references for the ANF and EFI is very similar. The differences in HR and FBI caused by using different references (Fig.10) are less important compared to differences from the different model versions (Fig.8). But the sensitivity of ETS to the model version and reference are roughly similar.

4.3 Verification of heavy precipitation forecasts

Figure 12 shows the 96-h forecasts of extreme precipitation regions from the (a) ANF and (b) EFI products, initiated at 0000 UTC 6 Jan. 2014. The shaded areas are the corresponding 72-

96h accumulated precipitation forecasts. Both products forecast the two major extreme precipitation regions, located over Baffin Island and from the Gulf of Mexico to the Atlantic Ocean, respectively. Overall, the patterns of extreme precipitation from the two products are very similar. The definition of extreme precipitation depends on local climatology. The figure illustrates the dependence of extreme precipitation on the geographic location. For example, the strong precipitation region over Washington State and British Columbia is not diagnosed as an extreme precipitation event. Conversely, a relatively weak precipitation area over Baffin Island is predicted as an extreme precipitation event.

Figure 13 compare the two products against the CCPA for another case over the CONUS. The 84-h forecasts of extreme precipitation regions were initiated at 0000 UTC 3 Dec. 2013. Again the forecasts from the two products are very similar and capture the major extreme precipitation region over the United States, although the forecasts underestimate the observed area of extreme precipitation. The verification scores demonstrate that the EFI predicts more extreme events with a slightly higher HR, FAR, and a similar ETS as the ANF. The proposed methodology will be applied to more cases to calculate the statistics of extreme precipitation prediction in the future.

5. Conclusions

In this work, we examine the ANF and EFI algorithms for observed extreme cold temperature and extreme heavy precipitation during the winter of 2013-2014. We develop a verification methodology in order to provide a tool to evaluate the relative performance of products from different methods (ANF and EFI), model versions (GEFS v10 and v11), forecasts (raw and bias-corrected), and different reanalysis climatology as well. We find a strong correlation between the ANF and EFI. For extreme cold event forecasts, an EFI of -0.78 is approximately equivalent to -2σ ANF (or ANF=0.05) and for extreme precipitation forecasts, EFI=0.687 corresponds to ANF=0.95. This provides a threshold to evaluate and compare the two different forecast algorithms.

The verification results show that both the ANF and EFI can predict extreme events. Verification statistics for extreme cold events in the winter of 2013-2014 indicate the EFI forecasts more extreme cold events than the ANF. The ANF produces a higher ETS value. The bias-corrected forecast has much better performance than the raw forecast when an 18-year control only reforecast was used as an approximate reference. This indicates a need for increasing the number of reforecast members to improve the extreme weather forecast. The work towards finding the optimized configuration of real-time GEFS reforecast runs are being conducted (Hamill et al. 2014; Guan et al. 2015). It will provide a better reference for the future applications. We also found that the upgrade of the GEFS model from v10 to v11 has a beneficial impact on the extreme cold weather forecast. Using a more recently developed climatology (CFSRR) as the reference gives a slightly better score than the 40-year reanalysis. A previously developed performance diagram (Roebber 2009) is also used to illustrate the verification results, further proving its usefulness as a visualization tool.

The current work also demonstrates that the verification methodology can be extended to extreme precipitation. We verified an extreme precipitation case that occurred in the winter of 2013-2014. The results indicated a potential wider application of the verification methodology. In the future, we will examine more extreme precipitation cases and calculate long-term statistics. Meanwhile, we will use the methodology to verify surface winds and surface pressure as well. The sensitivity of ANF-EFI relationship on forecast lead time will also be our focus.

Acknowledgements:

The authors would like to thank the members of the ensemble and post-processing team at NCEP/EMC for helpful suggestions and support for this work. Special appreciation goes to Dr. Yan Luo who kindly provided and helped with understanding the CCPA data and Bo Yang who provided the CFSRR data. The authors would also like to acknowledge the helpful advice and discussion of Dr. Bo Cui, Dr. Malaquias Pena, and Dr. Corey Guastini. Dr. Tara Jensen of NCAR is also thanked for providing the R-program used to generate performance diagrams.

References:

Buizza, R. and Palmer, T. N., 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 9, 2503-2518.

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

Doswell, C. A., Daviesjones, R., and Keller, D.L., 1990: On summary measures of skill in rare event forecasting based on contingency tables. Weather and Forecasting, 5, 576–585.

Dutra E, Diamantakis M, Tsonevsky I, Zsoter E, Wetterhall F, Stockdale T, Richardson D, Pappenberger F., 2013: The extreme forecast index at the seasonal scale. *Atmos. Sci. Lett.14, 256–262. (doi:10.1002/asl2.448)*

Geert Jan Van Oldenborgh, Rein Haarsma, Hylke De Vries, and Myles R. Allen, 2015: Cold Extremes in North America vs. Mild Weather in Europe: The Winter of 2013–14 in the Context of a Warming World. Bull. Amer. Meteor. Soc., 96, 707–714. doi: http://dx.doi.org/10.1175/BAMS-D-14-00036.1

Graham, R. A., and R. H. Grumm, 2010: Utilizing normalized anomalies to assess synoptic-scale weather events in the western United States. Wea. Forecasting, 25, 428–445.

Grumm, R. H, 2001: Standardized anomalies applied to significant cold season weather events: Preliminary findings. Wea. Forecasting, 16, 736–754.

Guan H., B. Cui, and Y. Zhu, 2015: Improvement of Statistical Postprocessing Using GEFS Reforecast Information. *Wea. Forecasting*, **30**, 841–854. doi: http://dx.doi.org/10.1175/WAF-D-14-00126.1

Hamill, T. M., 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.

Hamill, T. M., and Coauthors, 2014: A recommended reforecast configuration for the NCEP global ensemble forecast system. NOAA White paper, 24 pp, accessed 9 June 2015. [Available online at http://www.esrl.noaa.gov/psd/people/tom.hamill/White-paper-reforecast-configuration.pdf.]

Hosking, J. R. M., 1990: L -Moments: analysis and estimation of distributions sing linear combinations of order statistics." J. Roy. Stat. Soc. B 52, 105-124, 1990.

Hosking, J. R. M. and Wallis, J. R., 1997: Regional Frequency Analysis, an approach based on Lmoments, Cambridge University Press.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Evaluation of the impact of the stochastic 1 perturbation schemes on global ensemble forecast. Preprints, 19th Conference on Probability 2 and Statistics, New Orleans, LA, Amer. Meteor. Soc.

Hou, D., and Coauthors, 2014: Climatology-Calibrated Precipitation Analysis at Fine Scales: Statistical Adjustment of Stage IV toward CPC Gauge-Based Analysis. *J. Hydrometeor*, **15**, 2542–2557. doi: http://dx.doi.org/10.1175/JHM-D-11-0140.1

IPCC, 2011: IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation. Prepared by Working Group III of the Intergovernmental Panel on Climate Change [O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, C. von Stechow (eds)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1075 pp.

Kalnay, E., Kanamitsu M, Kirtler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White

G, Woollen J, Zhu Y, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Leetma A, Reynolds R, Jenne R, Joseph D. 1996: The NCEP/NCAR 40-year reanalysis project. Bull. Amer. Meteorol. Soc. 77: 437–471.

Kleist, D. T. and K. Ide, 2015: An OSSE-Based Evaluation of Hybrid Variational–Ensemble Data Assimilation for the NCEP GFS. Part I: System Description and 3D-Hybrid Results. *Mon. Wea. Rev.*, **143**, 433–451. doi: http://dx.doi.org/10.1175/MWR-D-13-00351.1.

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

Ma, J. H., Y. J. Zhu, R. Wobus, and P. X. Wang, 2012: An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. Adv. Atmos. Sci., 29(4), 782–794, doi:10.1007/s00376-012-1249-y.

Mason, I. B., 1982: A model for the assessment of weather forecasts. Aust. Meteo. Mag., 30, 291-303.

Matsueda, S. and Y. Takaya, 2013: Verification of the Extreme Forecast Index in JMA's operational one-month ensemble prediction system, CAS/JSC WGNE research activities in atmospheric and oceanic modelling, WGNE Blue Book, 2013. Avail. www.wcrpclimate.org/WGNE/BlueBook/2013/individualarticles/06_Matsueda_Satoko_EFI_Verification.pdf.

Petroliagis, T. I. and P. Pinson, 2012: Early indication of extreme winds utilizing the extreme forecast index. ECMWF Newsletter, 132, 13 – 19.

Roebber, P.J., 2009: Visualizing multiple measures of forecast quality. Wea. Forecasting, 24, 601-608.

Saha, S. and co-authors, 2010: The NCEP climate forecast system reanalysis. Bull. Amer. 11 Meteor. Soc., 91, 1015–1057.

Sela, J., 2010: The derivation of the sigma pressure hybrid coordinate Semi-Lagrangian model equations for the GFS, Office note 462. Avail. http://www.lib.ncep.noaa.gov/ncepofficenotes/files/on462.pdf

Schaefer, J. T., 1990: The critical success index as an indicator of forecasting skill. Wea. Forecasting, 5, 570-575.

Swets, J. A., 1988: Measuring the accuracy of diagnostic systems. Science, 240, 1285–1293.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: "Probability and Ensemble Forecasts." In book of: Forecast Verification: A practitioner's guide in atmospheric science. Ed.: I. T. Jolliffe and D. B. Stephenson. Wiley, 137-163.

Wang, X., D. Parrish, D. Kleist, and J. Whitaker, 2013: GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP Global Forecast System: Single-resolution experiments. *Mon. Wea. Rev.*, **141**, 4098–4117 (doi: 10.1175/MWR-D-12-00141.1).

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble 4 transform (ET) technique in the NCEP global operational forecast system. Tellus, 60A, 62–79.

Whitaker, Jeffrey S., Thomas M. Hamill, Xue Wei, Yucheng Song, Zoltan Toth, 2008: Ensemble Data Assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, 136, 463–482.

Whitaker, J. S. and T. M. Hamill, 2012: Evaluating methods to account for system errors in ensemble data assimilation. *Mon. Wea. Rev.*, 140, 3078-3089.

Zhou, X., Y. Zhu, D. Hou, Y. Luo, J. Peng, and R. Wobus, 2016: The NCEP Global Ensemble Forecast System with the EnKF Initialization. Submitted to MWR

Zhu, Y., 2007: Objective evaluation of global precipitation forecast, In special collection of: International Symposium on Advances in Atmospheric Science and Information Technology, Beijing, China, p3-8.

Zhu Y. and B. Cui, 2007: Web documentation, Dec. 4, 2007. http://www.emc.ncep.noaa.gov/gmb/yzhu/imp/i200711/3-Mean_spread.pdf

Zhu Y., R. Wobus, M. Wei, B. Cui, and Z. Toth, 2007: March 2007 NAEFS upgrade. [Available online at http://www.emc.ncep.noaa.gov/gmb/ens/ens_imp_news.html.]

Zhu, Y., et al., 2012: NCEP Implementation Documentation, February 14 2012: http://www.emc.ncep.noaa.gov/gmb/yzhu/html/imp/201109_imp.html

Zhu, Y., 2015: GEFS upgrade (V11). October 6, 2015: http://www.emc.ncep.noaa.gov/gmb/yzhu/imp/i201412/GEFS_sci_briefing.pdf.

Zsótér, E., 2006: Recent developments in extreme weather forecasting. ECMWF Newsletter, 107, 8 – 17.

Table:

Table 1. The contingency table used to evaluate forecasts of extreme cold events.

Figure Captions:

Figure 1. Comparisons of the ensemble mean ANF and EFI for the 96-hr 2m temperature forecast over North America. The raw foreacst and model climatology are used in producing the ANF and EFI. The solid line represents the best-fit curve. The foreacsts are initaited at 0000 UTC 1 Mar. 2015.

Figure 2. Comparison of the 50th percentile ANF and EFI for accumulated precipitation forecasts (72-96hr) over North America. The v11 raw foreacst and model climatology are used in producing the ANF and EFI. The solid line represents the best-fit curve. The foreacsts are initaited at 0000 UTC 6 Jan. 2014.

Figure 3. Extreme cold weather event observations (a), 96-hr EFI forecast (b), 96-hr ANF (c), and verification for both methods (d). The v11 raw forecast and v11 model climatology are used in producing the ANF and EFI. The foreacsts are initaited at 0000 UTC 5 Mar. 2015.

Figure 4. Extreme cold weather event observations or anomaly analysis (ANA) (a), 96-h raw EFI forecast (b), 96-h bias-corrected EFI forecast (c), and verification for the v11 RAW and v11 bias-corrected forecast (d). The 18-year control-only and CFSRR climatology are used in producing the raw and bias-corrected forecast products, respectively. The forecasts are initiated at 0000 UTC 2 Jan. 2014.

Figure 5. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different

algorithms (EFI and ANF) and forecasts (raw and bias-corrected) over North America. Blue and red bars are the v11 raw ANF and EFI, respectively; green and purple bars are the v11 bias-corrected ANF and EFI, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle.

Figure 6. Performance diagram summarizing the SR, POD, bias, and CSI. Solid and dashed lines represent CSI and bias scores, respectively. Shown are 96-h forecasts of extreme cold weather for 11 indiviual days from the raw ANF (blue dots), raw EFI (red dots), bias-corrected ANF (green dots), and bias-corrected EFI (purple dots). The four circles denote the corresponding 11-day scores.

Figure 7. Extreme cold weather event observations (a), EFI product from v10 (b) and v11 (c) 96-h bias-corrected foreacsts, and verification for both of model versions (d). The forecasts are initiated at 0000 UTC 2 Jan. 2014 and the reference climatology is CFSRR.

Figure 8. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms (ANF and EFI) and model versions (v10 and v11) over North America. Blue and red bars are the v10 bias-corrected ANF and EFI, respectively; green and purple bars are the v11 bias-corrected ANF and EFI, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle and the reference climatology is CFSRR.

Figure 9. Performance diagram as in Fig.6, but for the comparisons of the two model versions. Blue and red dots are the v10 bias-corrected ANF and EFI, respectively; green and purple dots are the v11 bias-corrected ANF and EFI, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle and the reference climatology is CFSRR.

Figure 10. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms and reference climatology over North America. Blue and red bars are the v11 bias-corrected ANF with NCEP/NCAR reanalysis and CFSRR as reference, respectively; green and purple bars are the v11 bias-corrected EFI with NCEP/NCAR reanalysis and CFSRR as reference, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle.

Figure 11. Performance diagram as in Fig. 6, but for the comparisons of the two reference climatology (30-year CFSR and 40-year reanalysis). Blues and green dots are ANF and EFI using the 40-year reanalysis as the reference; red and purple dots are ANF and EFI using the 30-year CFSRR as the reference.

Figure 12. The 96-h forecasts of extreme precipitation regions (red contours) from the ANF (a) and EFI products (b). The shaded areas are corresponding 72-96hr accumulated precipitation forecasts (mm). The contours in (a) and (b) represent ANF=0.95 and EFI=0.687, respectively. The forecasts are initiated at 0000 UTC 6 Jan. 2014.

Figure 13. The daily extreme precipitation distribution (60-84hr) for ANA (a), ANF (b), and EFI forecast (c), and verification for both of methods (d). The v11 forecasts are initiated at 0000 UTC 3 Dec. 2013.

Table 1. The contingency table used to evaluate forecasts of extreme cold events.

	Yes forecast	**No forecast**	**Total**
Yes observed	A	B	A+B
No observed	**C**	**D**	**C+D**

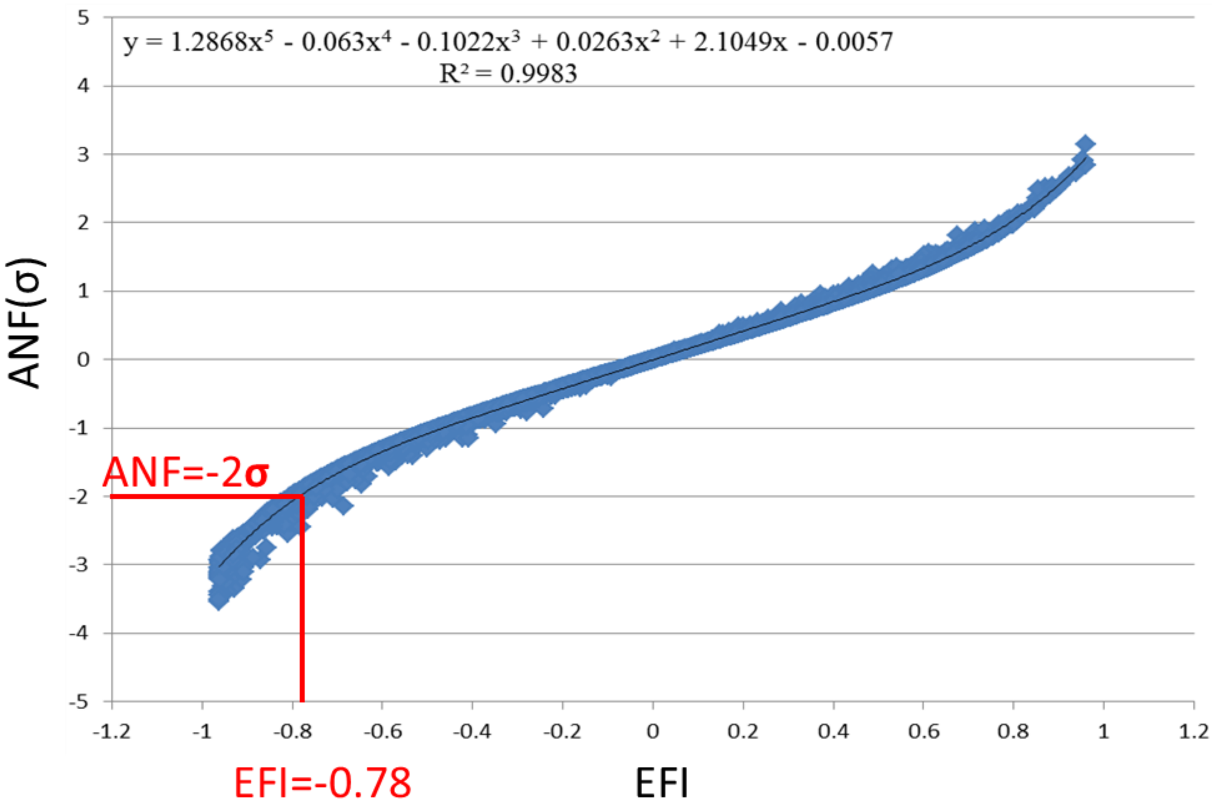

Figure 1. Comparisons of the ensemble mean ANF and EFI for 96-hr 2m temperature forecast over North America. The raw foreacst and model climatology are used in producing the ANF and EFI. The solid line represents the best-fit curve. The forecasts are initiated at 0000 UTC 1 Mar. 2015.

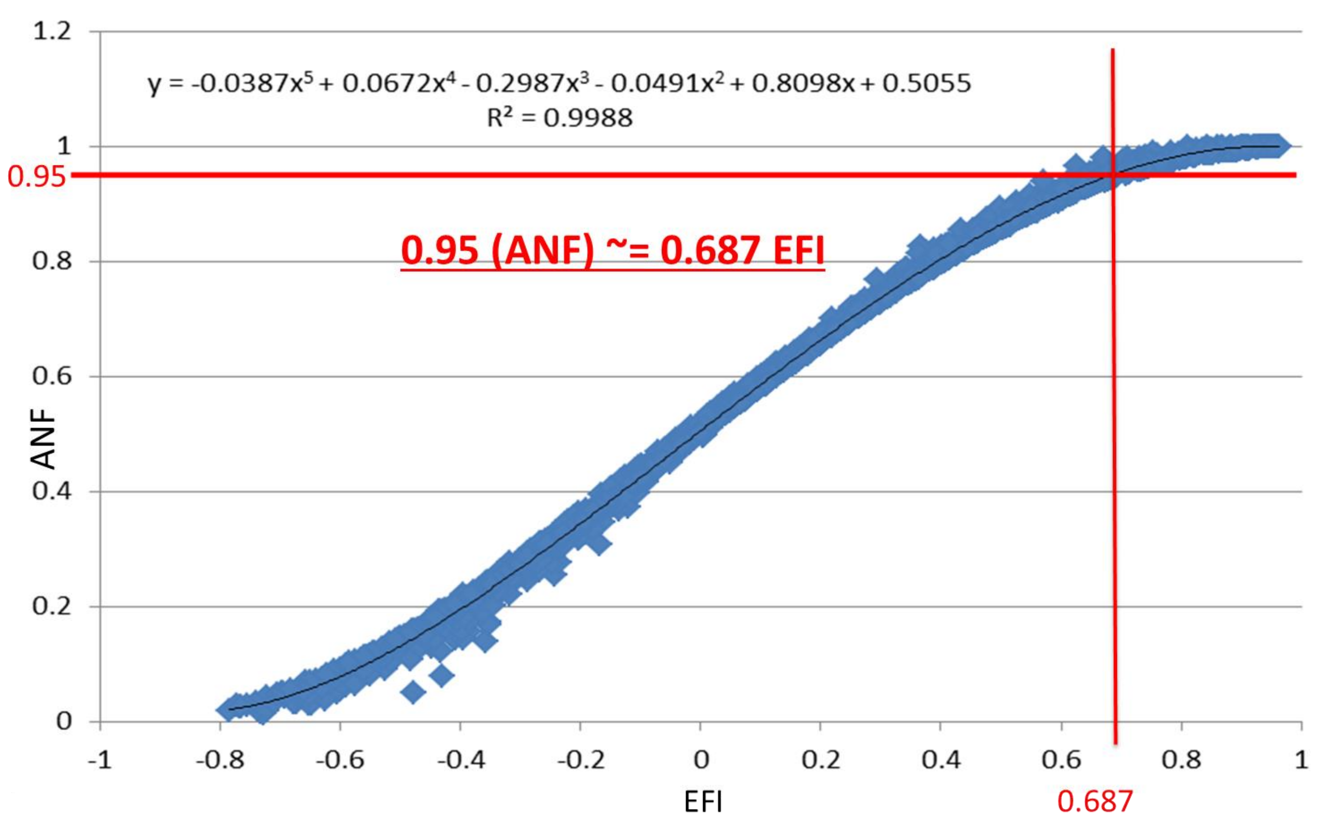

Figure 2. Comparison of the 50th percentile ANF and EFI for accumulated precipitation forecasts (72-96hr) over North America. The v11 raw foreacst and model climatology are used in producing the ANF and EFI. The solid line represents the best-fit curve. The foreacsts are initaited at 0000 UTC 6 Jan. 2014.

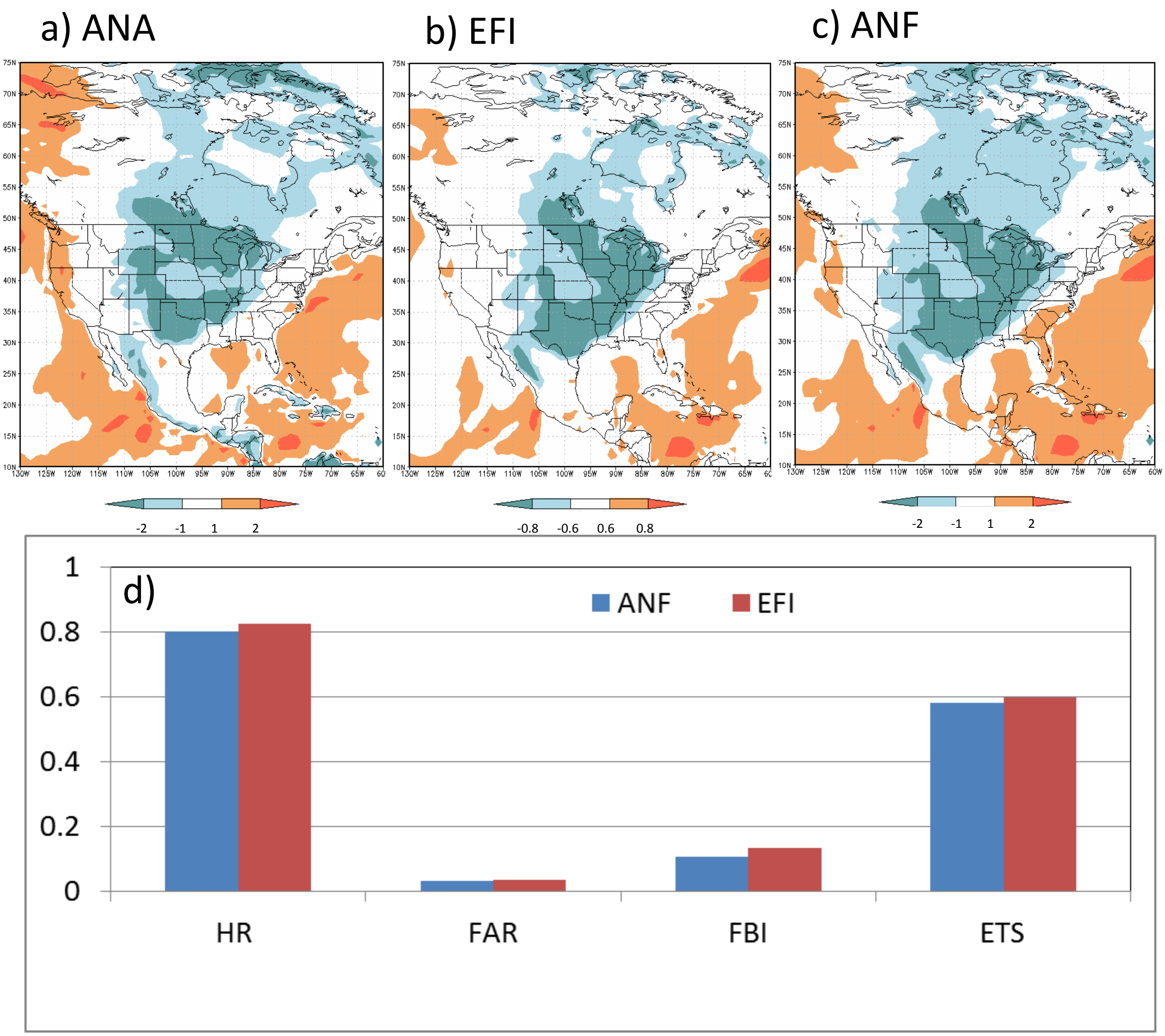

Figure 3. Extreme cold weather event observations or anormaly analysis (ANA) (a), EFI forecast (b), ANF (c), and verfication for both of methods (d). The v11 raw forecast and v11 model climatology are used in producing the ANF and EFI. The forecasts are initiated at 0000 UTC 5 Mar. 2015.

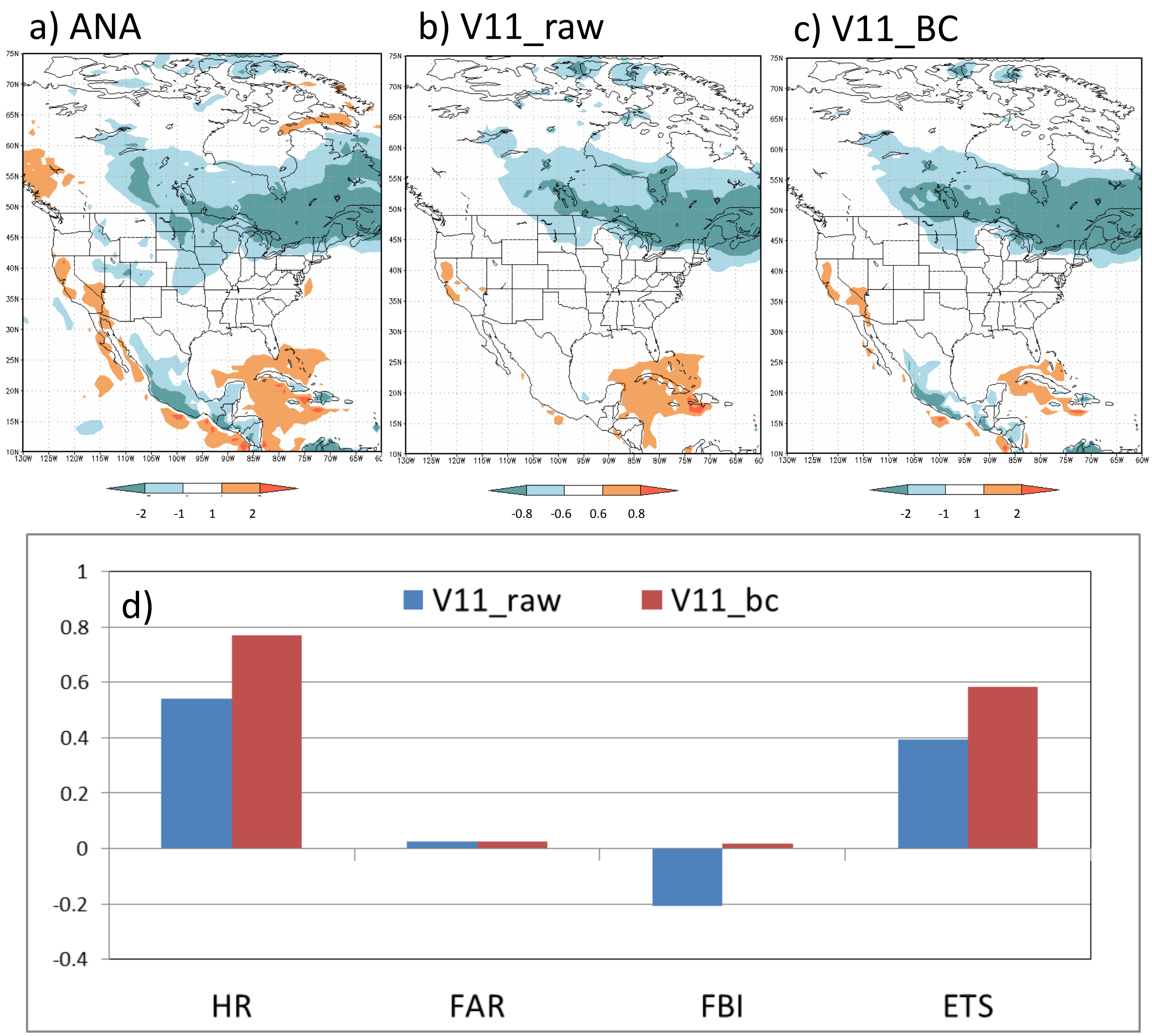

Figure 4. Extreme cold weather event observations or anomaly analysis (ANA) (a), 96-h raw EFI forecast (b), 96-h bias-corrected EFI forecast (c), and verification for the v11 raw and v11 bias-corrected forecasts (d). The 18-year control-only and CFSRR climatology are used in producing the raw and bias-corrected forecast products, respectively. The forecasts are initiated at 0000 UTC 2 Jan. 2014.

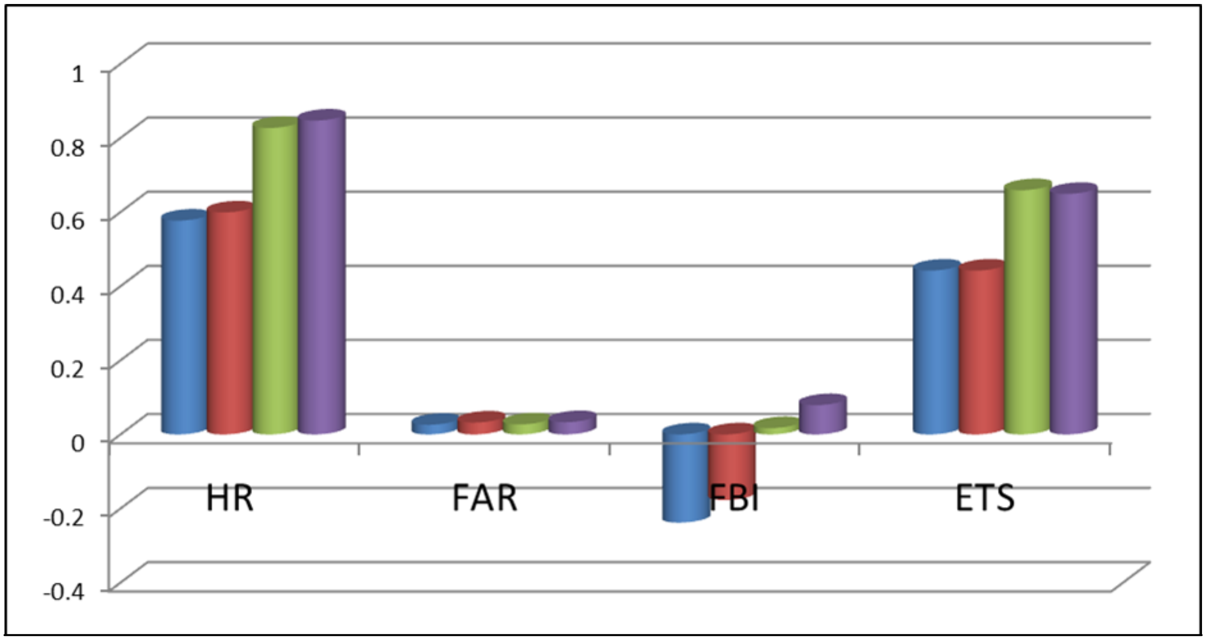

Figure 5. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms (EFI and ANF) and forecasts (raw and bias-corrected) over North America. Blue and red bars are the v11 raw ANF and EFI, respectively; green and purple bars are v11 bias-corrected ANF and EFI, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle.

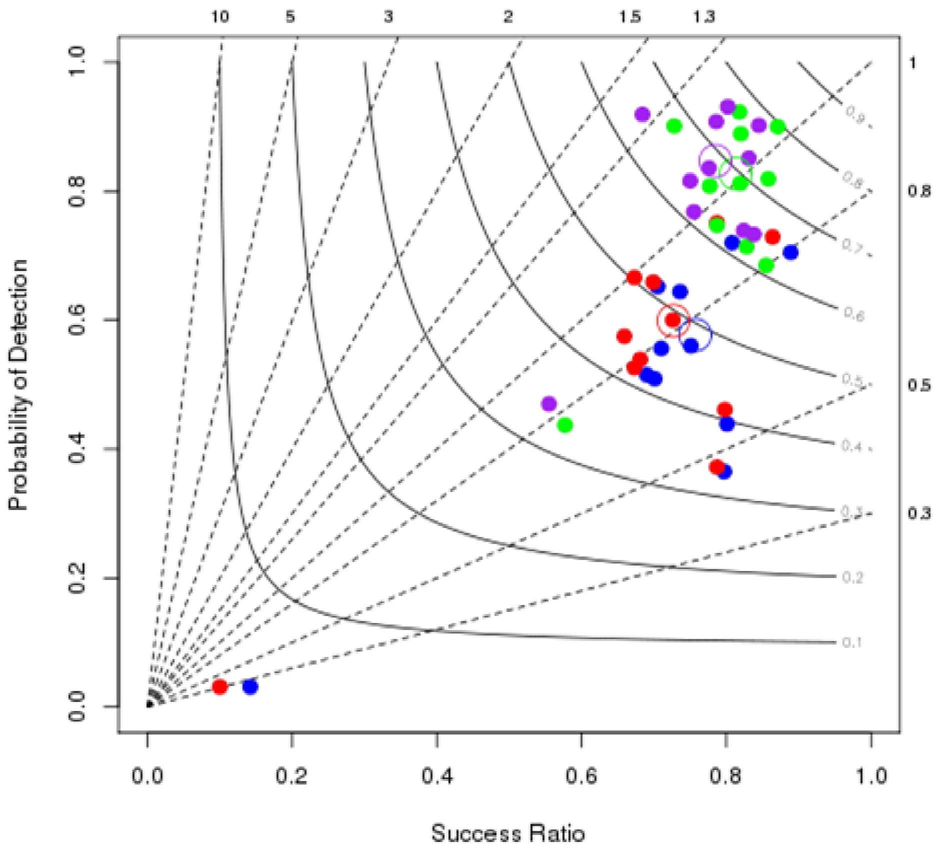

Figure 6. Performance diagram summarizing the SR, POD, bias, and CSI. Solid and dashed lines represent CSI and bias scores, respectively. Shown are 96-h forecasts of extreme cold weather for 11 indiviual days from the raw ANF (blue dots), raw EFI (red dots), bias-corrected ANF (green dots), and bias-corrected EFI (purple dots). The four circles denote the corresponding 11-day scores.

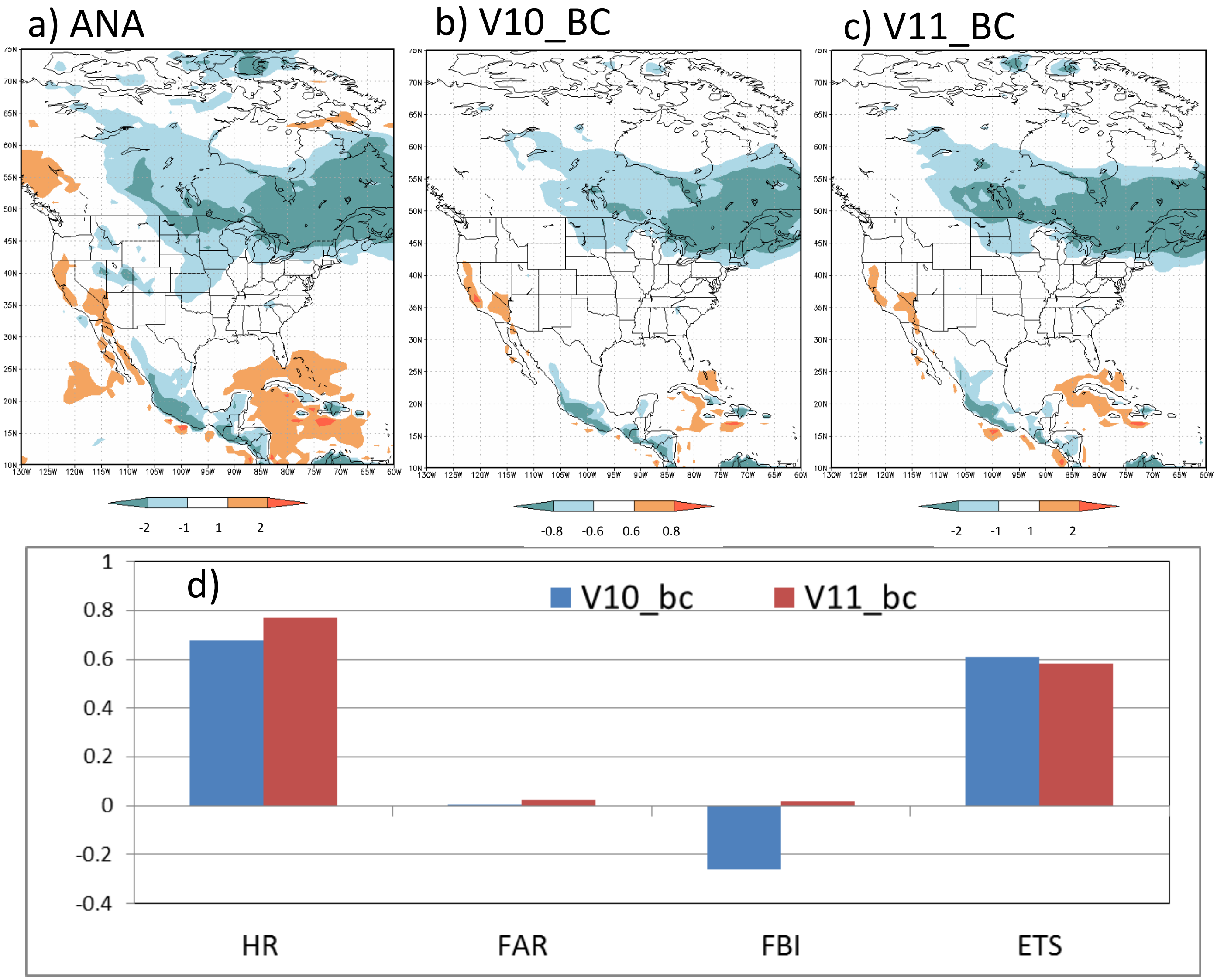

Figure 7. Extreme cold weather event observations (a), EFI product from v10 (b) and v11 (c) 96-h bias-corrected foreacsts, and verification for both of model versions (d). The forecasts are initiated at 0000 UTC 2 Jan. 2014 and the reference climatology is CFSRR.

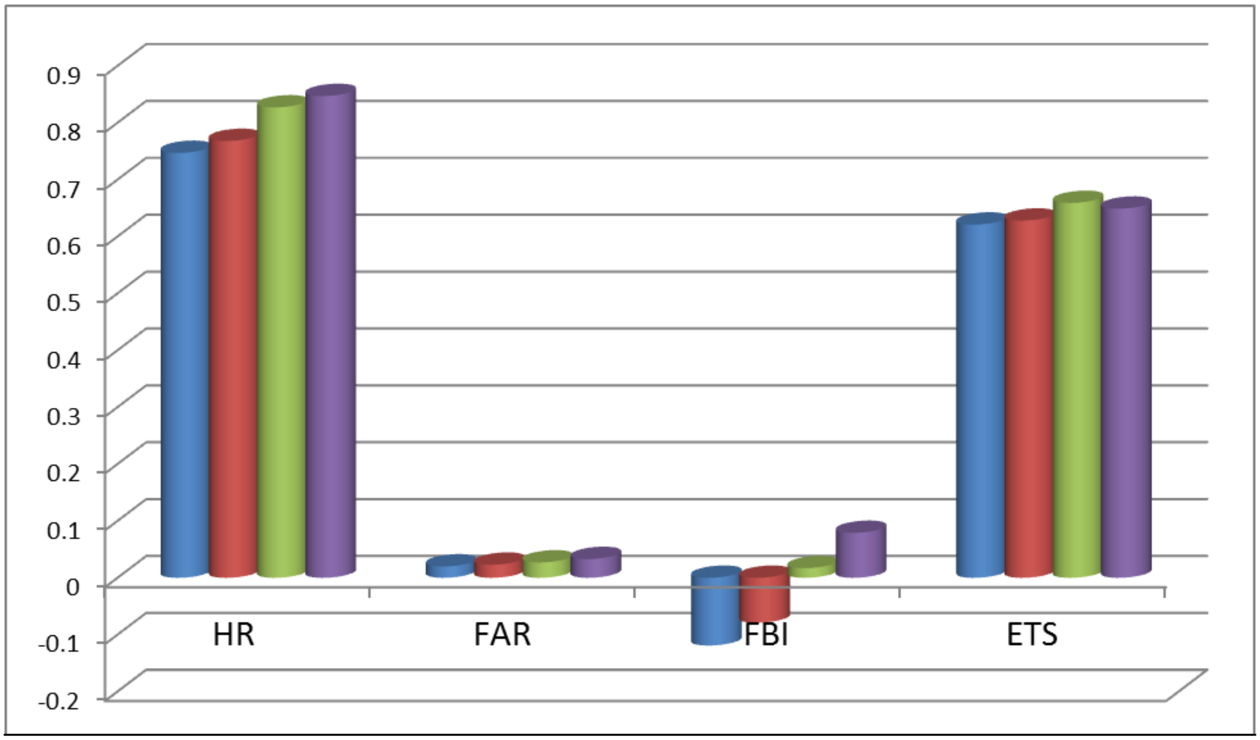

Figure 8. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms (ANF and EFI) and model versions (v10 and v11) over North America. Blue and red bars are the v10 bias-corrected ANF and EFI, respectively; green and purple bars are the v11 bias-corrected ANF and EFI, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle and the reference climatology is CFSRR.

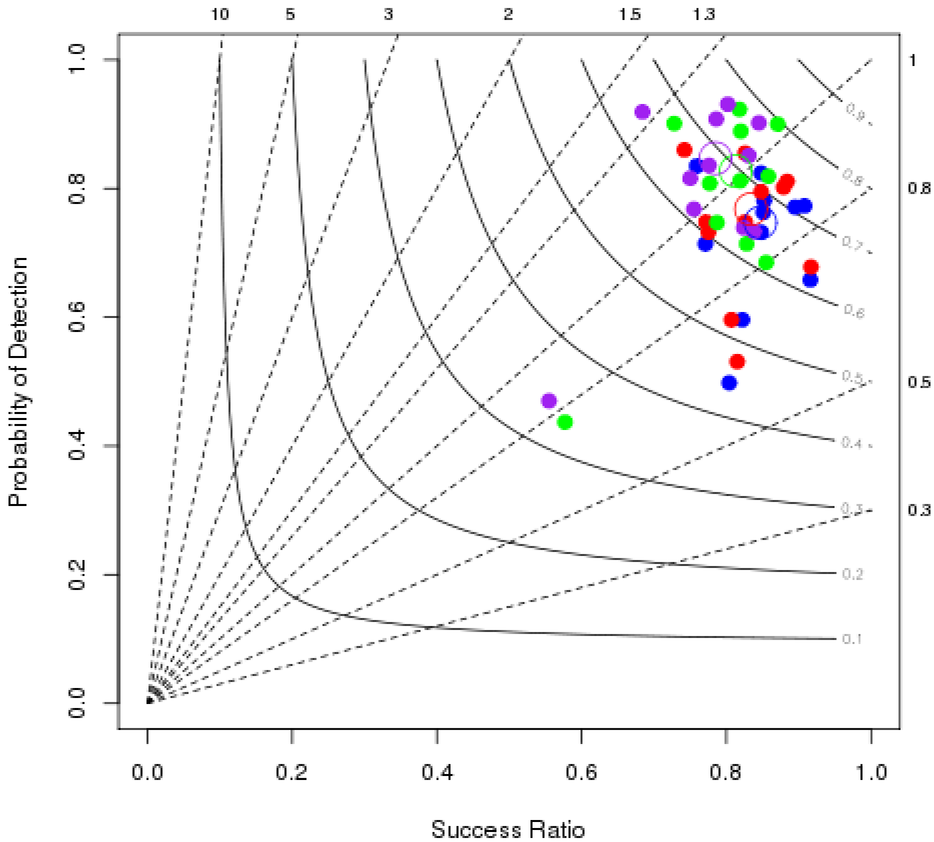

Figure 9. Performance diagram as in Fig.6, but for the comparisons between the two model versions. Blue and red dots are the v10 bias-corrected ANF and EFI, respectively; green and purple dots are the v11 bias-corrected ANF and EFI, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle and the reference climatology is CFSRR.

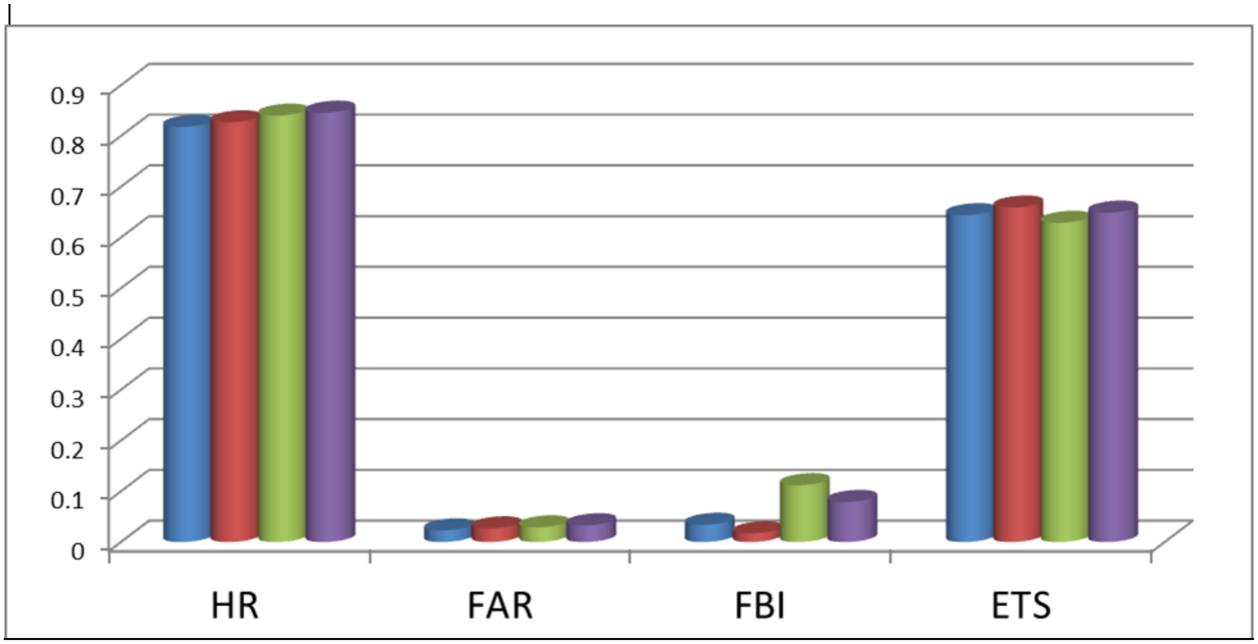

Figure 10. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms and reference climatology over North America. Blue and red bars are the v11 bias-corrected ANF with NCEP/NCAR reanalysis and CFSRR as reference, respectively; green and purple bars are the v11 bias-corrected EFI with NCEP/NCAR reanalysis and CFSRR as reference, respectively. All forecasts are 96-h forecasts from 0000 UTC cycle.

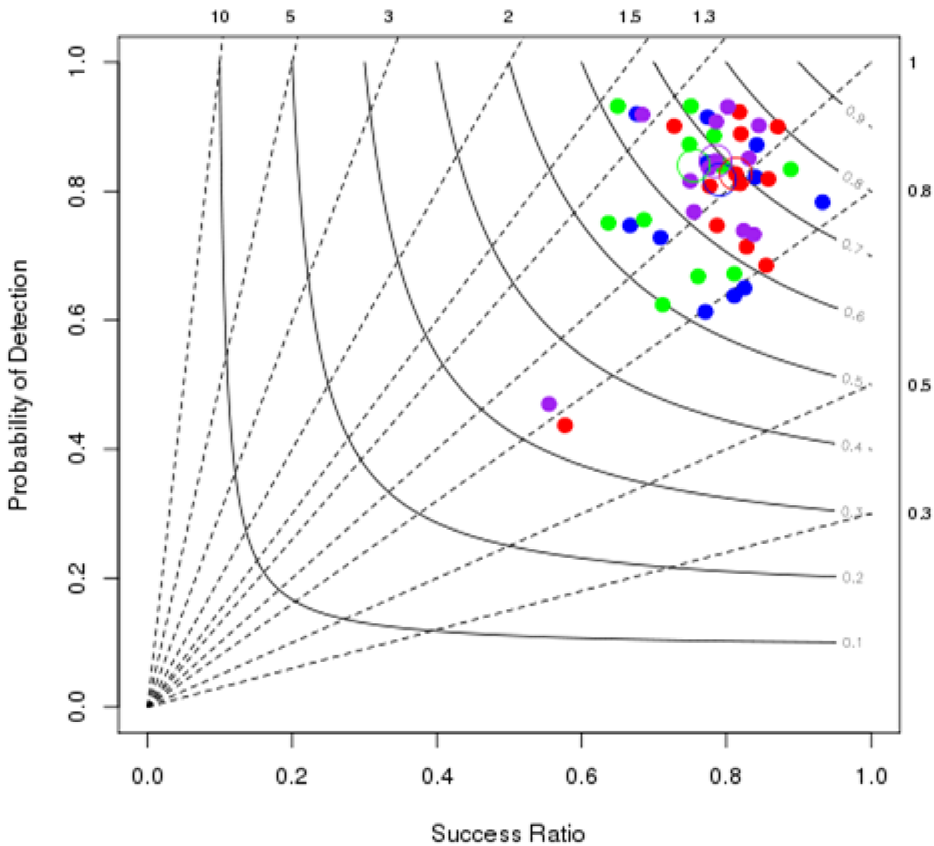

Figure 11. Performance diagram as in Fig. 6, but for the comparisons between the two reference climatology (30-year CFSR and 40-year reanalysis). Blue and green dots are ANF and EFI using the 40-year reanalysis as the reference; red and purple dots are ANF and EFI using the 30-year CFSR as the reference. All forecasts are 96-h forecasts from 0000 UTC cycle.

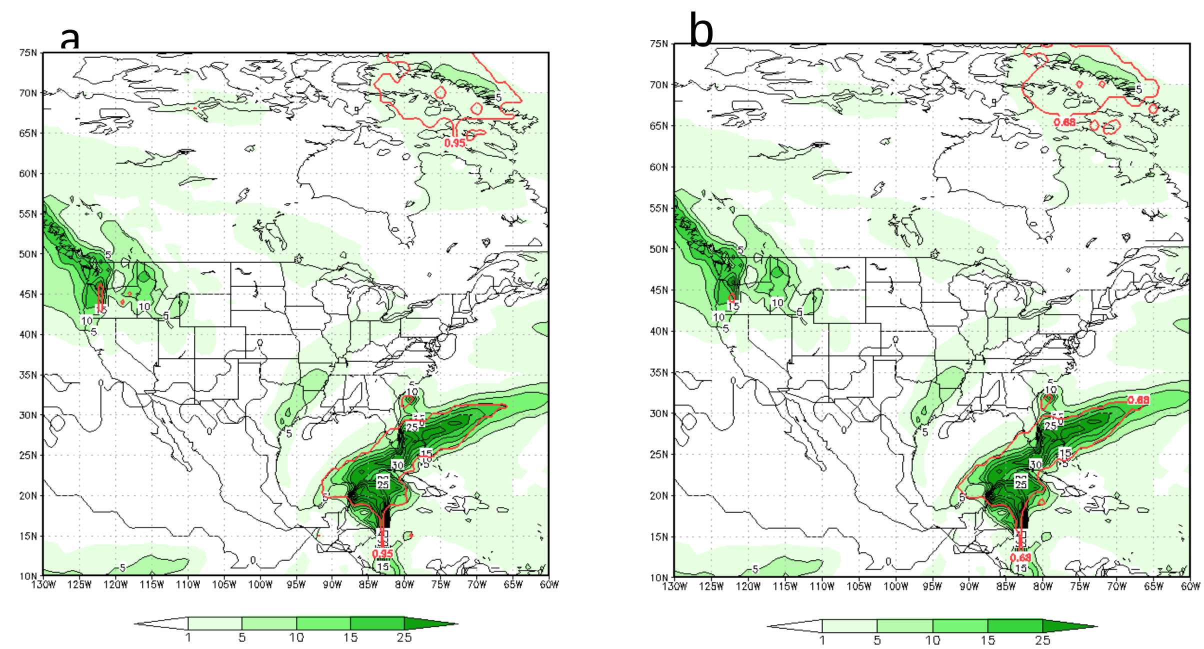

Figure 12. The 96-h forecasts of extreme precipitation regions (red contours) from the ANF (a) and EFI products (b). The shaded areas are the corresponding 72-96hr accumulated precipitation forecasts (mm). The contours in (a) and (b) represent ANF=0.95 and EFI=0.687, respectively. The forecasts are initiated at 0000 UTC 6 Jan. 2014.

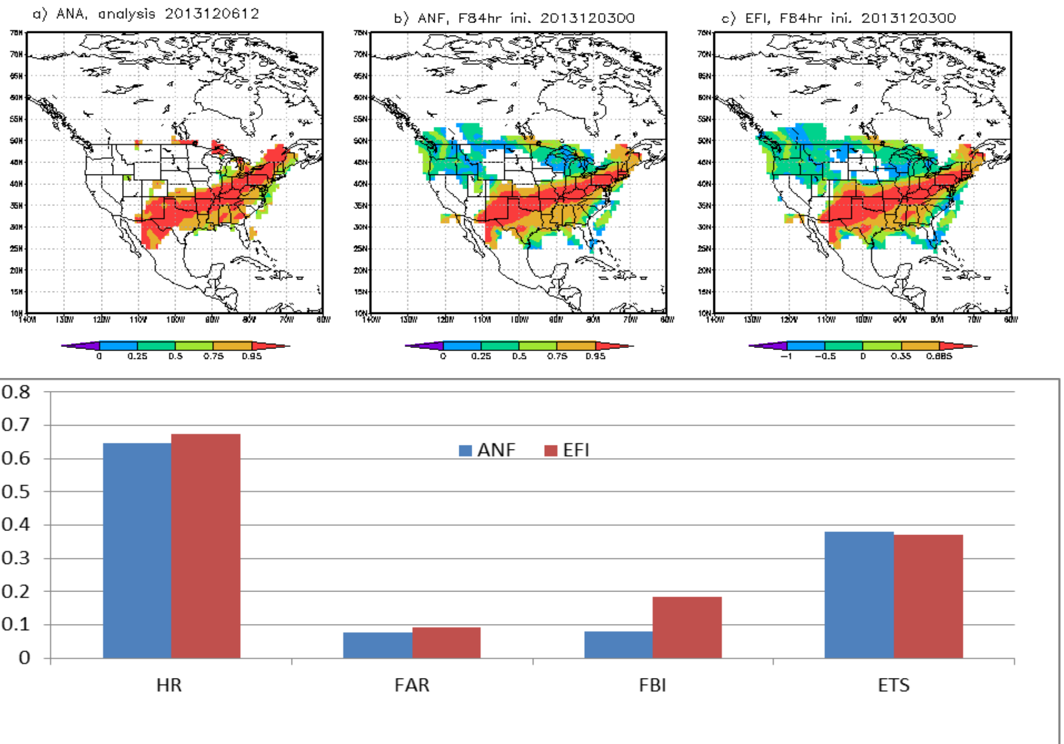

Figure 13. The daily extreme precipitation distribution (60-84hr) for ANA (a), ANF (b), and EFI forecast (c), and verification for both of methods (d). The v11 forecasts are initiated at 0000 UTC 3 Dec. 2013.