Development of Verification Methodology for Extreme Weather Forecasts

HONG GUAN

NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland, and System Research Group Inc., Colorado Springs, Colorado

YUEJIAN ZHU

NOAA/NWS/NCEP/Environmental Modeling Center, College Park, Maryland

(Manuscript received 1 July 2016, in final form 23 November 2016)

ABSTRACT

In 2006, the statistical postprocessing 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 forecasting (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 2013/14. 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 2013/14 indicate that ANF forecasts more extreme cold events with a slightly higher skill than 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-yr 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 a climatological base, a 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 billion U.S. dollars 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 to make early decisions and improve protection against risks.

A number of forecast methods have been developed and applied in identifying extreme weather events at various world forecast centers (Zhu and Cui 2007; Lalaurette 2003; Zsótér 2006; Dutra et al. 2013; 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 a model climate distribution. To increase the sensitivity of forecasts of extreme events, this index was further adapted in 2006 (Zsótér 2006) 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), the Canadian Meteorological Centre (CMC), and the Earth System Research Laboratory (ESRL) of the National and Oceanic and Atmospheric administration (NOAA).

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

DOI: 10.1175/WAF-D-16-0123.1

Anomaly forecasting (ANF) is a more natural method for forecasting extreme weather events. It measures the 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-yr reanalysis, a daily climatological distribution [probability distribution function (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) the 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-, 2-, and 3-sigma standard deviation ANFs for various meteorological elements. Furthermore, by comparing the forecast PDF with the climatological PDF, the users could easily identify an extreme weather event.

In this paper, we develop a verification methodology for comparing and evaluating the extreme weather forecast products from the ANF and EFI. After explaining the verification metrics, we evaluate products from different model versions (or model upgrades), different references, and products based on a raw forecast and bias-corrected forecast. We first introduce the model and datasets 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 datasets

In this study, Global Ensemble Forecast System (GEFS) version 10 (v10) (Zhu et al. 2012) and v11 (Zhou et al. 2016, manuscript submitted to *Mon. Wea. Rev.*) 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-yr control-only reforecast dataset.

The GEFS v10 was implemented on 14 February 2012 at NCEP. It consists of 21 members (one control member and 20 perturbed members) and is run four 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 perturbation (STTP) method (Hou et al. 2008). For the bias-corrected dataset, the model bias was removed using a decaying averaging postprocessing technique (Cui et al. 2012).

There are three major changes from v10 to v11. First, in v11, Euler's integration method is replaced by the semi-Lagrangian method in order to save computing time (Sela 2010). Second, the EnKF 6-h forecast is used as the basis for the ensemble initial perturbations instead of BV-ETR generation. The details of the EnKF technique can be found in references cited by Whitaker and Hamill (2012), Whitaker et al. (2008), Wang et al. (2013), and Kleist and Ide (2015). Third, the horizontal resolutions were 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-yr (1995–2012) control-only v11 reforecast was run at the 0000 UTC cycle every other day. The reforecast dataset was interpolated bilinearly to $1^\circ \times 1^\circ$ latitude–longitude grids from the native resolutions. The model native resolutions are about 34 and 55 km at midlatitudes for the first and last 8 days, respectively. From the $1^\circ \times 1^\circ$ dataset, the model climatology for each day and each grid point was generated. In calculating the climatology, we also include eight 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 ($9\,\text{yr} \times 3\,\text{day}\,\text{yr}^{-1} \times 9$ points) for each grid point.

The analysis climatology of 2-m temperature includes NCEP–NCAR 40-Year (1959–98) Reanalysis dataset (Kalnay et al. 1996) and NCEP's Climate Forecast System Reanalysis and Reforecast (CFSRR) 30-yr reanalysis dataset (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 the improved quality of the observations, a state-of-the-art model and assimilation system, and much higher spatial

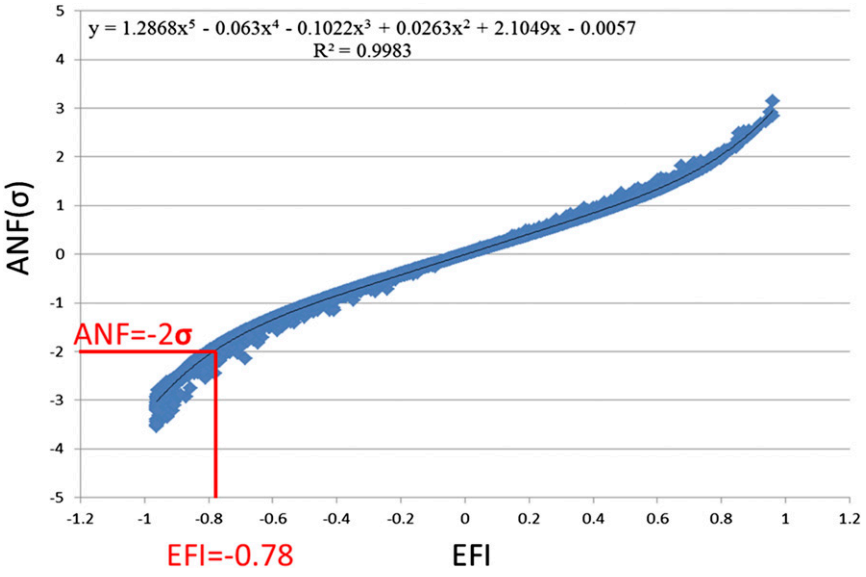

FIG. 1. Comparisons of the ensemble mean ANF and EFI for the 96-h 2-m temperature forecast over North America. The raw forecast and model climatology are used in producing the ANF and EFI. The solid line represents the best-fit curve. The forecasts are initiated at 0000 UTC 1 Mar 2015.

resolution. It has been pointed out that for the near-surface temperature the CFSRR produces a much finer structure than the NCEP–NCAR reanalysis (B. Yang 2015, personal communication).

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 reanalyses, the first four Fourier modes (higher smoothing) have been used to generate daily climatological means to include annual, semiannual, and seasonal cycles. Climatological standard deviations are linearly interpolated from monthly to daily means. For the NCEP–NCAR 40-yr reanalysis, the best analysis resolution is 2.5° × 2.5° globally. We have to interpolate the data to 1.0° × 1.0° to match the forecast resolution. The original resolution of the CFSRR is 1.0° × 1.0°.

The analysis climatology of precipitation was calculated based on climatology-calibrated precipitation analysis (CCPA; Hou et al. 2014) across the continental United States (CONUS). A gamma distribution was used to fit the precipitation distribution for each day of the year and each 1 × 1 grid point. The distribution parameters were determined via the L-moment method (Hosking 1990; Hosking and Wallis 1997). Details on the generation of climatology can be found online (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 the ensemble forecast $F_{en}(p)$ and the expected value of the climate distribution C:

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

In this work, we specifically calculate the ANFs for the ensemble mean and the 50th percentile for 2-m temperature and precipitation, respectively. For 2-m temperature, we calculate the value of ANF divided by one climatological standard deviation, the so-called standardized anomaly in Grumm (2001). For 24-h 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 to be a normal distribution $C = N(x, \mu, \sigma^2)$ and a 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 precipitation over North America. The x, μ, σ^2, k, and θ represent the location, mean, variance, shape factor, and scale parameter for the corresponding distributions, respectively.

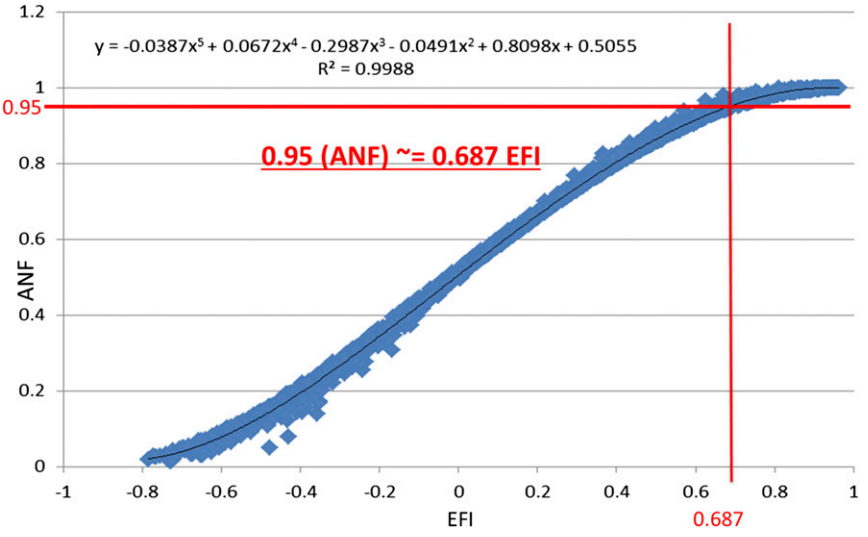

FIG. 2. Comparison of the 50th percentile ANF and EFI results for accumulated precipitation forecasts (72–96 h) over North America. The v11 raw forecast and model climatology are used in producing the ANF and EFI results. The solid line represents the best-fit curve. The forecasts are initiated at 0000 UTC 2 Jan 2014.

b. *EFI*

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

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

where p is the proportion of the 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 -1 and 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

a. *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 scatterplots of analysis anomalies and EFI, hit rate, false alarm rate, and relative operational characteristics (ROC) 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 the 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-yr CFSRR climatological data (Saha et al. 2010) and 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 the ensemble mean 96-h ANF and EFI for 2-m temperature on 1 March 2015 over North America. The ANF and EFI

TABLE 1. Contingency table used to evaluate forecasts of extreme events.

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

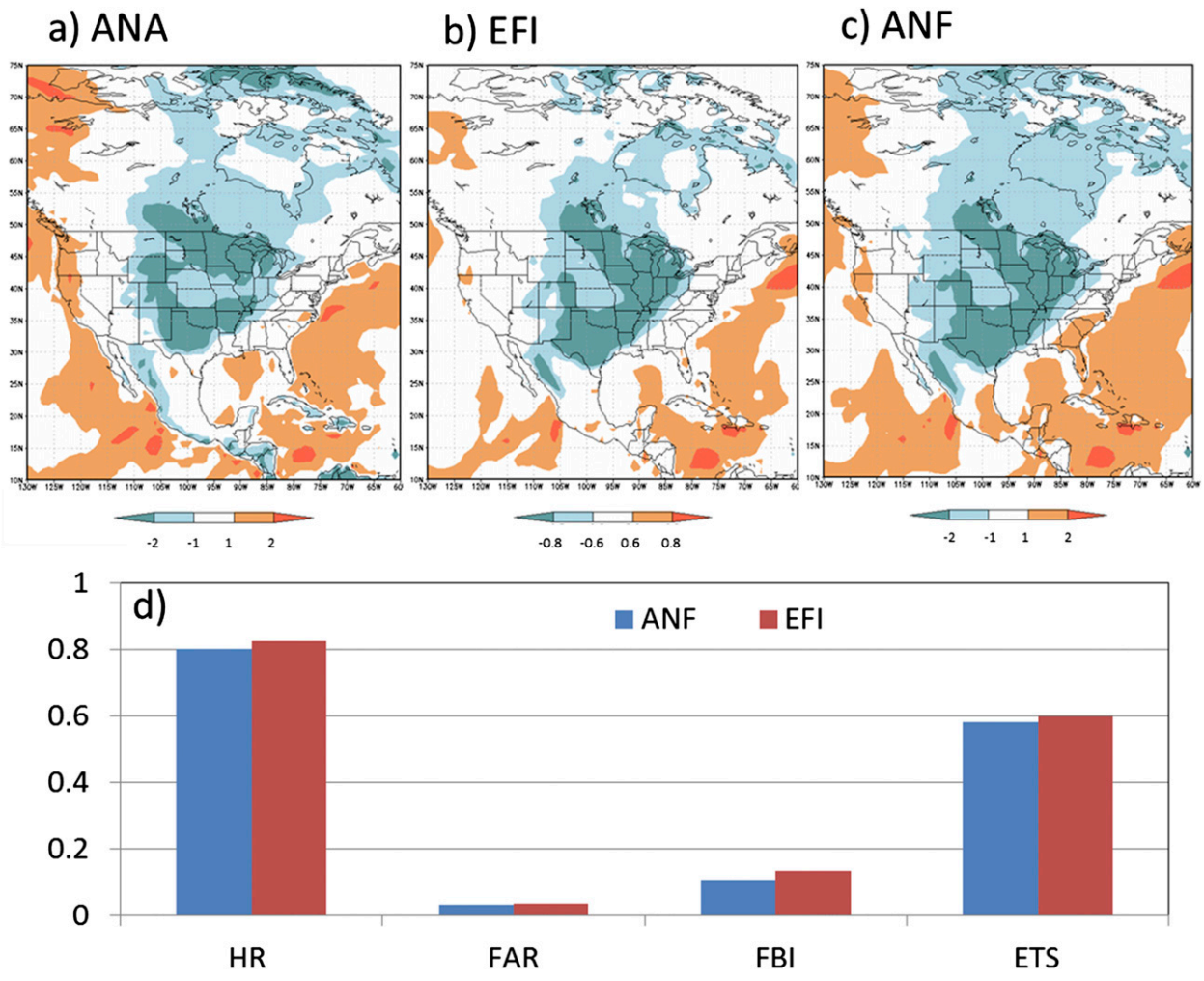

FIG. 3. (a) Extreme cold weather event observations or anomaly analysis (ANA), (b) 96-h EFI forecast, (c) 96-h ANF forecast, and (d) verification for both methods. The v11 raw forecast and v11 model climatology are used in producing the ANF and EFI results. The forecasts are initiated at 0000 UTC 1 Mar 2015.

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 fifth-order polynomial function through this sample dataset. 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 intercomparisons.

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 6 January 2014 over North America. Similar to the 2-m temperature, ANF and EFI are highly correlated and a fifth-order polynomial also best fits the dataset. However, instead of using σ as the ANF unit, here we use percentiles to express the precipitation ANF since a normal distribution cannot represent the asymmetric character of precipitation. The thresholds for ANF and EFI are taken as 0.95 and 0.687, 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 was assessed. Table 1 incorporates the model and the observation into a 2×2 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

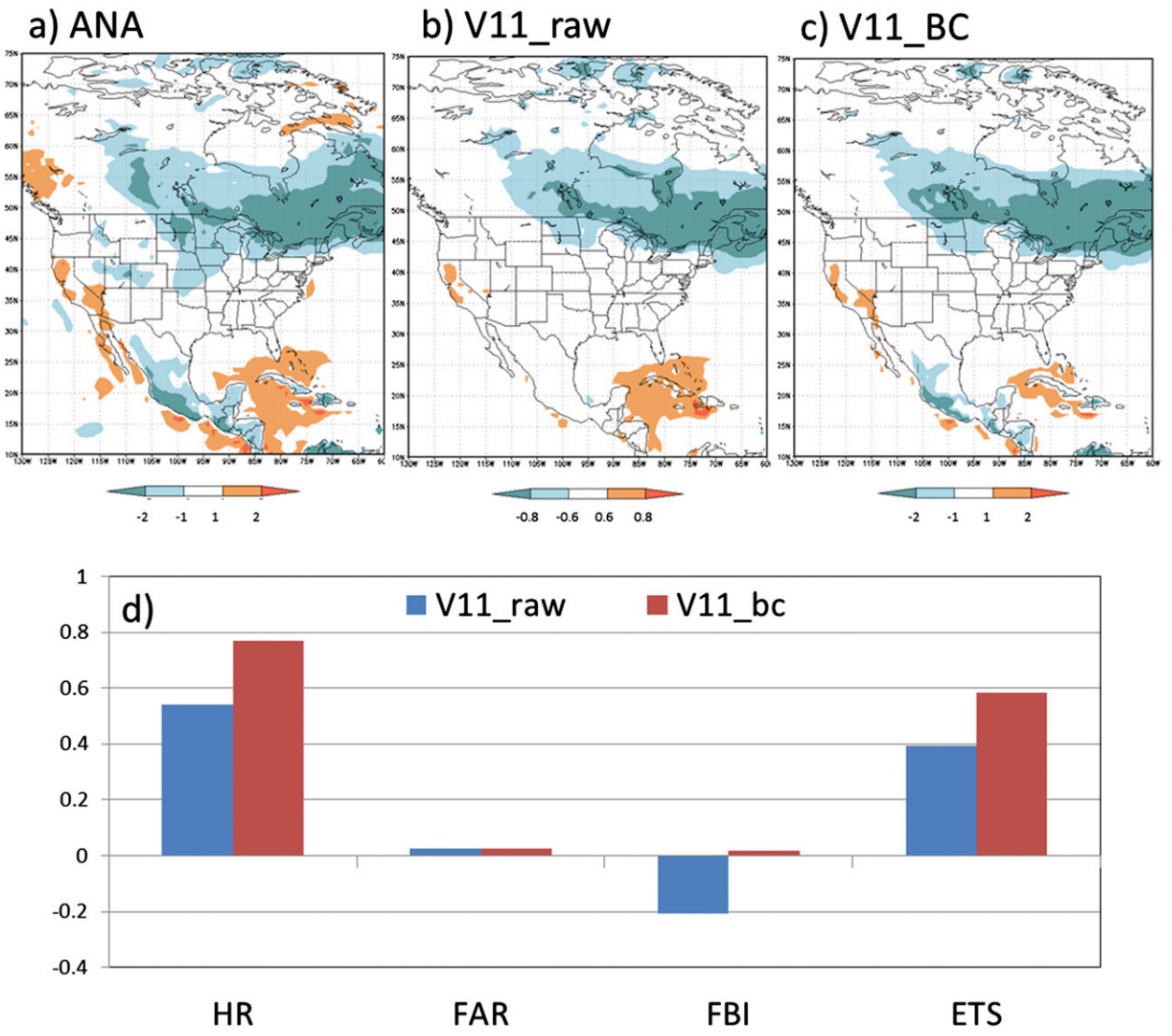

FIG. 4. (a) Extreme cold weather event observations or ANA, (b) 96-h raw EFI forecast, (c) 96-h bias-corrected EFI forecast, and (d) verification for the v11 RAW and v11 bias-corrected forecasts. The 18-yr control-only and CFSRR climatology are used in producing the raw and bias-corrected forecast products, respectively. The forecasts are initiated at 0000 UTC 29 Dec 2013.

bias (FBI), and equivalent threat scores (ETS) (Schaefer 1990) are defined as

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

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

$$\mathrm{FBI} = (A+C)/(A+B) - 1, \quad \text{and} \tag{5}$$

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

where

$$R(h) = [(A+C)(C+B)]/(A+B+C+D). \tag{7}$$

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 and Toth 2008).

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

$$\mathrm{CSI} = A/(A+B+C) \quad \text{and} \tag{8}$$

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

In section 4b, we also use a performance diagram to display the verification results for extreme cold events.

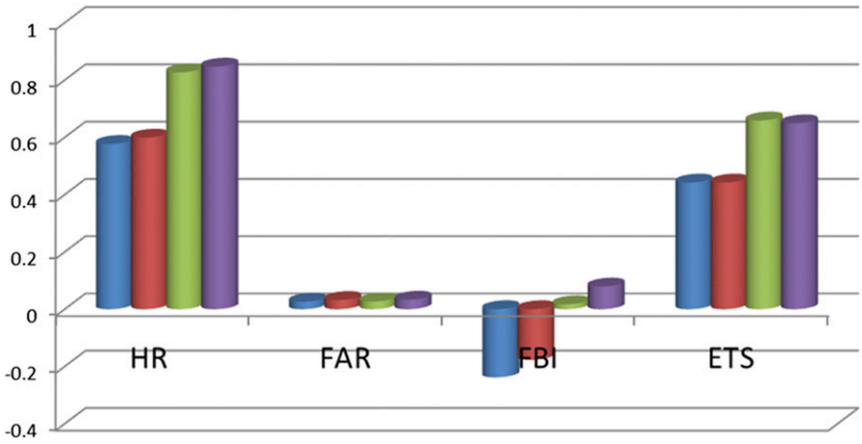

FIG. 5. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms (EFI and ANF) and forecasts (raw and bias corrected) across North America. Blue and red bars represent the v11 raw ANF and EFI results, respectively; green and purple bars are for the v11 bias-corrected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

b. Verification of extreme cold event forecasts

Using the verification methodology developed in section 4a, we compare the performance of the ANF and EFI products in forecasting extreme cold events for different model versions and forecasts. We also examine how using different analysis climatologies (NCEP–NCAR 40-yr reanalysis versus 30-yr 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 2013/14. This winter was considered to be colder and snowier than normal, as noted in Van Oldenborgh et al. (2015) and in 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 during the periods of 6–10 December 2013 and 29 December 2013–7 January 2014. Both cold waves caused extreme cold temperatures and broke daily precipitation and snowfall records across a considerable portion of North America.

1) VERIFICATION OF ANF AND EFI PRODUCTS

We show verifications of the EFI (Fig. 3b) and ANF (Fig. 3c) products against observations (Fig. 3a) across North America for the GEFS v11 raw forecasts for 0000 UTC 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; ETS, 0.58). The EFI predicts more extreme cold events than the ANF based on the FBI comparison. This may explain why EFI has slightly higher HR and ETS values. The FAR values (~0.03) are very similar 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

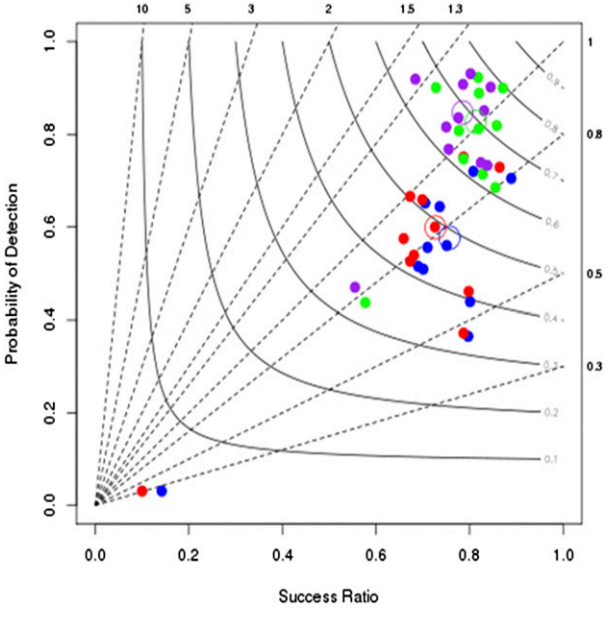

FIG. 6. Performance diagram summarizing the SR, POD, bias, and CSI results. Solid and dashed lines represent CSI and bias scores, respectively. Shown are 96-h forecasts of extreme cold weather for 11 individual days from the raw ANF (blue dots), raw EFI (red dots), bias-corrected ANF (green dots), and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

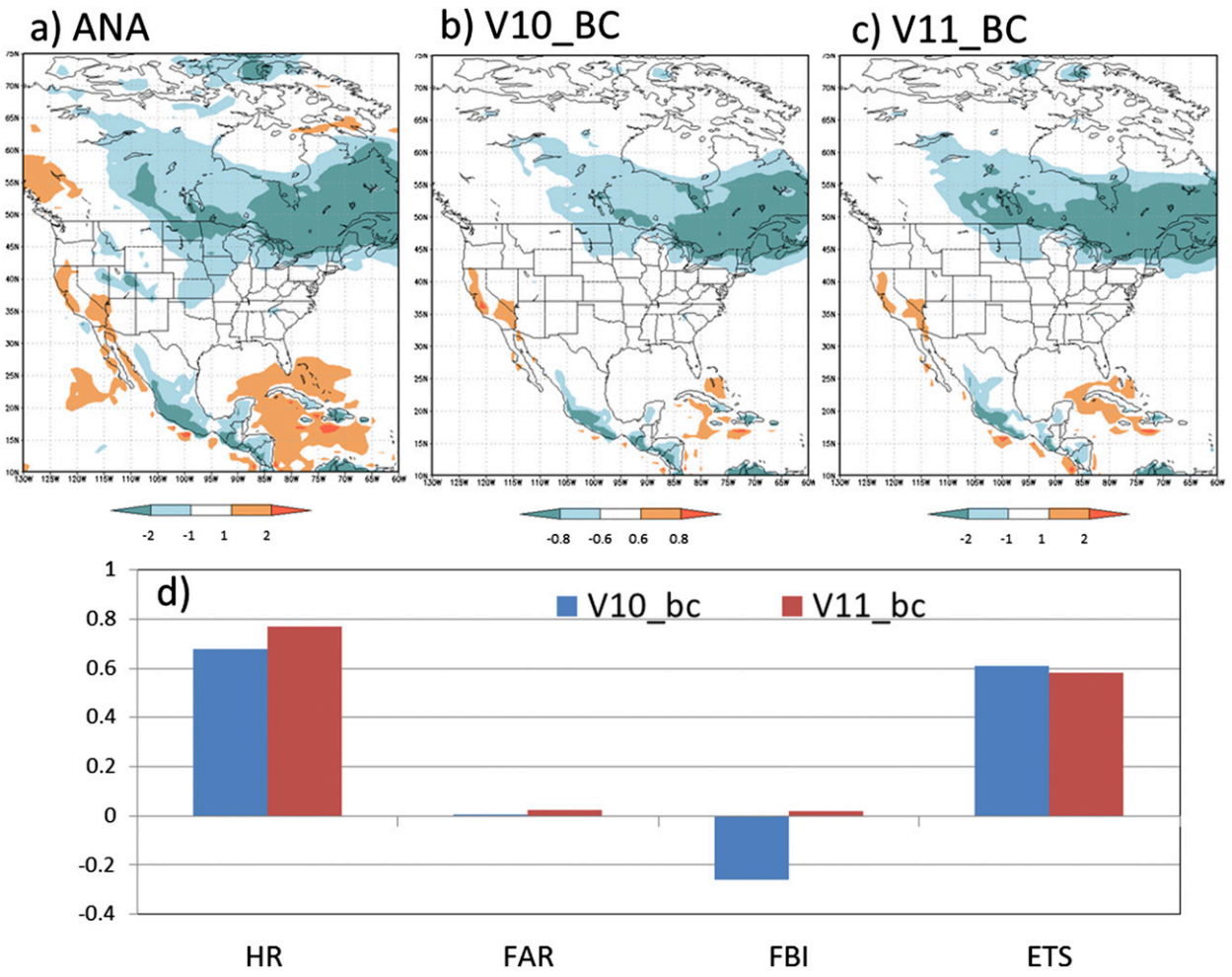

FIG. 7. (a) Extreme cold weather event observations, EFI product from (b) v10 and (c) v11 96-h bias-corrected forecasts, and (d) verification for both of the model versions. The forecasts are initiated at 0000 UTC 29 Dec 2013 and the reference climatology is CFSRR.

identifying the extreme cold area is another reason for the low FAR.

2) 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 29 December 2013. Both the raw and bias-corrected forecasts predict extreme cold weather across 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 the raw and bias-corrected forecasts in the ANF is also very similar to that of the EFI. Both methods demonstrate much better performance for the bias-corrected than the raw forecasts.

Figure 6 is the performance diagram for the above cases. A perfect forecast should have all four measures (HR, SR, bias, and CSI) equal to 1. In other words, a good forecast is closer to the top-right corner of the diagram. Obviously, the dots for the bias-corrected

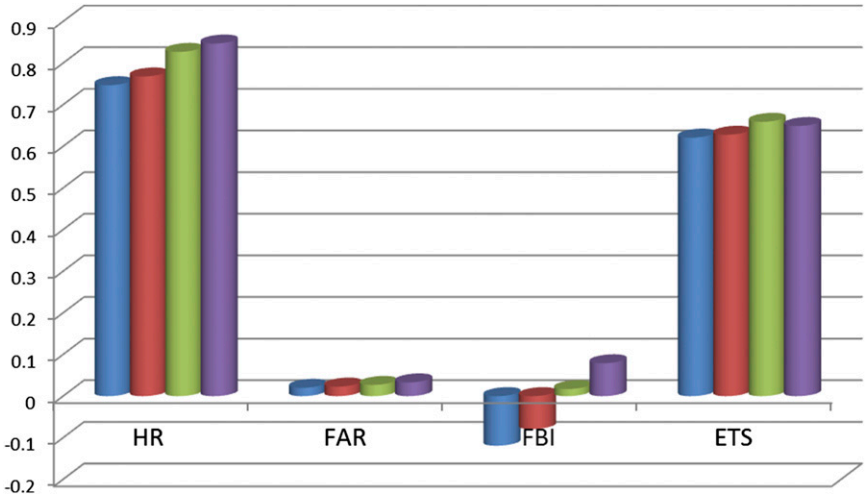

FIG. 8. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms (ANF and EFI) and model versions (v10 and v11) over North America. Blue and red bars represent the v10 bias-corrected ANF and EFI results, respectively; green and purple bars are for the v11 bias-corrected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

forecasts are more concentrated in the top right than for the raw forecasts. Overall, the bias-corrected ANF for the 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 the ensemble size on the probability forecast has been investigated by Buizza and Palmer (1998) and Ma et al. (2012). An increase in ensemble size is strongly beneficial to the forecast when the ensemble has fewer than 40 members. An effort is being made to create a model climatology using multimember reforecast runs. This would provide more robust model climatology and improve extreme weather forecasts.

3) VERIFICATION OF V10 AND V11 FORECAST PRODUCTS

Figure 7 shows the verification for the GEFS v10 and v11 bias-corrected forecasts for 0000 UTC 2 January 2014 using the EFI method. In general, both of the 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 the 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 has a higher HR, but the ETS is slightly smaller when compared with the v10.

The 11-day statistics are shown in Fig. 8. Overall, the v11 performs better than the v10 version with higher HR and ETS values. The v11 predicts more extreme events than are observed, while the v10 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 also be demonstrated in the performance diagram (Fig. 9). Overall, v11 is closer to the top-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.

4) VERIFICATION OF FORECAST PRODUCTS FOR A DIFFERENT REFERENCE CLIMATOLOGY

The current NCEP GEFS ANF product uses the 40-yr reanalysis as its reference climatology. To test the sensitivity of the ANF and EFI skill to their reference, we make verification comparisons with two different references (30-yr CFSRR and 40-yr reanalysis) in Figs. 10 and 11. The ANF and EFI calculated relative to the CFSRR climatology have slightly better HR, FBI, and ETS than those of 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

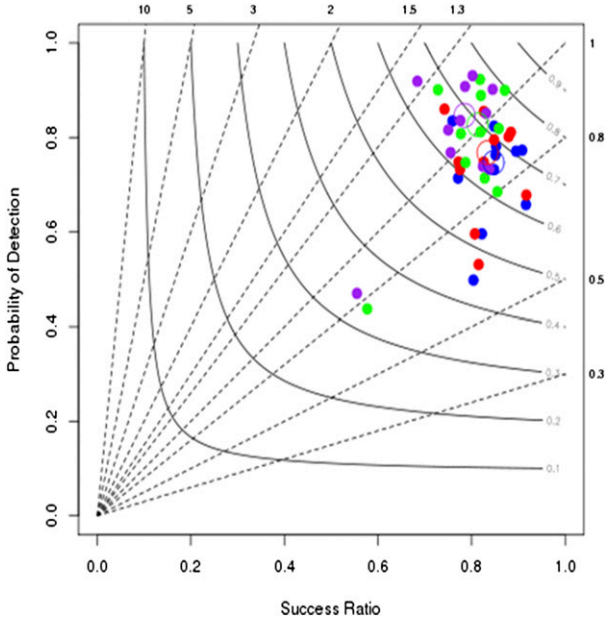

FIG. 9. Performance diagram as in Fig. 6, but for the comparisons of the two model versions. Blue and red dots represent the v10 bias-corrected ANF and EFI results, respectively; green and purple dots are for the v11 bias-corrected ANF and EFI results, respectively. All forecasts are 96-h forecasts from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

are closer to the top-right corner than for the reanalysis reference, indicating a slightly higher accuracy when a more sophisticated analysis is used. The sensitivities of the verification scores to the references for the ANF and EFI are 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.

c. *Verification of heavy precipitation forecasts*

The 96-h forecasts of extreme precipitation regions from the ANF (Fig. 12a) and EFI (Fig. 12b) products are shown, initiated at 0000 UTC 6 January 2014. The shaded areas are the corresponding 72–96-h 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. Figure 12 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 compares 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 December 2013. Again, the forecasts from the two products are very similar and capture the major extreme precipitation region over the United States. The verification scores demonstrate that the EFI predicts more extreme events

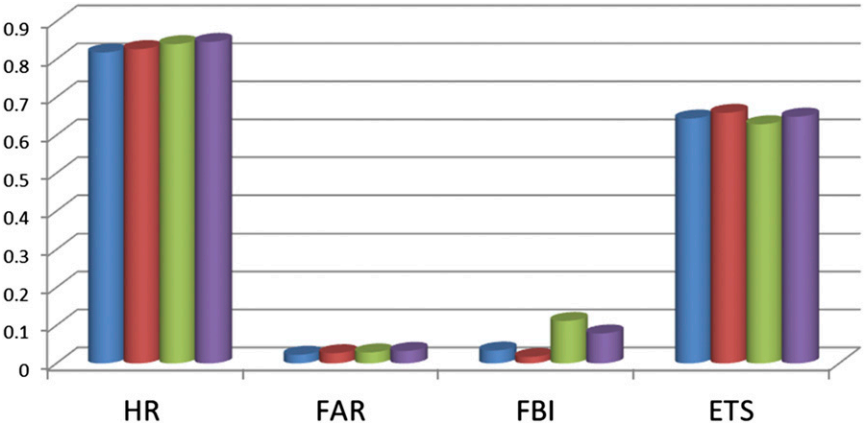

FIG. 10. The 2-m temperature histograms of HR, FAR, FBI, and ETS for 11 days with different algorithms and reference climatologies over North America. Blue and red bars represent the v11 bias-corrected ANF results with the NCEP–NCAR reanalysis and CFSRR as a reference, respectively; green and purple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

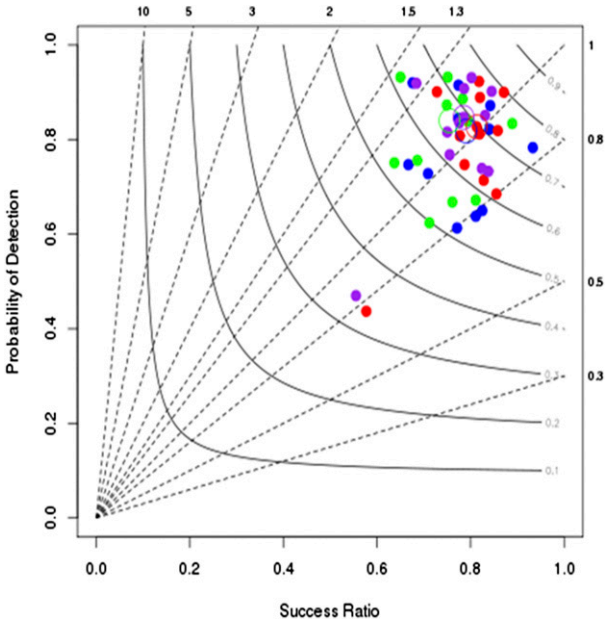

FIG. 11. Performance diagram as in Fig. 6, but for comparisons of the two reference climatologies (30-yr CFSRR and 40-yr reanalysis). Blues and green dots represent ANF and EFI results using the 40-yr reanalysis as the reference; red and purple dots are for the ANF and EFI results using the 30-yr CFSRR as the reference.

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 2013/14. 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 climatologies 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 for evaluating and comparing 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 during the winter of 2013/14 indicate the EFI forecasts more extreme cold events than the ANF. The ANF produces a higher ETS value. The bias-corrected forecast shows much better performance than the raw forecast when an 18-yr 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 toward finding the optimized configuration of real-time GEFS reforecast runs is 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

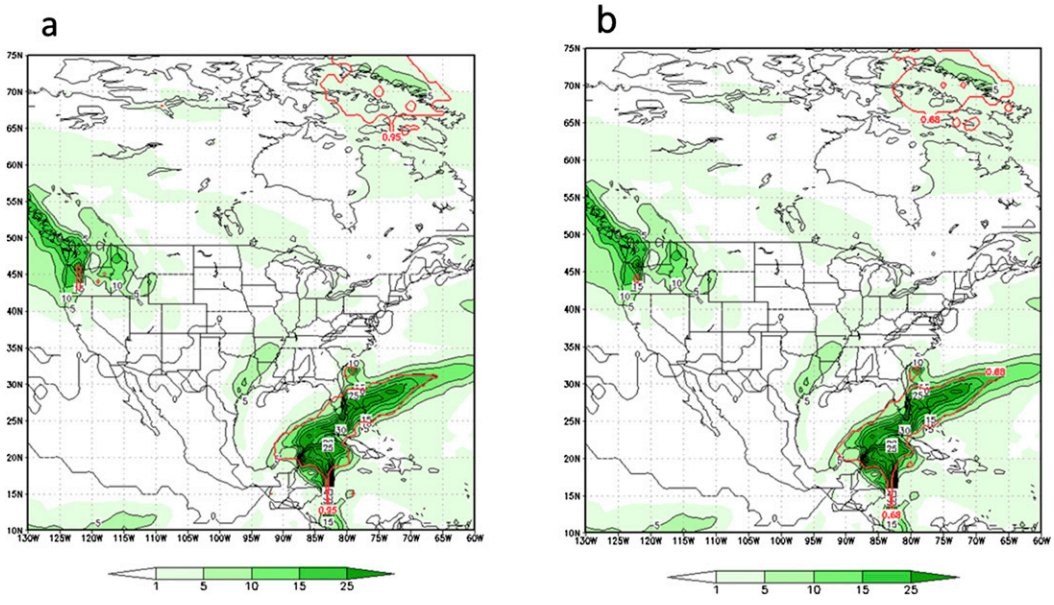

FIG. 12. The 96-h forecasts of extreme precipitation regions (red contours) from the (a) ANF and (b) EFI products. The shaded areas are corresponding 72–96-h accumulated precipitation forecasts (mm). The contours in (a) and (b) represent ANF = 0.95 and EFI = 0.687, respectively. The forecasts are initiated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

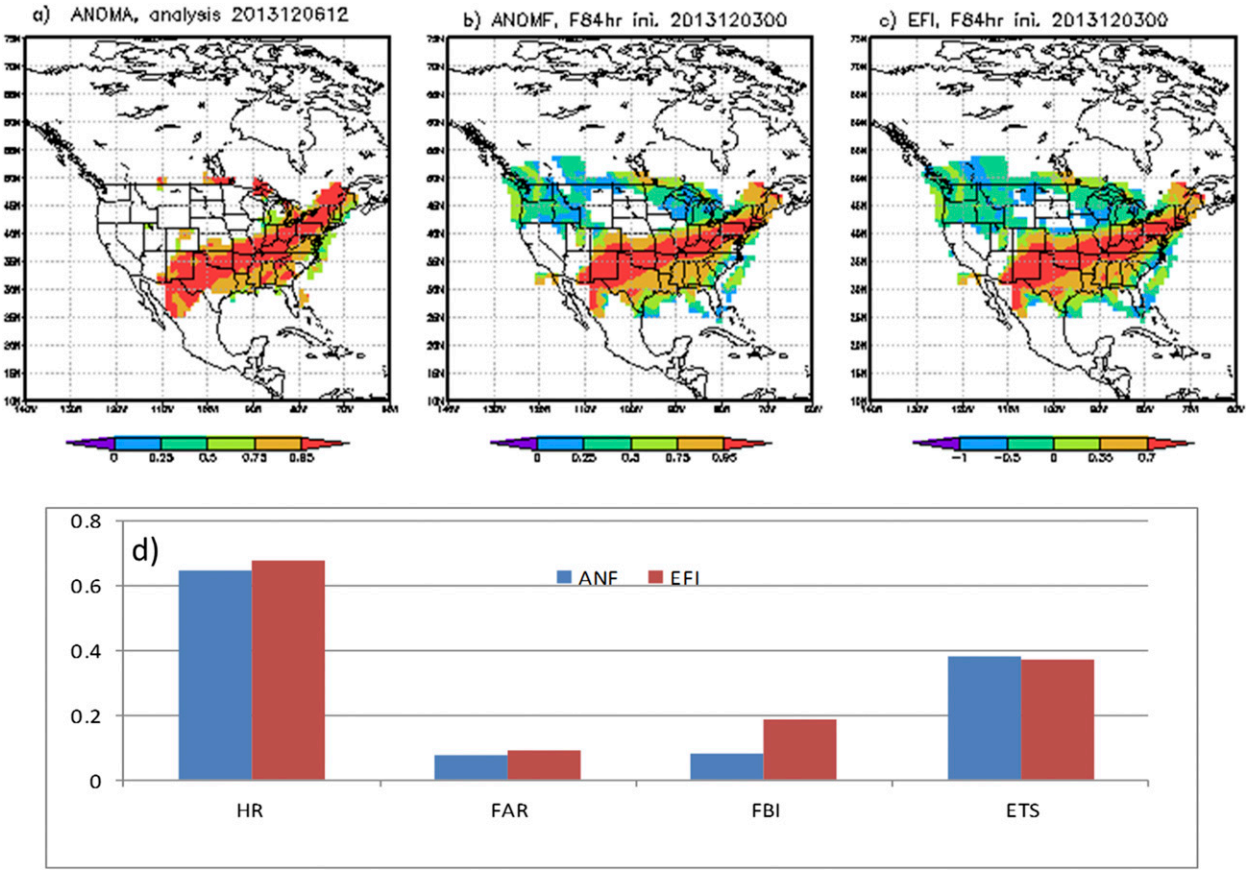

FIG. 13. The daily extreme precipitation distribution (60–84 h) for (a) ANA and the (b) ANF and (c) EFI forecasts, and (d) verification for both methods. The v11 forecasts are initiated at 0000 UTC 3 Dec 2013.

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-yr 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 during the winter of 2013/14. 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 the ANF–EFI relationship on forecast lead time will also be our focus.

Acknowledgments. The authors thank the members of the ensemble and postprocessing 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 our understanding of the CCPA data, and Bo Yang, who provided the CFSRR data. The authors would also like to acknowledge the helpful advice and discussion of Drs. Bo Cui, Malaquias Pena, and Corey Guastini. Dr. Tara Jensen of NCAR is also thanked for providing the R program used to generate performance diagrams. This work was completed as a part of regular work duties at NOAA/NWS/NCEP/EMC.

REFERENCES

Buizza, R., and T. N. Palmer, 1998: Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.*, **126**, 2503–2518, doi:10.1175/1520-0493(1998)126<2503:IOESOE>2.0.CO;2.

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., III, R. Davies-Jones, and D. L. Keller, 1990: On summary measures of skill in rare event forecasting based on contingency tables. *Wea. Forecasting*, **5**, 576–585, doi:10.1175/1520-0434(1990)005<0576:OSMOSI>2.0.CO;2.

Dutra, E., M. Diamantakis, I. Tsonevsky, E. Zsoter, F. Wetterhall, T. Stockdale, D. Richardson, and F. Pappenberger, 2013: The

extreme forecast index at the seasonal scale. *Atmos. Sci. Lett.*, **14**, 256–262, doi:10.1002/asl2.448.

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, doi:10.1175/2009WAF2222273.1.

Grumm, R. H., 2001: Standardized anomalies applied to significant cold season weather events: Preliminary findings. *Wea. Forecasting*, **16**, 736–754, doi:10.1175/1520-0434(2001)016<0736:SAATSC>2.0.CO;2.

Guan, H., B. Cui, and Y. Zhu, 2015: Improvement of statistical postprocessing using GEFS reforecast information. *Wea. Forecasting*, **30**, 841–854, doi: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 dataset. *Bull. Amer. Meteor. Soc.*, **94**, 1553–1565, doi:10.1175/BAMS-D-12-00014.1.

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

Hosking, J. R. M., 1990: L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *J. Roy. Stat. Soc.*, **52B**, 105–124.

——, and J. R. Wallis, 1997: *Regional Frequency Analysis: An Approach Based on L-Moments*. Cambridge University Press, 244 pp.

Hou, D., Z. Toth, Y. Zhu, and W. Yang, 2008: Evaluation of the impact of the stochastic perturbation schemes on global ensemble forecast. *19th Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., 1.1. [Available online at https://ams.confex.com/ams/88Annual/techprogram/paper_134165.htm.]

——, 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:10.1175/JHM-D-11-0140.1.

IPCC, 2011: *Special Report on Renewable Energy Sources and Climate Change Mitigation*. Cambridge University Press, 1075 pp.

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

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: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, doi:10.1256/qj.02.152.

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**, 782–794, doi:10.1007/s00376-012-1249-y.

Mason, I. B., 1982: A model for the assessment of weather forecasts. *Aust. Meteor. 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. *Research Activities in Atmospheric and Oceanic Modelling*, A. Zadra, Ed., WCRP Rep. 10/2013. [Available online at https://www.wcrp-climate.org/WGNE/BlueBook/2013/individual-articles/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*, No. 132, ECMWF, Reading, United Kingdom, 13–19. [Available online at http://www.ecmwf.int/en/elibrary/14590-newsletter-no132-summer-2012.]

Roebber, P. J., 2009: Visualizing multiple measures of forecast quality. *Wea. Forecasting*, **24**, 601–608, doi:10.1175/2008WAF2222159.1.

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

Sela, J., 2010: The derivation of the sigma pressure hybrid coordinate Semi-Lagrangian model equations for the GFS. NCEP Office Note 462, 31 pp. [Available online at 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, doi:10.1175/1520-0434(1990)005<0570:TCSIAA>2.0.CO;2.

Swets, J. A., 1988: Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293, doi:10.1126/science.3287615.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., John Wiley and Sons, 137–163.

Van Oldenborgh, G. J., R. Haarsma, H. De Vries, and M. 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:10.1175/BAMS-D-14-00036.1.

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 transform (ET) technique in the NCEP global operational forecast system. *Tellus*, **60A**, 62–79, doi:10.1111/j.1600-0870.2007.00273.x.

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, doi:10.1175/MWR-D-11-00276.1.

——, ——, X. Wei, Y. Song, and Z. Toth, 2008: Ensemble data assimilation with the NCEP Global Forecast System. *Mon. Wea. Rev.*, **136**, 463–482, doi:10.1175/2007MWR2018.1.

Zhu, Y., 2015: GEFS upgrade (V11). NWS/Environmental Modeling Center. [Available online at http://www.emc.ncep.noaa.gov/gmb/yzhu/imp/i201412/GEFS_sci_briefing.pdf.]

——, and B. Cui, 2007: NAEFS mean, spread and probability forecasts. NOAA/NCEP Rep., 4 pp. http://www.emc.ncep.noaa.gov/gmb/yzhu/imp/i200711/3-Mean_spread.pdf.]

——, and Z. Toth, 2008: Ensemble based probabilistic forecast verification. *19th Conf. on Probability and Statistics in the Atmospheric Sciences*, New Orleans, LA, Amer. Meteor. Soc., 2.2. [Available online at https://ams.confex.com/ams/pdfpapers/131645.pdf.]

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

——, D. Hou, M. Wei, R. Wobus, J. Ma, B. Cui, and S. Moorthi, 2012: GEFS upgrade—AOP plan—Major implementation. NWS/Environmental Modeling Center. [Available online at http://www.emc.ncep.noaa.gov/gmb/yzhu/html/imp/201109_imp.html.]

Zsótér, E., 2006: Recent developments in extreme weather forecasting. *ECMWF Newsletter*, No. 107, ECMWF, Reading, United Kingdom, 8–17. [Available online at http://www.ecmwf.int/sites/default/files/elibrary/2006/14618-newsletter-no107-spring-2006.pdf.]