Extreme Precipitation in Models: An Evaluation

GREGORY R. HERMAN AND RUSS S. SCHUMACHER

Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado

(Manuscript received 12 May 2016, in final form 12 July 2016)

ABSTRACT

A continental United States (CONUS)-wide framework for analyzing quantitative precipitation forecasts (QPFs) from NWP models from the perspective of precipitation return period (RP) exceedances is introduced using threshold estimates derived from a combination of NOAA Atlas 14 and older sources. Forecasts between 2009 and 2015 from several different NWP models of varying configurations and spatial resolutions are analyzed to assess bias characteristics and forecast skill for predicting RP exceedances. Specifically, NOAA's Global Ensemble Forecast System Reforecast (GEFS/R) and the National Severe Storms Laboratory WRF (NSSL-WRF) model are evaluated for 24-h precipitation accumulations. The climatology of extreme precipitation events for 6-h accumulations is also explored in three convection-allowing models: 1) NSSL-WRF, 2) the North American Mesoscale 4-km nest (NAM-NEST), and 3) the experimental High Resolution Rapid Refresh (HRRR). The GEFS/R and NSSL-WRF are both found to exhibit similar 24-h accumulation RP exceedance climatologies over the U.S. West Coast to those found in observations and are found to be approximately equally skillful at predicting these exceedance events in this region. In contrast, over the eastern two-thirds of the CONUS, GEFS/R struggles to predict the predominantly convectively driven extreme QPFs, predicting far fewer events than are observed and exhibiting inferior forecast skill to the NSSL-WRF. The NSSL-WRF and HRRR are found to produce 6-h extreme precipitation climatologies that are approximately in accord with those found in the observations, while NAM-NEST produces many more RP exceedances than are observed across all of the CONUS.

1. Introduction

Heavy precipitation and associated flooding and flash flooding have an enormous impact on many different facets of society. As a weather hazard, with 81 average annual deaths, floods are responsible for more fatalities in the United States over the last 30 years than any other single severe weather hazard, including tornadoes, hurricanes, lightning, and other windstorms (NWS 2002). Floods also heavily damage and destroy buildings, roads, crops, and other property; in 2014, flash floods were responsible for more economic damage than any other weather hazard, with nearly $2.5 billion in reported flash flood damages occurring that year (NWS 2015). Though some damages from extreme rainfall and flooding are inevitable, appropriate preparedness can greatly alleviate damages and almost completely eliminate flood fatalities (NWS 2002). As such, both accurate forecasts and proper understanding of the physical mechanisms underlying extreme precipitation and flood processes are of immense value to society.

A critical first step in addressing this societal challenge requires assessing our current capability for forecasting these high-impact flooding and flash flooding events. Flash flood forecasting is an exceptionally difficult forecast problem (e.g., Doswell et al. 1996). The hydrologic response to rainfall is known to be highly sensitive to the precise locations and precipitation amounts; antecedent conditions, including soil moisture and other surface properties; and even precipitation timing, duration, and distribution (e.g., Hapuarachchi et al. 2011). Additionally, extreme precipitation, particularly in the warm season, is widely considered to be one of the most poorly forecast variables in numerical weather prediction (NWP; e.g., Fritsch and Carbone 2004; Sukovich et al. 2014). Given the great sensitivity and immense challenge in forecasting the underlying forcing for hydrologic models, it is frequent practice in operational settings to attempt to estimate the impacts of heavy precipitation directly

Corresponding author address: Gregory R. Herman, Dept. of Atmospheric Science, Colorado State University, Fort Collins, CO 80523.
E-mail: gherman@atmos.colostate.edu

DOI: 10.1175/WAF-D-16-0093.1

from quantitative precipitation forecasts (QPFs) from NWP models (e.g., Smith et al. 2005). Flash flood potential based on QPFs can be assessed in several different ways. One approach attempts to assess the local climatology and antecedent conditions to produce flash flood guidance (FFG) thresholds; exceedance of these thresholds at their respective points is believed to produce flash flooding (e.g., Schmidt et al. 2007). This approach, performed at River Forecast Centers (RFCs), has the advantage of accounting for preceding rainfall and anomalous soil moisture, but suffers from ambiguity and subjectivity; for example, large discontinuities in threshold estimates across RFC boundaries are not uncommon (e.g., Ortega et al. 2009). A much simpler approach simply considers whether some fixed, large precipitation threshold will be exceeded over a specified accumulation interval (AI), such as 50 mm in 3 h. This is simple to apply and evaluate, and traditionally QPFs are verified in this manner (e.g., Brooks and Stensrud 2000; Hamill and Whitaker 2006; Rossa et al. 2008; Marsh et al. 2012; Baxter et al. 2014; Novak et al. 2014; Sukovich et al. 2014; Scheuerer and Hamill 2015). However, this approach is deeply flawed from the consideration of extreme precipitation impacts. For example, 50 mm in a 3-h period may be fairly routine in parts of the southeastern United States and result in minimal impacts there, while the same rainfall amount may result in damaging flooding over parts of the Intermountain West, as in the case of the flash flood in southern Utah during September 2015 that killed 13 (NCEI 2015). Relatedly, using this framework for forecast verification over the hydrometeorologically diverse regions of the continental United States (CONUS) may be misleading; verification results may be biased toward the model's performance in the climatologically wetter regions as a result of the varying climatology. This verification issue, raised by Hamill and Juras (2006) and others, can only be fully resolved if the verifying event is equally likely to occur at any location. A robust framework for statistical analysis and verification can be found through the use of precipitation return periods (RPs) and associated return period thresholds (RPTs; e.g., Stevenson and Schumacher 2014; Herman 2016). RPT exceedances are also found to have a much better correspondence with precipitation impacts and flash flood potential than the use of any fixed threshold across the meteorologically and hydrologically diverse regions of the CONUS (e.g., Reed et al. 2007). For these reasons, this study will develop and utilize an RP-based verification framework for examining model performance in forecasting extreme precipitation.

To improve forecasting of extreme precipitation and flash flooding, both at present and in future development, it is necessary to understand how contemporary modeling systems perform in different extreme precipitation scenarios in order to best correct for model biases and ascertain which forecast information should be given the most credence (e.g., Fritsch and Carbone 2004). This study will attempt to build on the state of knowledge in this area by using a variety of methods to both qualitatively and quantitatively assess the ability of operational and experimental/research NWP models—of varying resolutions and degrees of sophistication—to forecast locally extreme rainfall events. These events will be characterized in this study by an RP *N* and AI *T*: an *N*-year *T*-hour event. Here, *T* is a descriptor of the type of event, namely the span of time for precipitation to accumulate, while *N* characterizes the rarity of the event. The datasets used for this research and the methodology employed to assess model performance in forecasting locally extreme precipitation will be described in section 2. In section 3, the return period threshold estimates to be used for this analysis are presented, and a brief assessment of the climatology of extreme precipitation over the CONUS as discerned by RPT exceedances is made. Model analysis and verification is then presented, first within a broad, qualitative context in order to approximate model performance for several different numerical models. Later, the presented results will attempt to quantify model bias characteristics and model skill for a subset of models. Findings will be summarized and important conclusions highlighted in section 4.

2. Data and methods

a. Forming return period threshold grids

Analysis of model performance in the return period/recurrence interval framework first requires establishing the actual numerical thresholds—RPTs—corresponding to the RPs of interest for all locations of interest. This paper seeks to assess model performance in all regions of the CONUS; thus, nationwide threshold grids are required. RPs of 1, 2, 5, 10, 25, 50, and 100 yr are evaluated for the verification performed herein. Despite the challenges and large degrees of uncertainty associated with characterizing events occurring with expected occurrences of once per hundreds or even thousands of years from only decades of observational data, because of the immense importance of local RP estimates for hydrology and other applications, substantial work has been conducted using a long record of observations—primarily gauge data—to estimate RPTs for various AIs (e.g., Tye and Cooley 2015). Over the past several years, the National Oceanic and Atmospheric Administration (NOAA), and specifically the Hydrometeorological Design Studies Center (HDSC), has made a major effort to generate comprehensive and up-to-date RPT estimates via the Atlas 14 project. The project seeks to develop updated CONUS-wide RPT estimates for RPs from 1 to 1000 and AIs from minutes to months. This product is an update of work done by Hershfield published in Technical Paper 40 (TP-40; Hershfield 1961), which spanned much of the United States east of the continental divide, and NOAA's Atlas 2, released in 1973 for the western states (Miller et al. 1973). Both TP-40 and Atlas 2 fit Gumbel distributions (Gumbel 1960)—a right-skewed distribution and special case of the generalized extreme value (GEV) distribution—to station gauge data for 6- and 24-h accumulation intervals to derive 2–100-yr RPT estimates. Topographically aware formulas were then derived and applied to extend those estimates to all points (Hershfield 1961; Miller et al. 1973). In addition to having several decades of new data with increased station density to improve the precipitation frequency estimates, Atlas 14 uses more sophisticated methods for deriving estimates than its predecessors. A suite of different distributions were fit to the precipitation data, using the method of L moments for parameter estimation; goodness-of-fit tests such as the Kolmogorov–Smirnov approach were conducted and used to assess the optimal choice of distribution. To date, all Atlas 14 updates have selected the GEV distribution as the distribution that most often had an acceptable fit to the observational data, and have chosen to apply it uniformly so as to avoid large spatial discontinuities. A more sophisticated regionalization technique was employed to use data from multiple nearby stations to inform a point rainfall frequency estimate. RPs also extended from 1 up to 1000 yr, and estimates are available for AIs ranging from minutes to months (Bonnin et al. 2004, 2006; Perica et al. 2011, 2013). At the time this study was conducted, Atlas 14 had updated previous RPT estimates for the majority of the CONUS; however, the northwest (Washington, Oregon, Idaho, Montana, and Wyoming), northeast (New York, Vermont, New Hampshire, Maine, Massachusetts, Connecticut, and Rhode Island), and Texas had not yet received published updated estimates from Atlas 14.[1] Atlas 2 estimates were used for the northwestern states, while TP-40 estimates were used in Texas and the aforementioned northeastern states. Each of these estimates have considerable uncertainty, especially at higher RPs; however, the quantitative uncertainty associated with the RPT estimates is beyond the scope of this study and will not be considered in determining RP exceedance events.

CONUS-wide RPT estimates for the 6- and 24-h AIs are then used as the basis for determining the climatology of locally extreme precipitation events as discerned from National Centers for Environmental Prediction (NCEP) stage IV precipitation analysis, a high-resolution (~4.75 km grid spacing), multisensor product using both rain gauge observations and radar-derived rainfall estimates to generate CONUS-wide analyses for 1-, 6- and 24-h AIs (Lin and Mitchell 2005). This product was selected because of its relatively long, consistent analysis record in conjunction with its rather high resolution across the CONUS, which is essential for acquiring acceptable estimates of highly localized extreme precipitation amounts. However, despite quality control (QC) by NWS RFCs to assure stray radar artifacts and other spurious anomalies do not appear in the final product, numerous deficiencies were encountered requiring further QC attention (e.g., Hitchens et al. 2013; Stevenson and Schumacher 2014; Nelson et al. 2016); additional QC procedures are described in more detail in the appendix. The product also exhibits other limitations that cannot be readily remedied via QC procedures. In areas of complex terrain, radar beam blockage may yield regions with no or poor radar-derived precipitation estimates (e.g., Zhang et al. 2011); unless sufficient rain gauges are located in these areas to compensate the lacking radar data, some areas within regions of complex terrain may have poor rainfall estimates, and, correspondingly, extreme precipitation occurrences in these areas may not be recorded in the stage IV product. Additionally, beam refraction from low-level surface inversions can degrade the accuracy of rain-rate estimates wherever and whenever they are present. These signals can be difficult or impossible to identify and correct during quality control.

b. *Model analysis and verification*

Further, raw QPFs of several diverse NWP models were used both to assess individual model characteristics and biases in the forecasting of extreme precipitation and also to quantify model skill. Depending on model data availability, either the 9 June 2009–30 August 2014 period or the shorter 12 August 2014–11 August 2015 period was used for evaluation and

[1] The northeastern states did receive updated Atlas 14 estimates in October 2015.

TABLE 1. Details of the NWP models used in this research. Information about horizontal and vertical resolution, and the various parameterizations used in each model, is included.

Model	Horizontal grid spacing	No. of vertical levels	Initial conditions/ boundary conditions	Cumulus	Planetary boundary layer	Microphysics	Land surface	Shortwave radiation	Longwave radiation
GEFS/R	T254	42	GFS	Simplied Arakawa–Schubert version 2 (SAS2)	Pan–Mahrt (PM)	Zhao–Carr (ZC)	Noah	RRTM	RRTM
NAM-NEST	4 km	60	NAM	—	MYJ	Ferrier	Noah	Lacis–Hansen (LH)	LH
NSSL-WRF	4 km	35	NAM	—	MYJ	WSM6	Noah	Dudhia	RRTM
HRRR	3 km	50	RAP/HRRR	—	Mellor–Yamada–Nakanishi–Niino (MYNN)	Thompson	RUC–Smirnova	Goddard	RRTM

comparison. The Global Ensemble Forecast System Reforecast, version 2 (GEFS/R), and National Severe Storms Laboratory Weather Research and Forecasting (NSSL-WRF) model were evaluated for the 24-h AI over the longer 2009–14 period for 12–36-h forecasts off of each model's 0000 UTC initialization. The GEFS/R dataset contains reforecasts of the February 2012 version of the GEFS from December 1984 to the present. Though only the control member is used here, it is an 11-member ensemble run at a coarse, cumulus parameterization-requiring T254L42 resolution (~40-km-equivalent horizontal grid spacing at 40° latitude), rerun once daily for the 0000 UTC forecast cycle. The NSSL-WRF, by contrast, is a limited-area model with a domain spanning the CONUS; it is a gridpoint model with 4-km horizontal grid spacing and no cumulus parameterization. The GEFS/R, by design, has no model changes of any kind during this analysis period except for changes to the data assimilation input in association with operational upgrades made in May 2012 (Hamill et al. 2013); the NSSL-WRF has only one change of note: an update of the WRF version from 3.1.1 to 3.4.1 in April 2013. The long record of unchanged model data allows for a robust diagnosis of model performance characteristics. A broader comparison of several convection allowing models (CAMs) was conducted for 6-h AIs over the shorter 2014–15 period comparing the NSSL-WRF, the North American Mesoscale 4-km nest (NAM-NEST), and the experimental version of the High Resolution Rapid Refresh (HRRR) run by the Earth Systems Research Laboratory (ESRL). The beginning of this period coincides with a major update to the NAM-NEST, which significantly altered the model's QPF bias characteristics (NWS 2014). The HRRR did experience model changes during this analysis period. The Rapid Refresh, which provides the HRRR initial conditions, was updated on 1 January 2015. In association with the annual spring update, on 10 April 2015, the HRRR WRF version was changed from 3.5.1 to 3.6.1, the data assimilation system was upgraded to a hybrid three-dimensional variational (3DVAR)–ensemble Kalman filter (EnKF) implementation, and several changes were made to the model physics; most notably, the Thompson microphysics parameterization was upgraded to an aerosol-aware formulation, with an added coupled parameterization to represent small-scale boundary layer clouds (Alexander et al. 2015; Benjamin et al. 2016). The developers implemented these changes to alleviate warm season warm and dry biases over the Great Plains and also to alleviate observed biases in QPFs. More details about each of these models, including a full list of parameterizations, appear in Table 1. Prior to analysis, all model data were regridded onto the stage IV Hydrologic Rainfall Analysis Projection (HRAP) grid using a first-order conservative scheme, preserving the total amount of precipitation on each grid.

Over the extended period analysis, quantitative assessments of model skill were also conducted by means of an aggregated fractions skill score (FSS; Roberts and Lean 2008). For an evaluation radius r, over K evaluation points and D evaluation times, the FSS is given by

$$\mathrm{FSS} = 1.0 - \frac{\sum_{d=1}^{D}\left[\sum_{j=1}^{K}(O_{jd} - M_{jd})^2\right]}{\sum_{d=1}^{D}\left[\sum_{j=1}^{K}(O_{jd}^2 + M_{jd}^2)\right]},$$

where

$$O_{jd} = \frac{1}{(2r+1)^2}\sum_{y=\mathrm{lat}_j - r}^{\mathrm{lat}_j + r}\ \sum_{x=\mathrm{lon}_j - r}^{\mathrm{lon}_j + r} E_{yxd} \quad \text{and}$$

$$M_{jd} = \frac{1}{(2r+1)^2}\sum_{y=\mathrm{lat}_j - r}^{\mathrm{lat}_j + r}\ \sum_{x=\mathrm{lon}_j - r}^{\mathrm{lon}_j + r} Z_{yxd}$$

with

$$E_{yxd} = \begin{cases} 1 & P_{yxd} \geq \theta_{yx} \\ 0 & P_{yxd} < \theta_{yx} \end{cases} \quad \text{and} \quad Z_{yxd} = \begin{cases} 1 & Q_{yxd} \geq \theta_{yx} \\ 0 & Q_{yxd} < \theta_{yx} \end{cases} .$$

Here, P_{yxd} and Q_{yxd} denote, respectively, the observed and forecasted precipitation at latitude y and longitude x accumulated over the period corresponding to observation record d, while θ_{yx} refers to the critical precipitation threshold, in this case, the N-year T-hour RPT of interest for the corresponding location. The FSS can be considered a measure of forecast skill that allows for spatial error in the placement of observed features or exceedances—up to a specified tolerance as prescribed by the evaluation radius—without penalty. As the evaluation radius reduces to zero, the FSS reduces to the Brier skill score with a reference forecast that exhibits no skill. An FSS of 1.0 indicates a perfect set of forecasts, while an FSS of 0.0 is the lowest possible score, and is indicative of a set of forecasts with no skill.

3. Results

a. *Threshold analysis*

The composite threshold maps for a 24-h AI appear in Fig. 1. As expected, threshold estimates increase monotonically with increasing RP. For the 1-yr RP, several parts of the country, in particular areas of the arid and Intermountain West, have 24-h RPTs of less than 25 mm, or 1 in. A few locations even have 2-yr 24-h RP thresholds below 1 in. In contrast, much wetter regions of the country such as the Pacific coastal mountains and southeast Gulf Coast region experience 1-yr average recurrence intervals (ARIs) at much higher precipitation thresholds of around 150 and 100 mm, respectively. Spanning nearly an order of magnitude, this highlights the stark contrast the RP framework brings relative to the traditional fixed threshold analysis approach. The relative regional relationships between RP thresholds tend to stay fairly similar at different RPs, with the Intermountain West remaining the lowest and the Pacific and Gulf coasts remaining the highest with respect to the required precipitation accumulation for a fixed frequency of occurrence. At the 100-yr RP, there are parts of the West with thresholds below 50 mm—lower than the 1-yr RP in other parts of the CONUS—and other places where the thresholds are in excess of 500 mm. Finally, close inspection reveals some spatial discontinuities when changing data sources for RP estimates. For example, TP-40 RP estimates for the South and Northeast appear to have been higher than for the updated Atlas 14 estimates; this can be clearly seen by inspection of the Texas–Oklahoma and New York–Pennsylvania borders in Fig. 1d. Though the complex terrain makes the comparison more difficult, Atlas 2 estimates in the Northwest appear to be much lower than their updated neighbors in places; inspect for example the eastern borders of Montana and Wyoming in Fig. 1a.

Corresponding grids for the 6-h AI appear in Fig. 2. Unsurprisingly, many of the patterns seen here are very similar to those seen in Fig. 1, just with all-around lower thresholds due to the shortened AI. The 1-yr RP thresholds range from roughly 10 to 100 mm, now with the highest thresholds seen in the Gulf Coast area rather than the Pacific coast; similarly, 100-yr RP thresholds produce accumulations that range from near 30 mm to approximately 300 mm. However, the biggest qualitative differences between Figs. 1 and 2 relate to regional differences in the nature of extreme precipitation events. In the West, extreme precipitation events are predominantly long-duration stratiform cases in which abundant moisture is advected from the ocean to the land and then that moisture precipitates out from lift by the coastal topography (e.g., Rutz et al. 2014; Villarini 2016). Because extreme precipitation events in this area tend to be of long-duration, low-to-moderate intensity, there is a large difference between the 6- and 24-h AI thresholds. In contrast, many extreme precipitation events in the East are driven by convective cells or convective systems (e.g., Brooks and Stensrud 2000; Schumacher and Johnson 2005, 2006; Stevenson and Schumacher 2014) and tend to be shorter-duration, higher-intensity events than seen in the West. As such, the difference between the 6- and 24-h thresholds is not as large, since many of the heavy precipitation events occur predominantly within a 6-h window. For example, the 1-yr RP for the Olympic Mountains of Washington reach nearly 200 mm for the 24-h AI, but this amount is reduced to roughly 80 mm at the same location for the 6-h AI. Compare this with central Iowa, where the 1-yr 24-h thresholds are near 65 mm, but the 1-yr 6-h thresholds are down to only near 55 mm. Many of the data source contrasts still exist in Fig. 2; the New York–Pennsylvania difference is significantly amplified in the 6-h thresholds compared with the 24-h threshold differences.

b. *Climatological and bias analysis*

1) 6-H ACCUMULATIONS: CAMS

A qualitative sense of the climatology of locally extreme precipitation over the United States may be readily discerned from the inspection of Figs. 3d, 4d,

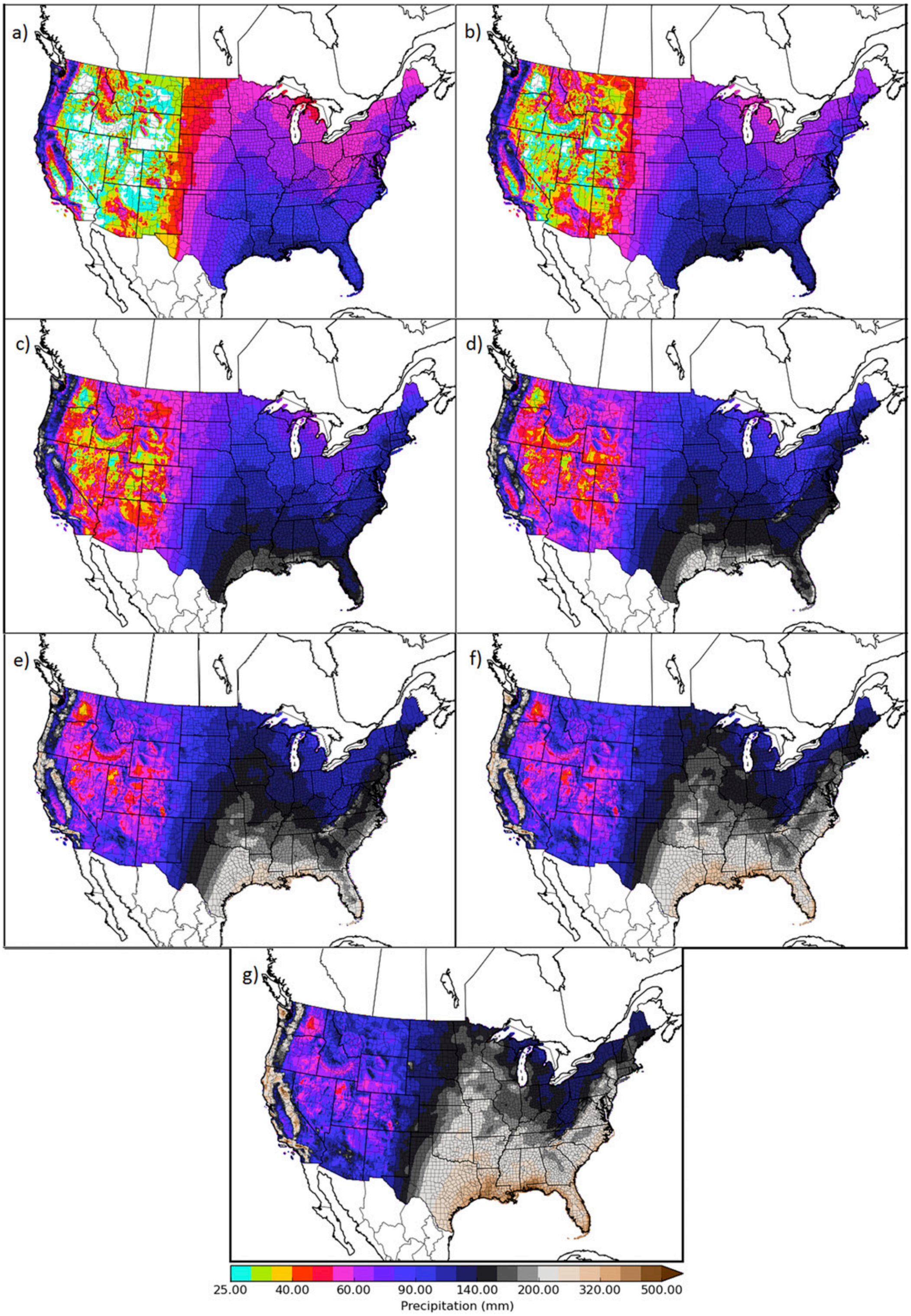

FIG. 1. Return period thresholds over the CONUS for a 24-h accumulation interval: (a) 1-, (b) 2-, (c) 5-, (d) 10-, (e) 25-, (f) 50-, and (g) 100-yr return period thresholds. Threshold sources come from a combination of Atlas 14, TP-40, and Atlas 2 data as described in the text.

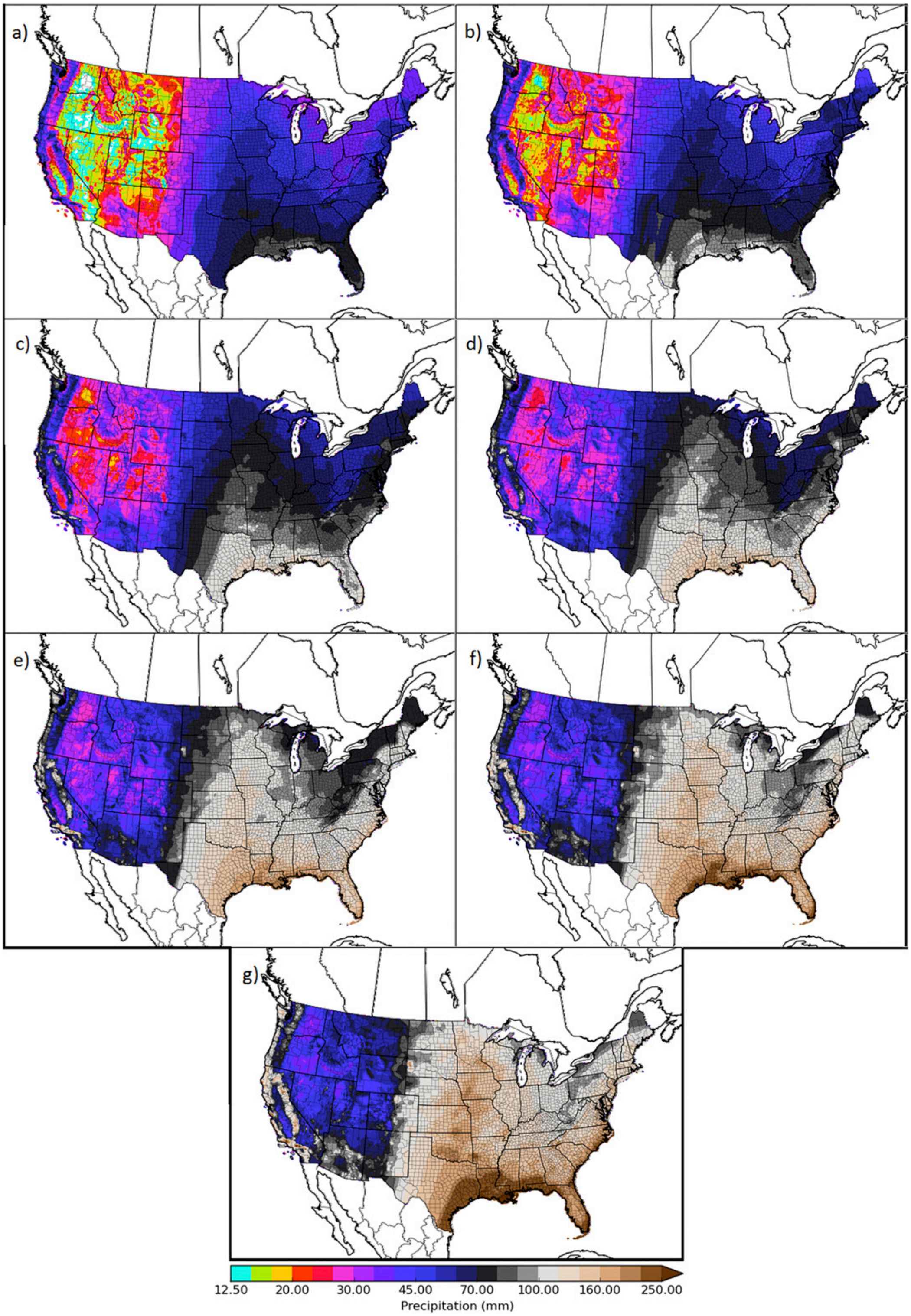

FIG. 2. As in Fig. 1, but for a 6-h accumulation interval.

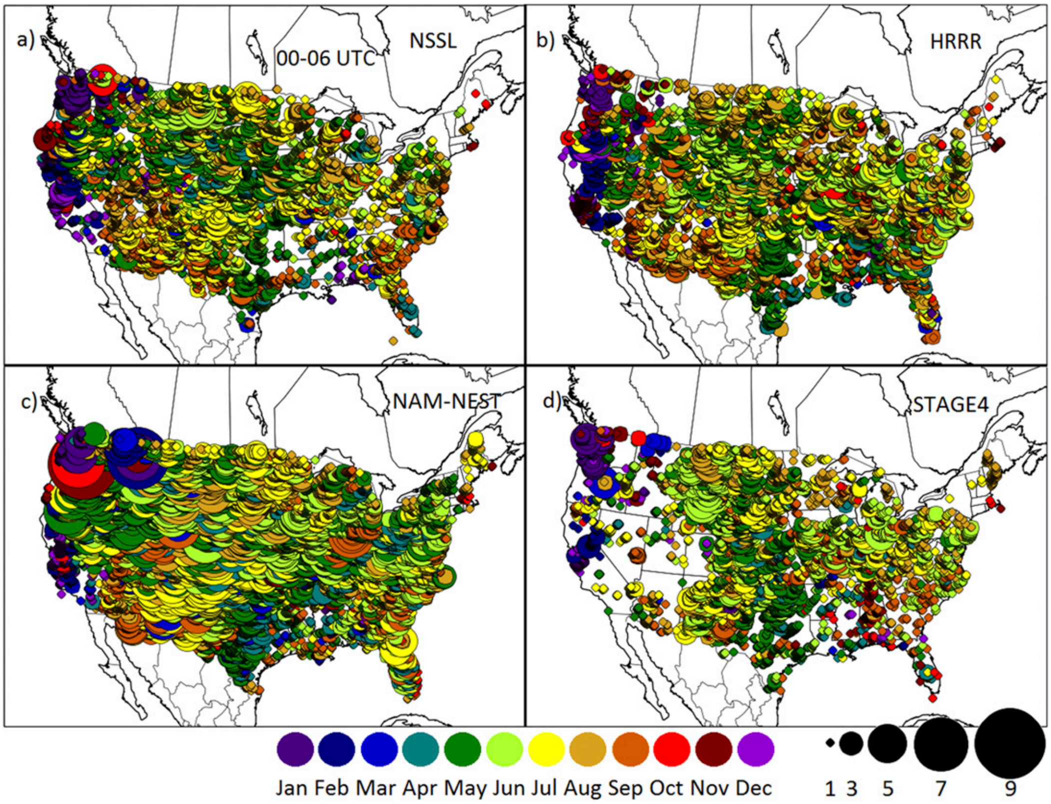

FIG. 3. Forecasted and observed exceedances of the 2-yr return period precipitation threshold for a 6-h accumulation interval, as illustrated in Fig. 2b, over the 0000–0600 UTC period from 12 Aug 2014 through 11 Aug 2015. Circles indicate an observed or forecasted event at the location of the circle center; circle size is proportional to the number of events, as indicated by the black circle legend in the figure's bottom-right corner. Forecasted events from the (a) NSSL-WRF, (b) HRRR, and (c) NAM-NEST forecasts initialized at 0000 UTC. These results correspond to forecast lead times of 24–30 h for NSSL-WRF and NAM-NEST in (a) and (c) and 0–6 h for HRRR in (b). The 24–30-h precipitation accumulations come from 0000 UTC initializations. (d) Observed RPT exceedances based on stage IV analysis. Circle colors indicate the mode month of event occurrence as depicted in the figure legend. For clarity, every other grid point in each dimension is assessed; only-one quarter of the total number of grid points is analyzed.

5d, and 6d , illustrating observed 2-yr RPT exceedances for four different 6-h accumulation periods over the 12 August 2014–11 August 2015 analysis period. The climatology depicted here is in general accord with previous findings (e.g., Schumacher and Johnson 2006; Hitchens et al. 2013; Stevenson and Schumacher 2014). Most events, both forecast and observed, occur during the cool season months of October–March in the Pacific coast states. In contrast, to the east of the Rockies, the vast majority of events occur in the warm season months from April through September. In particular, the central United States from Texas up through North Dakota is seen to experience most events during the early part of the warm season—primarily from April through July—while almost all events over the eastern states are seen between June and September. The southeast United States has perhaps the most diverse collection of identified events seasonally, with several events being identified in both the warm and cool seasons (e.g., Moore et al. 2015). However, there is a distinct maximum in identified events during the mid-to-late tropical cyclone season, from August through October, as many extreme precipitation events in this region are associated with tropical cyclone activity (e.g., Schumacher and Johnson 2006). Over the plains and the Midwest, it is also noteworthy that, as identified in previous studies and observations, precipitation systems tend to shift north climatologically throughout the warm season, with many of the observed events in the upper Midwest occurring in August. Finally, some data anomalies are worth noting. There is a stark decline in the number of observed events crossing from Pennsylvania into New York and further into New England; this is also reflective of the dramatic increase in 2-yr

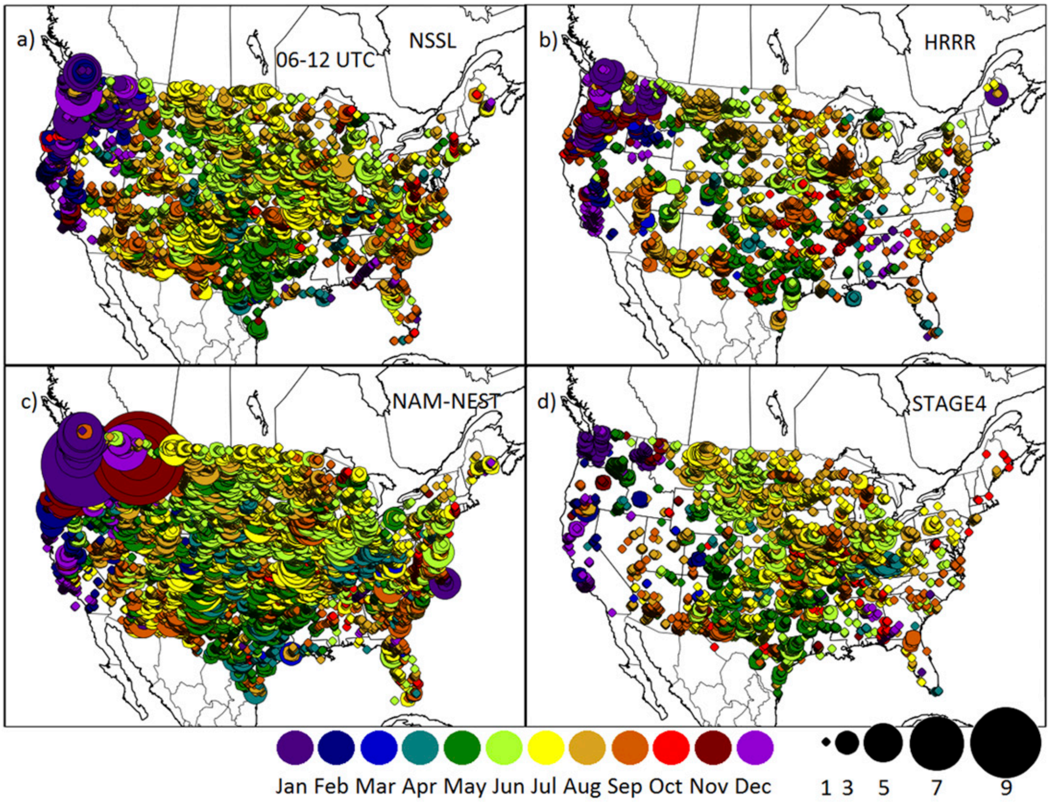

FIG. 4. As in Fig. 3, but for the 0600–1200 UTC period. NSSL-WRF, NAM-NEST, and HRRR forecasts are taken from the 6–12-h precipitation forecast from the 0000 UTC initialization.

6-h RP thresholds noted in Fig. 2b. Given that the number of events in Pennsylvania and surrounding states to the south and west is in rough proximity to the number of 2-yr events one would anticipate seeing over a 1-yr period for a given 6-h interval, this suggests that the 6-h thresholds derived from TP-40 data are likely too high. There are also only a small number of identified observed events in the southwest United States in Arizona, Utah, Nevada, and southeast California. Unlike the northeast U.S. disparity, this local minimum in events is seen only in the stage IV verification and not in the model forecasts. This minimum is likely attributable not to unrealistic thresholds in the Atlas 14 threshold estimates, but may instead be due to poor precipitation verification estimates due to the complex terrain and small population, yielding poor radar and gauge coverage in this region (e.g., Nelson et al. 2016).

Sunrise—the beginning of a meteorological day—occurs over the CONUS predominantly during the 1200–1800 UTC timespan depicted in Fig. 5d. Comparing Fig. 5d with Figs. 3d, 4d, and 6d, one notes that this period, likely because of the limited time for solar heating effects to operate, corresponds with a local minimum in extreme precipitation events across the CONUS, as indicated by a minimum in the number of plotted circles. Spatially, observed events are approximately uniformly distributed during this period, with events observed over all regions, but no particular areas of event concentration. Proceeding to the 1800–0000 UTC period illustrated in Fig. 6d, a substantial increase in events is observed, and two geographic regions appear particularly susceptible to extreme precipitation over this period. First, a concentration of observed events is seen along the Rocky Mountains, from New Mexico north through Montana. The Rocky Mountains serve as a forcing for convective initiation, and the time coincides with the onset of substantial solar radiation over the region (e.g., Schumacher and Johnson 2006); the combination results in many convective storms over this region during this time interval and an associated enhanced frequency of extreme precipitation accumulations. Additionally, a secondary maximum is observed near the Atlantic coast, from Florida through Maine, with a local minimum in observed events over the Mississippi valley between the two maxima. Progressing temporally to the 0000–0600 UTC period (Fig. 3d), an eastward progression of the extreme-precipitation-producing storms

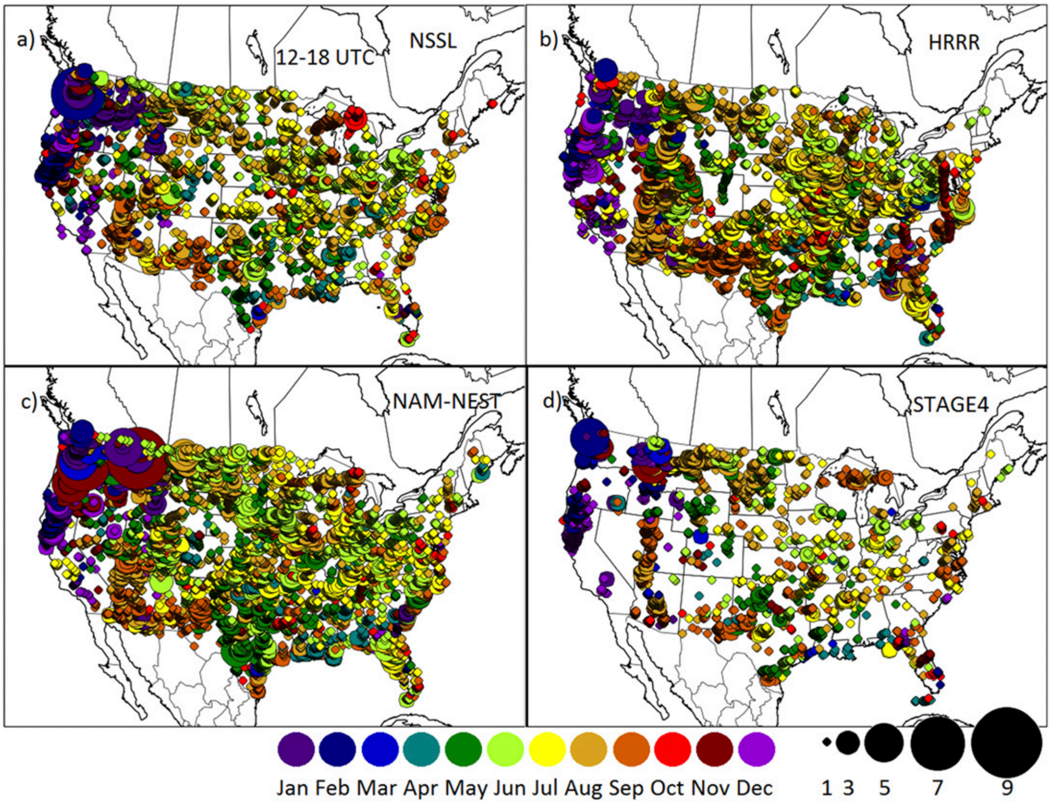

FIG. 5. As in Fig. 3, but for the 1200–1800 UTC period. NSSL-WRF and NAM-NEST forecasts are taken from the 12–18-h precipitation forecast from the 0000 UTC initialization. HRRR forecasts are taken from the 0–6-h forecast from the 1200 UTC initialization.

noted in the previous period is observed, with many events now observed from western Texas north through North Dakota. To the east, more events are observed over the Mississippi and especially the Ohio River valley as well, without the Atlantic maximum seen over the previous period. Further eastward progression of the Rocky Mountains–sourced storms is indicated in the final 0600–1200 UTC period (Fig. 4d), with many events observed farther to the east over Arkansas, Missouri, Iowa, Illinois, and Indiana, in addition to areas of the central plains farther to the west. A minimum in extreme precipitation events is seen over the far eastern states during this period. No clear diurnal cycle to the 6-h locally extreme precipitation events can be discerned from these figures for the states to the west of the Rocky Mountains, with predominantly cold season events observed to be approximately uniformly distributed across these four periods. The same observations of the climatology of 6-h AI locally extreme precipitation over CONUS can be seen by inspection of the 50-yr RP through Figs. 7d, 8d, 9d, and 10d . Many of the minima and maxima appear to be more pronounced at this higher RP, with minima being especially evident as a result of the increased event rarity.

By comparing panels a–c with panel d in Figs. 3–10, qualitative aspects of the climatology and bias characteristics of extreme precipitation in different numerical models may be deduced. The seasonality of local extreme precipitation events in the CAMs assessed here appears to be in approximate agreement with the observations at each location, though springtime events appear to be especially overforecast in many of the CAMs in the Intermountain West, Southeast, and the mid-Atlantic regions. However, the diurnal cycles exhibited in the CAMs do have some distinct differences with respect to the identified regions of minimum and maximum threat at different times throughout the day when compared with the stage IV analysis. First, a phase and intensity bias in areas of complex terrain that has been noted previously in the literature (e.g., Schwitalla et al. 2008; Weckwerth et al. 2014), in which convection is initiated too early and is too strong on the windward side of topography, appears prominently here in Figs. 5 and 9. Other differences are particularly

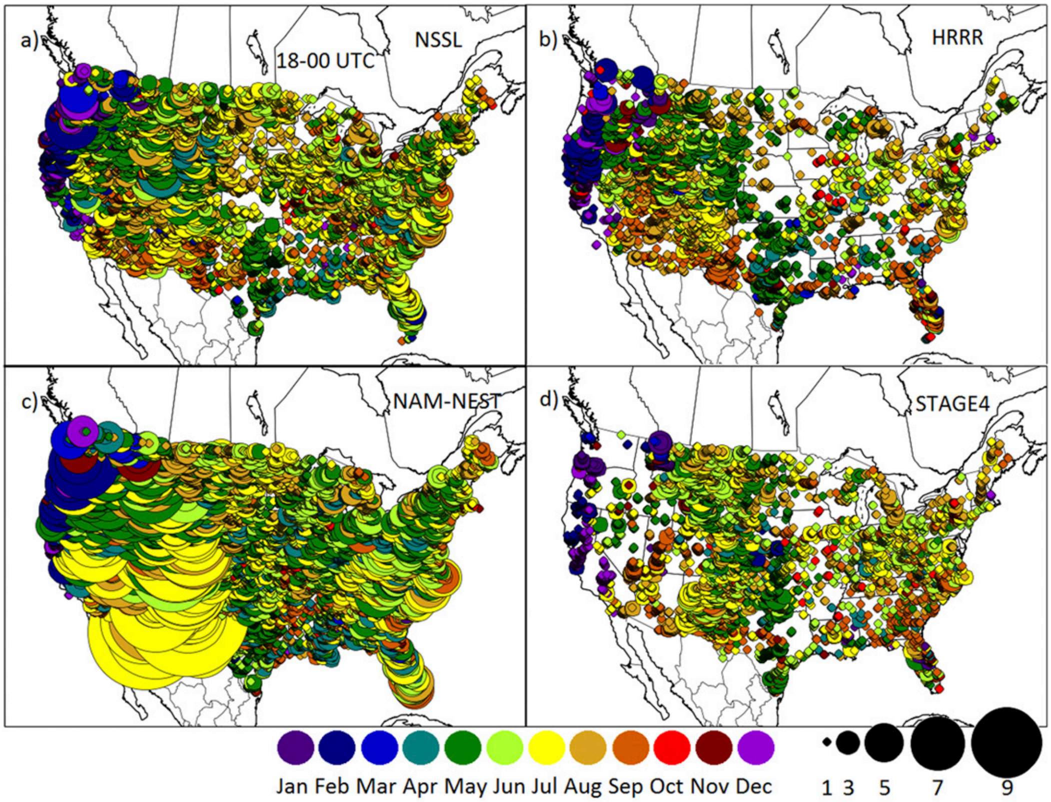

FIG. 6. As in Fig. 3, but for the 1800–0000 UTC period. NSSL-WRF and NAM-NEST forecasts are taken from the 18–24-h precipitation forecast from the 0000 UTC initialization. HRRR forecasts are taken from the 6–12-h forecast from the 1200 UTC initialization.

evident at the 50-yr RP, for example in Figs. 7 and 10. The event concentration to the immediate east of the Rocky Mountains seen in the observations in Fig. 7d is not apparent, or at least much less so, in the three CAMs analyzed here. Similarly in Fig. 10, the Rocky Mountains and Atlantic maxima seen in stage IV are evident to an extent in each of the CAMs, especially via the minimum depicted over the Great Plains, but the trend is still muted compared with what is observed. Further, each of the models depicts a maximum over the Intermountain West that is not corroborated in the observations; this may be attributed to a combination of aforementioned observational deficiencies with the stage IV product in this region, in addition to a tendency in the models to overproduce and overly intensify orographic convection (e.g., Schwitalla et al. 2008; Weckwerth et al. 2014). The most apparent model–observation contrast seen in the model comparison is the tendency of the NAM-NEST to overforecast occurrences of locally extreme rainfall events at both the 2-yr and especially the 50-yr return periods all across the CONUS. While the NSSL-WRF and HRRR appear to forecast only a few more events than are observed, particularly over the western states, the NAM-NEST forecasts far more events than are observed for all four 6-h accumulation periods. The tendency to forecast extreme precipitation events with such high frequency clouds the ability to definitively discern its extreme precipitation diurnal cycle. This overforecast tendency is most pronounced during the 1800–0000 UTC period, as evidenced by the circle sizes produced in Fig. 10c, particularly over the West and Southwest. A more quantitative assessment of the CAM bias characteristics follows.

The frequency bias characteristics of the CAMs compared in Figs. 3–10 is summarized quantitatively in Fig. 11. Confirming what was seen in Figs. 3–10, for all four times of day, the most events are seen during the warm season months, with fewer events per month witnessed during the cool season. In the 0000–0600 and 0600–1200 UTC periods, the number of events per month varies on average by approximately an order of magnitude. This result is especially pronounced at the 50-yr RP, as many cool season months have few or no observed events during the 1-yr analysis period. In

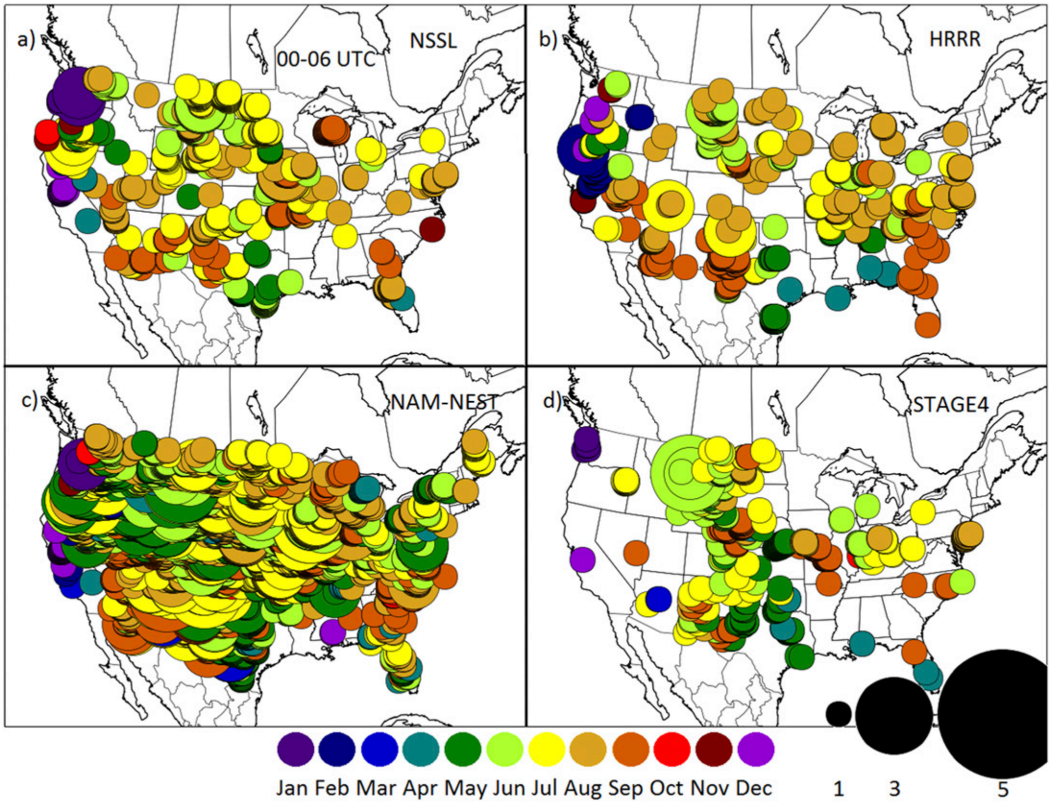

FIG. 7. As in Fig. 3, but for the 50-yr return period thresholds.

contrast, for the 1200–1800 UTC period, which is by far the least driven by warm season convection, the difference is much smaller, with the December event count even being higher than many of the warm season months in the stage IV analysis. As previously indicated, the 1200–1800 UTC period is found to have the smallest number of observed events, with 0000–0600 UTC being the largest. Also, despite large month-to-month variability, the average ratio of 2-yr events to 50-yr events is approximately 25, as one would expect a priori.

Comparing the total forecasted RP exceedances between the models (Fig. 11), the NAM-NEST does indeed forecast many more events than are observed. During the 0000–0600 UTC period, the NAM-NEST predicts more events than were observed for each month at the 2-yr RP; in the warm season months, the NAM-NEST forecasts roughly 3–4 times as many events as are observed during the 0000–0600 UTC period, while for the 50-yr RP, the quotient is typically 5–6 and is as large as 10—an order of magnitude difference—for some warm season months. The bias is roughly similar for the 0600–1200 and 1200–1800 UTC periods, but is further amplified in the 1800–0000 UTC period, with most of the warm season months having an order of magnitude more forecasted events than are observed, and in some cases the quotient is larger than 20 for the 50-yr RP. There are several possible reasons for these substantial observed biases; given that the extent of these biases was not seen in previous versions of the NAM-NEST (not shown), the magnitude of the biases is likely attributable to the August 2014 model updates, including the switch to the Ferrier–Aligo microphysics parameterization (NWS 2014). Except in some cool season months, the NSSL-WRF forecasted fewer events than the NAM-NEST each month over the analysis period for both the 2- and 50-yr RPs. For most warm season months, it is seen to be slightly positively biased relative to stage IV at the 2-yr RP, but the difference is typically less than a factor of 2. Though the month-to-month results are more variable for the 50-yr RP, the NSSL-WRF exhibits an even smaller positive bias, with several warm season months having fewer forecasted events than were observed. Too few 50-yr events were observed during the cool season months to draw any definitive conclusions about the bias characteristics of the CAMs during this period. The HRRR bias characteristics appear to be quite good as well, with the HRRR event count lines

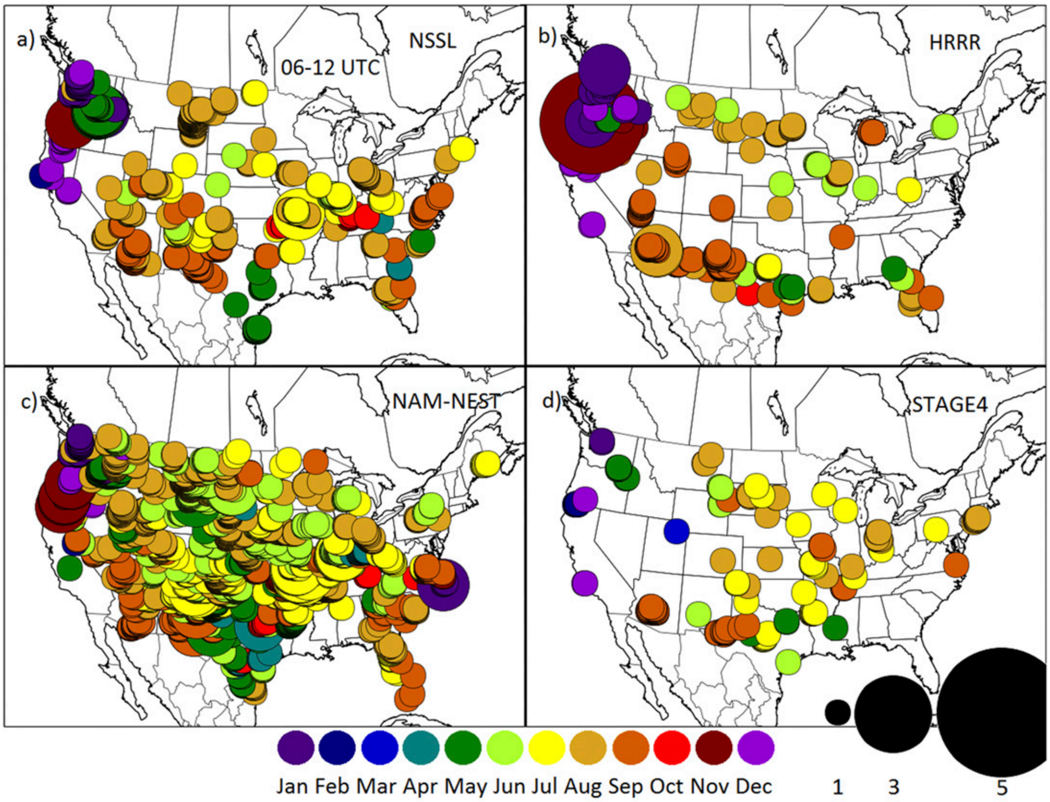

FIG. 8. As in Fig. 4, but for the 50-yr return period thresholds.

most closely tracking the stage IV lines during most period–month combinations. The most notable exception occurs during the 1200–1800 UTC period. Here, while the NAM-NEST and NSSL-WRF substantially overpredict the number of exceedances in May and June, particularly at the 50-yr RP, the HRRR predicts a much closer number of exceedances to what is observed. However, from August onward, the HRRR tends to predict more events than observations or either of the other analyzed CAMs for this 6-h period. This dichotomy may be attributable to the aforementioned HRRR model changes that occurred during this period of record, leading to improved bias characteristics later in the analysis period during April–July.

2) 24-H ACCUMULATIONS

Analysis of 24-h 1200–1200 UTC precipitation events using the longer 9 June 2009–30 August 2014 period is depicted in Fig. 12, with the GEFS/R replacing the NAM-NEST and HRRR for the 24-h analysis. It is immediately apparent that the coarse GEFS/R model forecasts far fewer events at both the 10-yr and particularly the 100-yr RPs than are actually observed. This is in stark contrast to the CAMs in the 6-h analysis, which all tended to be positively biased. This is likely attributable to the GEFS/R being unable to resolve many small-scale processes that contribute to the development of locally extreme precipitation events (e.g., Schwartz et al. 2009; Clark et al. 2010). Closer inspection of, for example, Fig. 12d confirms this: almost all of the GEFS/R forecasted events occur over the West Coast during the cool season, where the vast majority of heavy precipitation events occur in association with synoptic-scale systems supplying ample moisture to the region with stratiform precipitation occurring in association with lift from major, large-scale topographic features (e.g., Villarini 2016). All of these processes can be adequately handled even by a coarse model, and in fact, the Oregon and California coastal mountains are one of the only areas where the GEFS/R is seen (cf. Figs. 12c and 12e) to overforecast extreme precipitation by forecasting events that did not verify. Outside of the synoptic driven West Coast systems, the events that GEFS/R appears to be able to forecast appropriately extreme precipitation amounts only in cases with very strong synoptic-scale forcing, such as Tropical Cyclones Irene and Lee in August and September 2011, respectively, whose tracks and

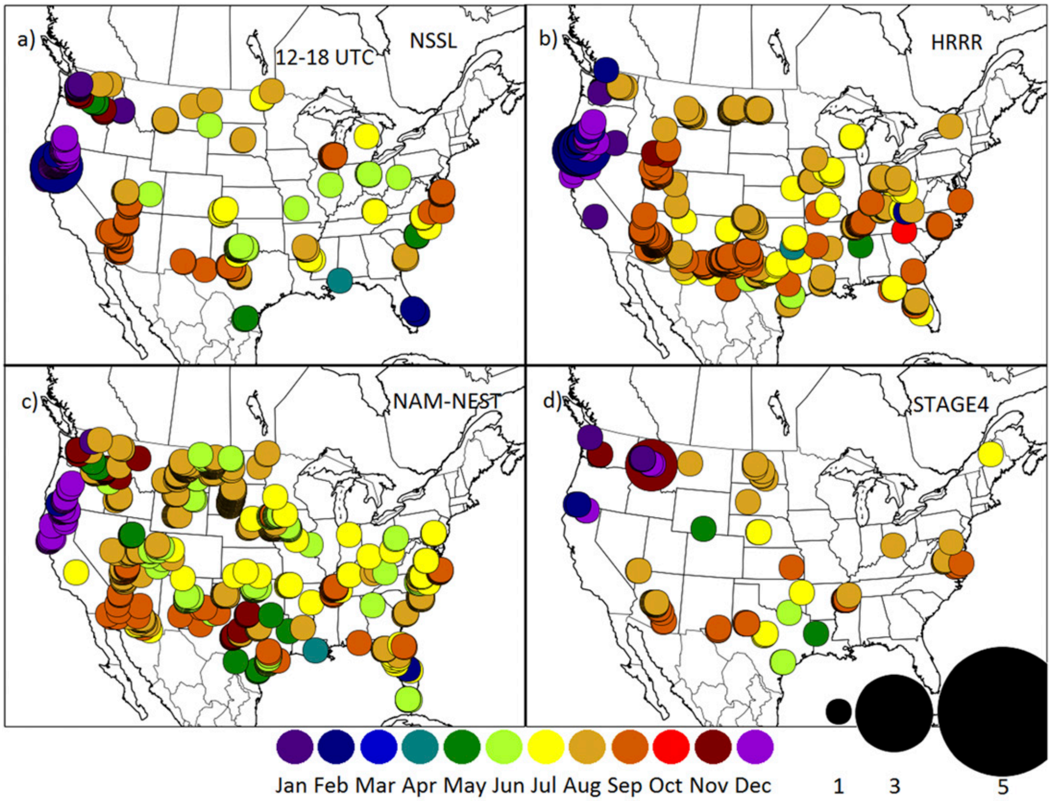

FIG. 9. As in Fig. 5, but for 50-yr return period thresholds.

associated swaths of heaviest precipitation are clearly outlined in Fig. 12d and less clearly in Fig. 12f (NWS 2012a,b). Other events include the major September 2009 Southeast flooding, which involved exceptional synoptic-scale moisture transport into the region (NWS 2010); the Arizona flooding of January 2010 (Parzybok et al. 2011; Neiman et al. 2013); and a stretch of exceptionally wet systems in Montana during May and June 2013. The intensity of other synoptic-scale systems that produced 100-yr 24-h precipitation events, such as Hurricane Sandy in October 2012 (see red circles in Fig. 12f; Blake et al. 2013; NWS 2013), were not forecasted by GEFS/R. The NSSL-WRF (Fig. 12b) still performed better with some of the tropical cyclones, more adequately forecasting the rainfall amounts associated with both Hurricane Sandy and also with Tropical Storm Debby, which affected northern Florida in June 2012 (Kimberlain 2013). However, no system on the plains or in the Midwest, regardless of whether it was observed to have actually occurred, is forecasted by the GEFS/R model, in spite of the fact that many events were observed in the stage IV analysis. The NSSL-WRF has much more robust forecast characteristics, correctly forecasting many August/September events in New Mexico and a smattering of isolated events throughout the warm season across the plains. The relative minimum of events in the southern plains (Fig. 12e) is also well captured (Fig. 12a). As in the 6-h analysis, the NSSL-WRF does tend to forecast many events in the Intermountain West that do not verify, but again, this may be at least partially attributable to poor precipitation verification in this region. The general seasonality of model events is in accord with the true 24-h locally extreme precipitation analysis as determined by stage IV precipitation analysis.

Event count analysis for the 24-h accumulation interval is illustrated in Fig. 13, with lines for the 1-, 5-, 25-, and 100-yr RPs included. The seasonal cycle of events is similar to that of the 6-h accumulation intervals, with more events per month typically observed during the warm season months, but the signal is considerably attenuated. The signal does, however, amplify with increasing return period; that is, the difference between the number of events during the cool season and warm season months increases with increasing RP. At low RPs, the bias characteristics of both the GEFS/R and NSSL-WRF are both fairly good—within a factor of 2 from unity each month. However, at higher RPs, some discernable biases

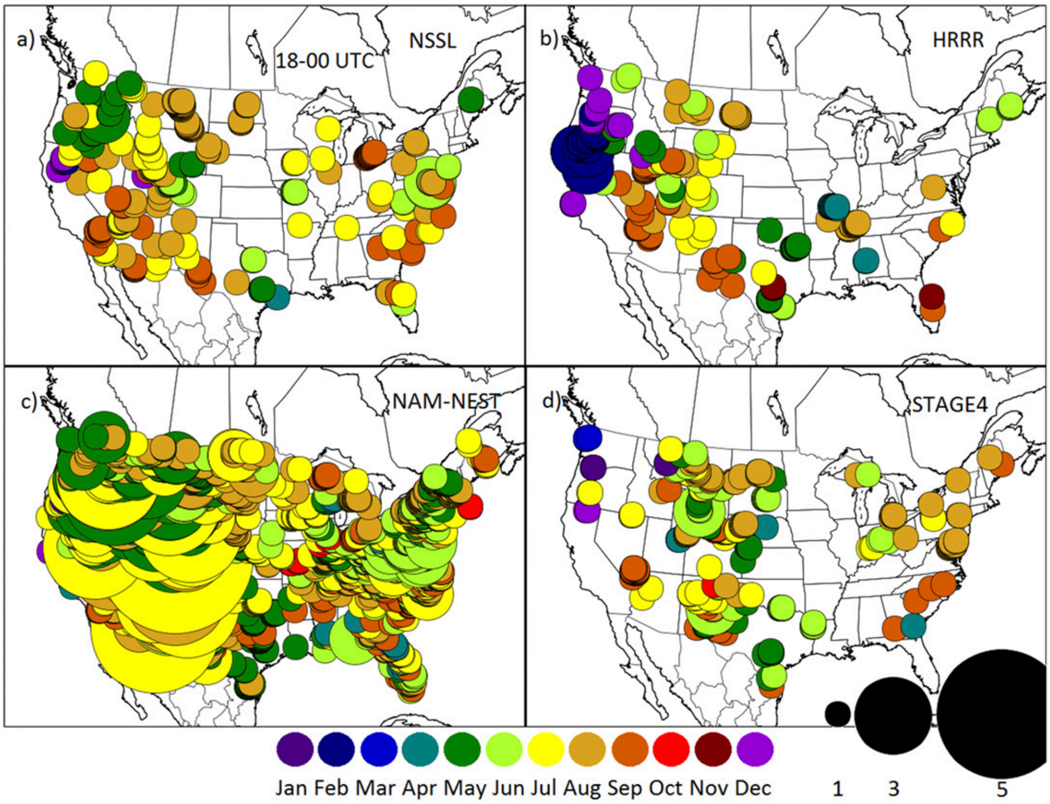

FIG. 10. As in Fig. 6, but for 50-yr return period thresholds.

begin to emerge. The NSSL-WRF overforecasts high-RP 24-h locally extreme precipitation events during the cool season months, with observed frequency biases of approximately 3–5 during this period. Figure 12 shows that GEFS/R greatly underpredicts high-RP events. Figure 13 reveals that the overall bias characteristics, while almost always underpredicting, are well within an order of magnitude difference in most months. In July, however, GEFS/R underforecasts the number of events by an exceptional two orders of magnitude relative to the relatively large number of events observed during that month over the +5-yr analysis period.

c. *Skill analysis*

1) 6-H ACCUMULATIONS: NSSL-WRF

The remaining analysis quantitatively evaluates model skill for the NSSL-WRF over the longer 9 June 2009–30 August 2014 analysis period; other CAMs are not assessed in this capacity because of a lack of data availability and consistency over this period. Fraction skill score results at each RP assessed in this study for 6-h NSSL-WRF forecasts between forecast hours 12 and 36 on the 0000 UTC initialization are depicted in Fig. 14. FSS values generally increase with increasing evaluation radius, and the highest scores are seen at lower RPs, which, being more common, are typically more predictable. At all evaluation radii and RPs, the highest scores by a considerable amount are observed with the 12–18-h forecast period. This is likely attributable to the combination of 1) the fact that, as previously noted, a lower proportion of events in this 6-h period is convectively driven than the other periods, and the events caused by smaller convective cells with weaker large-scale forcing tend to be inherently less predictable (e.g., Doswell et al. 1996); and 2) this is the earliest forecast period analyzed with respect to forecast initialization time and thus has had the least time for nonlinear error growth. In a similar vein, the FSSs at all evaluation radii are very low for the 50- and 100-yr RPs at the 24- and 30-h lead times, probably because of these periods being the farthest from the model initialization and these both being convectively active times of day. The low values at high-evaluation radii also highlight that errors were not merely in spatial displacement but were primarily to completely missing events that occurred and forecasting widespread extreme events that did not verify. The FSSs are considerably better for the 12- and 18-h forecast lead times,

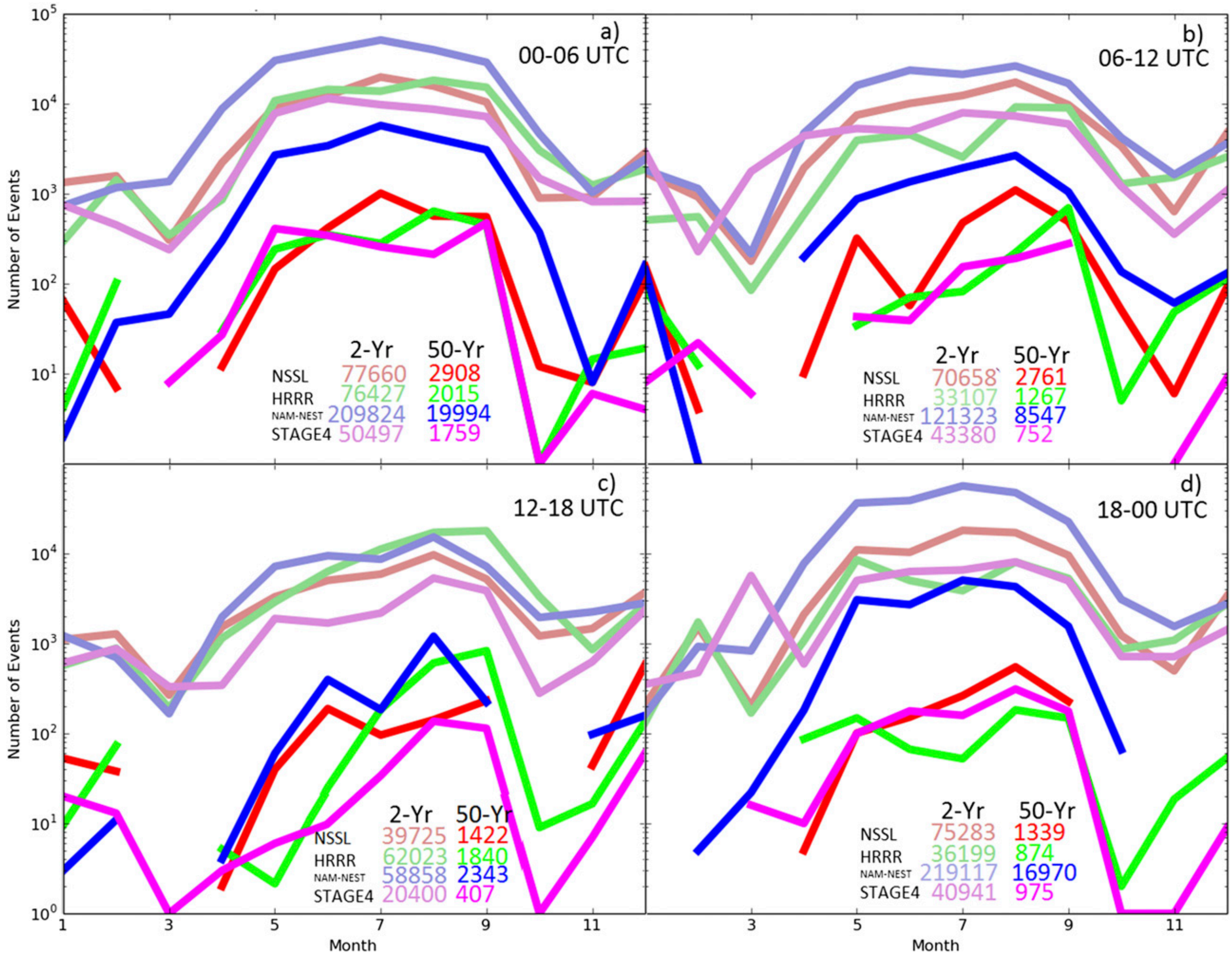

FIG. 11. Total number of events forecasted or observed over the 12 Aug 2014–11 Aug 2015 verification period for 6-h precipitation accumulations for the NSSL-WRF, HRRR, and NAM-NEST models compared against stage IV precipitation analysis. Shown are counts for the (a) 0000–0600, (b) 0600–1200, (c) 1200–1800, and (d) 1800–0000 UTC periods. Event count is plotted on a logarithmic scale, for return periods of both 2 and 50 yr. A discontinuity in a line indicates that no event was forecasted or observed for the data source and return period in question over the verification period for that month. Some runs of the HRRR were not run or did not complete; HRRR event counts have been rescaled in proportion to the number of missing dates in each respective month. Numbers indicate the total number of identified events over all months for each period, with text colors having the same associations as the line colors.

better even than the 5-yr RP verification for the last periods at the high-evaluation radii.

A graphical representation of NSSL-WRF model skill for 6-h locally extreme rainfall forecasts over the CONUS appears in Fig. 15. These plots have been aggregated over all four 6-h forecast periods. The general trend of decreasing forecast skill with increasing RP is immediately apparent by inspection of the figure. However, these figures allow one to discern the impact of a particular region's forecast on and in comparison with overall model performance. At low RPs, several events often impact a given region over the analysis period, resulting in a smoother FSS field, as seen in Fig. 15a. In contrast, at the high RPs, for example in Fig. 15d, often none or just one event occurred over a given region during the 5-yr analysis period, resulting in the highlighting of areas where the model handled one event well. Figures 15a and 15b illustrate that the broad areas of extreme precipitation, as determined by low RP exceedances, in the California and New England systems that occurred during this period were very well handled by the NSSL-WRF model, with skill scores not too far from unity. The same two events reappear in Figs. 15c and 15d as well, indicating that the severity of the systems was also appropriately forecast by the model. Two other regions are also highlighted in the high-RP panels, Montana and the mid-Atlantic, in association

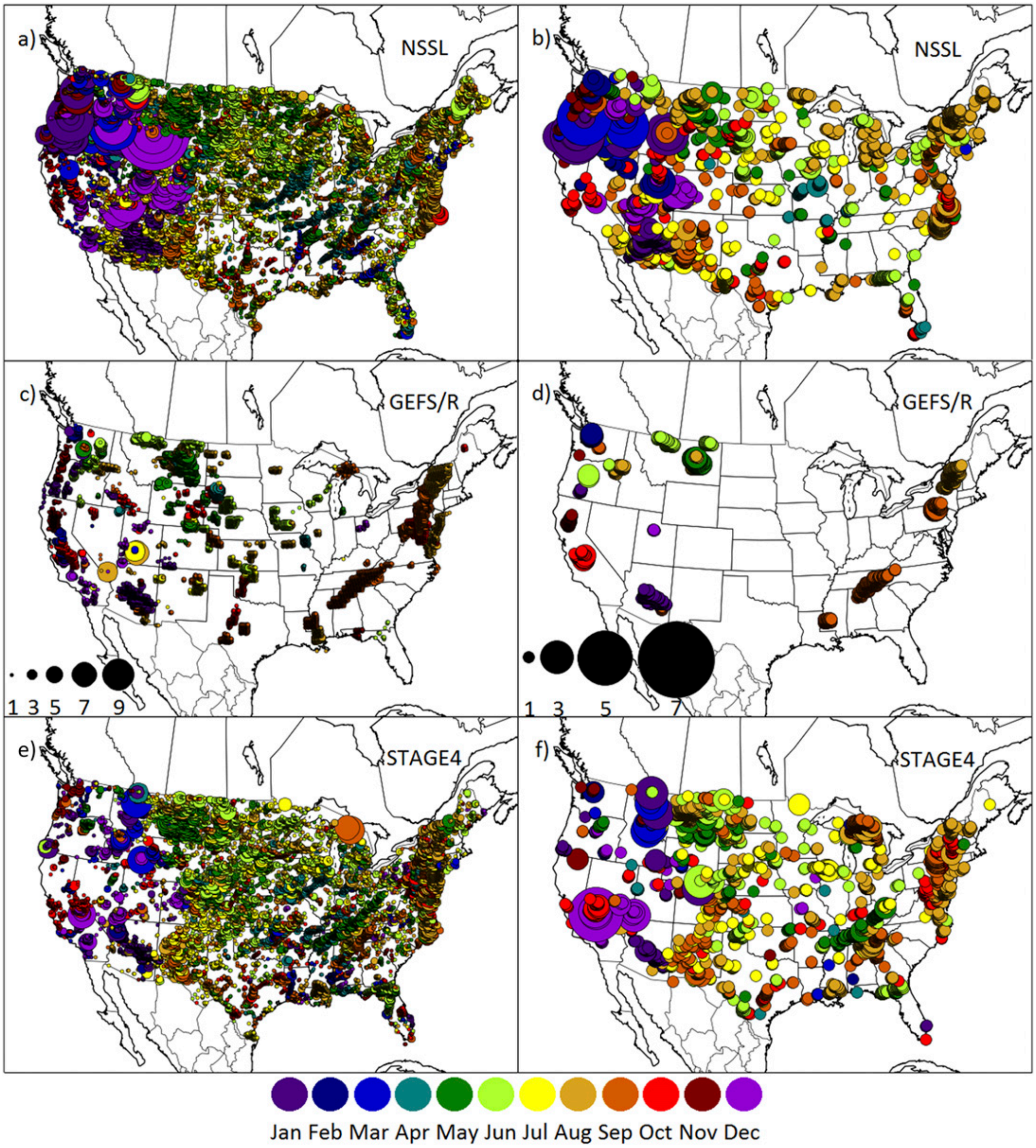

FIG. 12. (a),(c),(e) Forecasted and (b),(d),(f) observed events of exceedances of 10- and 100-yr return period thresholds for a 24-h accumulation interval from 9 Jun 2009 through 30 Aug 2014. Circles indicate an observed or forecasted event at the location of the circle center; larger circles indicate more collocated events, as specified by the black circles in the figure legends. Results in (a),(b) and (c),(d) correspond to forecasted threshold exceedances from forecast hours 12–36 of the 0000 UTC initialization of the NSSL-WRF and GEFS/R, respectively. Results in (e),(f) correspond to observed threshold exceedances based on stage IV precipitation analysis during the same evaluation period. Circle colors indicate the mode month of event occurrence as depicted in the figure legend. For clarity, every other grid point in each dimension is assessed; only one-quarter of the total number of grid points is analyzed.

with the aforementioned May Montana floods and Tropical Storm Lee, respectively; this suggests that the NSSL-WRF may have forecast the severity of the events well, but did not perform quite as well when forecasting the spatial extent of the locally extreme rainfall. Other events, such as Tropical Storm Debby, were reasonably well forecast during the 25-yr RP but not at the 100-yr RP. On the less extreme spectrum of extreme rainfall events, areas of the mountain and desert West and Southwest were not well handled; this was seen qualitatively in Figs. 3–10. For the most extreme event threshold, the Florida peninsula and the central plains were the areas with the worst verifying forecasts.

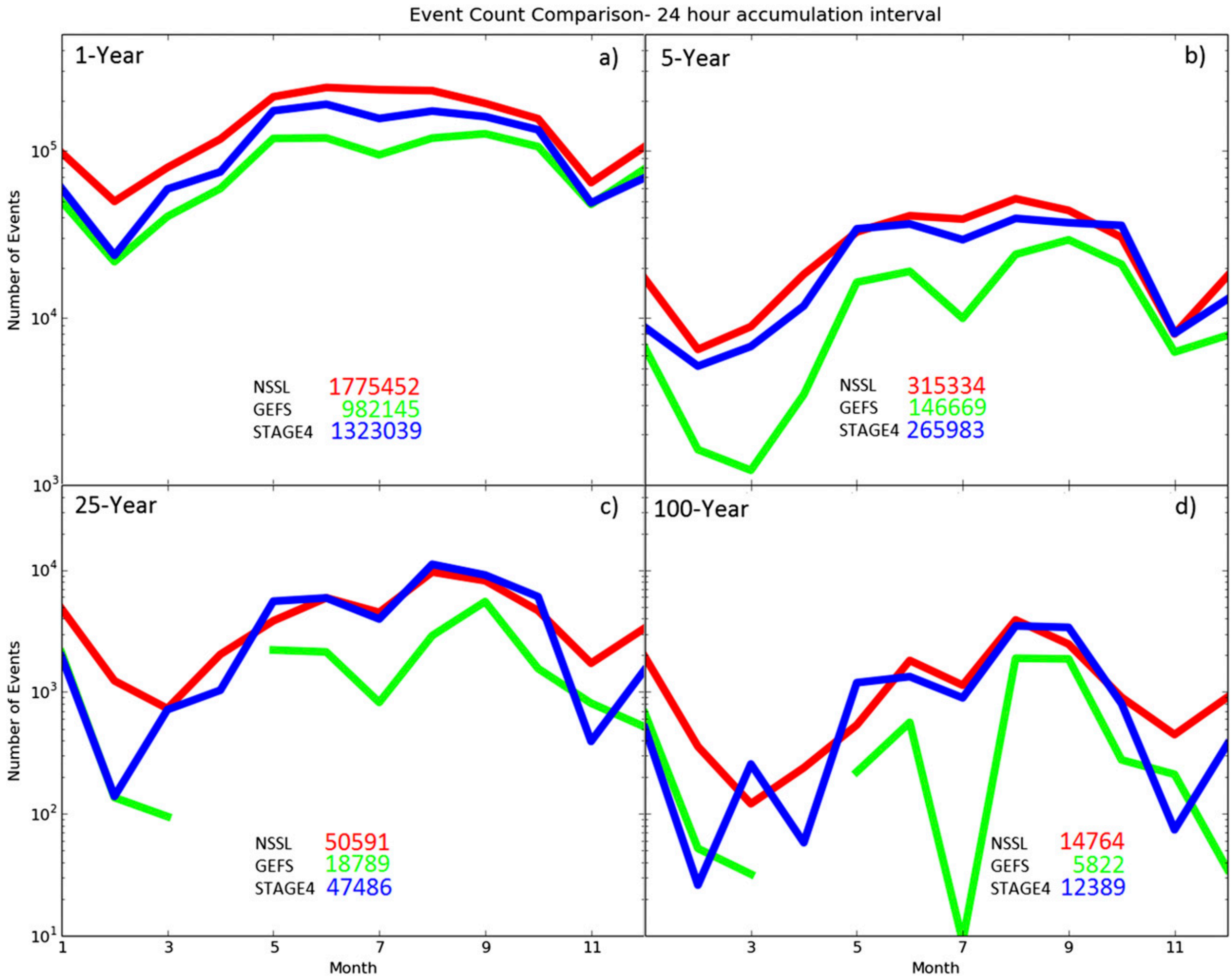

FIG. 13. Total number of events forecasted or observed over the 6 Jun 2009–30 Aug 2014 verification period for 24-h 1200–1200 UTC precipitation accumulations for the NSSL-WRF and GEFS/R models compared against stage IV precipitation analysis. Event count is plotted on a logarithmic scale; return periods of 1, 5, 25, and 100 yr are shown for each data source. A discontinuity in a line indicates that no event was forecasted or observed for the data source and return period in question over the verification period for that month.

Figure 16 provides regional skill analysis for different times of day as opposed to the different RPs shown in Fig. 15. Much of the variance in the Northeast comes from the largely coincidental timing of tropical cyclone landfall for the storms that impacted the region during this period. However, some trends can be more definitively discerned. Though 2-yr, 6-h events were observed during all four periods (see Figs. 3d, 5d, 7d, and 9d) in the Pacific Northwest, skill was observed to be notably higher during 0600–1800 than 1800–0600 UTC. In regions that experience extreme precipitation events from various types of systems and meteorological conditions, such as the southeast United States, one sees substantially elevated skill during the less convective 1200–1800 UTC period in comparison with the other three. To the east of the Rockies, in particular eastern Colorado and western Kansas and Nebraska, the lowest skill is seen during the 0600–1200 UTC period, likely attributable to the locally anomalous time of occurrence of these types of events; typically, ongoing convection is well to the east of this region at this time (e.g., Stevenson and Schumacher 2014). The highest forecast skill in the upper Midwest is seen during the 0000–0600 UTC period, while across the CONUS in the desert Southwest, skill is higher at this time and lowest 6 h later in the 0600–1200 UTC period. Isolated regions of relatively high and low skill appear at various locations and times in the lower Mississippi valley in association with particular extreme-rain-producing storm systems that were well and poorly forecast, respectively.

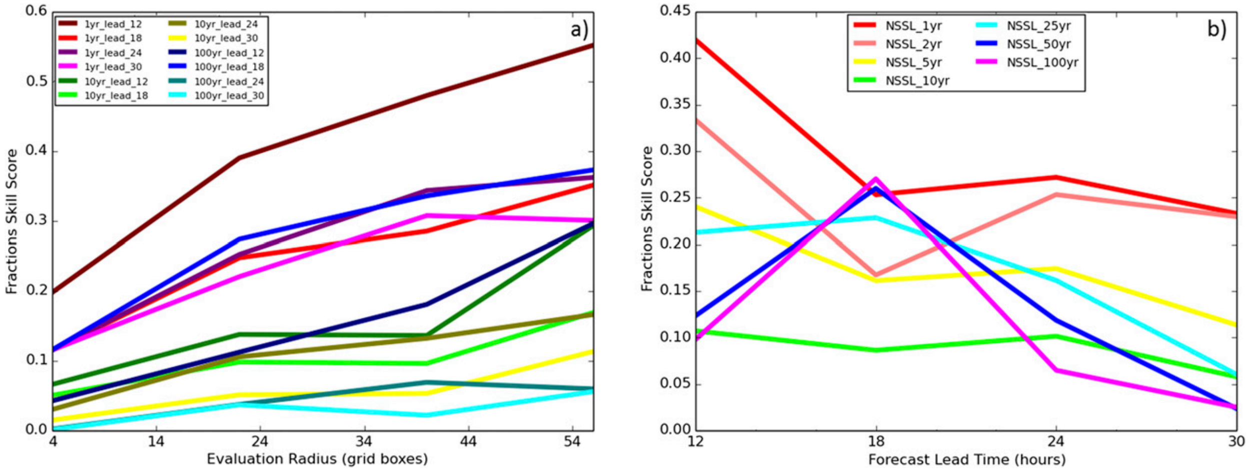

FIG. 14. (a) Aggregated FSSs for the NSSL-WRF for the 6-h accumulation interval for the 1-, 10-, and 100-yr return periods. (b) FSS vs lead time for all return periods, at an evaluation radius of 24 grid boxes (~110 km). Verification is performed over the 9 Jun 2009–30 Aug 2014 period. Forecasts are taken from the 0000 UTC initialization; ergo, lines indicated to have a lead of 12 correspond to the 1200–1800 UTC forecast period, leads of 18 to the 1800–0000 UTC period, 24 to the 0000–0600 UTC period, and 30 to the 0600–1200 UTC period.

2) 24-H ACCUMULATIONS

Summary aggregated FSS analysis for 24-h accumulations at all RPs is provided for both the GEFS/R and NSSL-WRF models in Fig. 17. FSS values again generally increase with increasing evaluation radius and decrease with increasing RP. Comparing Figs. 14 and 17, skill scores are generally higher for the 24-h accumulation interval compared with any of the 6-h periods. The higher skill at longer accumulation intervals is likely attributable to the decreased sensitivity to temporal and to a lesser extent spatial displacement error, in addition to a larger proportion of 24-h events occurring in association with longer-lived, larger-scale processes as opposed to 6-h events, which occur at higher frequency in association with isolated convective cells (e.g., Stevenson and Schumacher 2014). At low and very high RPs, the NSSL-WRF appreciably outperforms the GEFS/R at all evaluation radii examined here. All else being equal, the improved model resolution and ability to explicitly simulate convection without use of a cumulus parameterization should tend to produce more realistic and accurate representations of precipitation processes, thereby yielding these results. However, at the middle RPs, from 10 to 50 yr, the GEFS/R is competitive with and often even outperforms the NSSL-WRF. This may have to do with the types of systems that are found to exceed thresholds in this frequency range, but not higher or lower: to generate precipitation of this rarity, strong large-scale forcing is required. Generating the most extreme events, with a 100-yr RP or greater, may often require a combination of large-scale forcing and meso- and smaller-scale forcings, which cannot be adequately simulated by GEFS/R and other models of similar horizontal resolution.

Regional representations of model performance for 24-h events appear in Figs. 18 and 19 for the NSSL-WRF and GEFS/R, respectively; regional model performance is directly compared via Fig. 20. Inspecting the 1-yr RP verification in Figs. 18a and 19a, it is apparent that slightly elevated skill is seen over much of the West and East Coasts, while skill over the plains, the Midwest, and the Southeast is reduced. Though local fluctuations are seen, the FSS field remains relatively smooth, with skill scores remaining between 0.1 and 0.9. Larger FSS gradients are observed at the 5-yr RP in association with forecast quality of individual events; California, Arizona, Montana, the mid-Atlantic, the Iowa–Missouri–Illinois region, and various parts of the southeast United States begin to stand out as areas of locally enhanced forecast skill. These regions are further highlighted moving to the 25-yr RP in Fig. 18c; Montana and the Florida panhandle in association largely with Tropical Storm Debby are particularly notable, with FSSs approaching unity. The mid-Atlantic, California, and Arizona skill results occur in association primarily with particular extreme precipitation events discussed with Fig. 12. Previous research has found that the highlighted region of enhanced skill in the Midwest here coincides with a local maximum in Lagrangian persistence (e.g., Germann et al. 2006), that is, the tendency for storm motion, both speed and direction, to remain the same. The elevated Lagrangian persistence here is reasonably

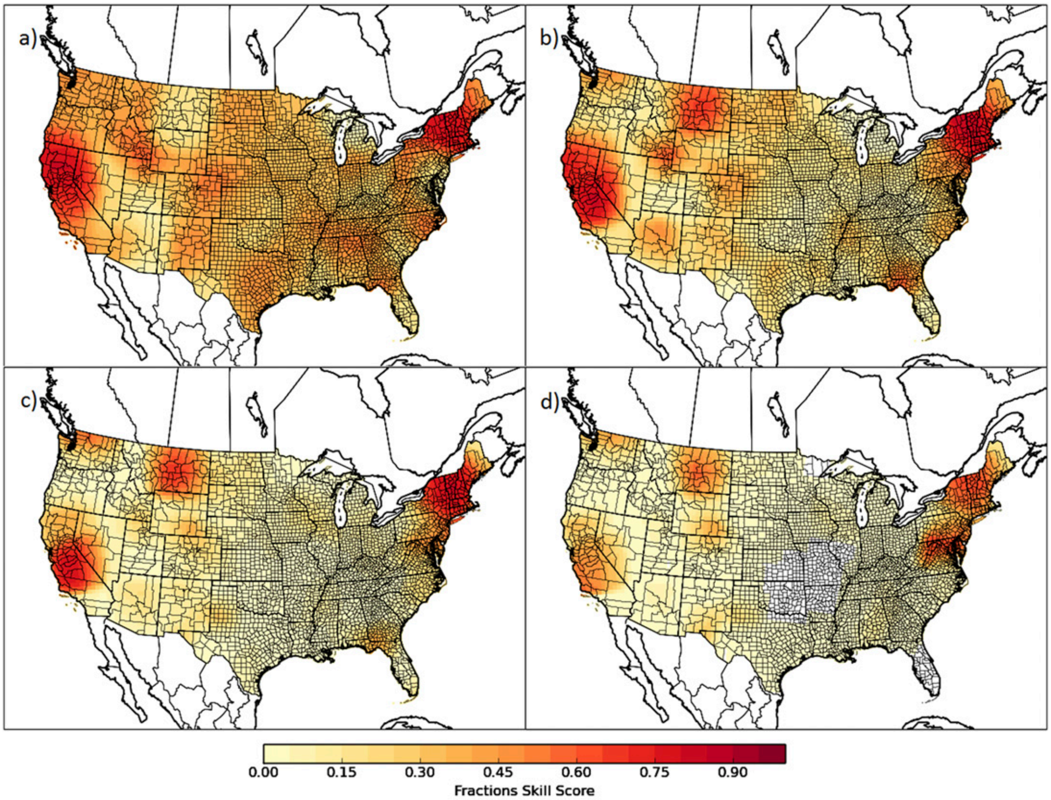

FIG. 15. Gridded aggregated FSSs for the NSSL-WRF for 6-h precipitation forecasts aggregated over each of the 12–18-, 18–24-, 24–30-, and 30–36-h forecast periods for the 0000 UTC model initialization. Verifications on the (a) 1-, (b) 5-, (c) 25-, and (d) 100-yr return periods. Verification is performed over the 9 Jun 2009–30 Aug 2014 period. FSSs correspond to an evaluation radius of 40 grid boxes on the stage IV HRAP grid.

associated with locally enhanced precipitation forecast skill. The 100-yr RP sees a further degradation of forecast skill overall, with a very noisy assessment of regional skill due to the model performance in single events.

The trends in the GEFS/R are rather similar. Comparing the two at the 1-yr RP (Fig. 20a), it is apparent that, despite point-to-point fluctuations, both the GEFS/R and NSSL-WRF perform about equally well in the western states. In the eastern two-thirds of the CONUS, with the exception of some areas of areas near the Gulf Coast, the NSSL-WRF exhibits higher skill than the GEFS/R, particularly over the convection-dominated regions of the plains, the Midwest, and the Mississippi valley. The region of enhanced Lagrangian persistence in the Midwest also sees the local area of largest skill difference between the NSSL-WRF and GEFS/R, with the NSSL-WRF performing notably better. It is logical that, with the higher model resolution, the NSSL-WRF has more realistic representations of convection and is, thus, able to better take advantage of the enhanced persistence, as opposed to the GEFS/R, which does not benefit much from this property. Inspecting the 5- and 25-yr comparisons in Figs. 20b and 20c, particular events can be identified: the NSSL-WRF better handled Tropical Storm Debby, the New Mexico flooding of September 2013, and to a lesser extent the Montana floods; meanwhile, GEFS/R had better forecasts for the aforementioned January 2010 and September 2009 events in Arizona and the southeast United States, respectively, in addition to better TC forecasts in the mid-Atlantic and New England. With the exception of the northeast CONUS, where the model performances become more similar, these differences are exacerbated at the 100-yr RP (Fig. 20d).

4. Discussion and conclusions

NWP model verification and diagnosis from the fixed-frequency recurrence interval/return period perspective was performed for a suite of dynamical models of varying spatial scales, from the global, convection-parameterized GEFS/R to regional, convection-permitting models such as the NSSL-WRF, NAM-NEST, and HRRR.

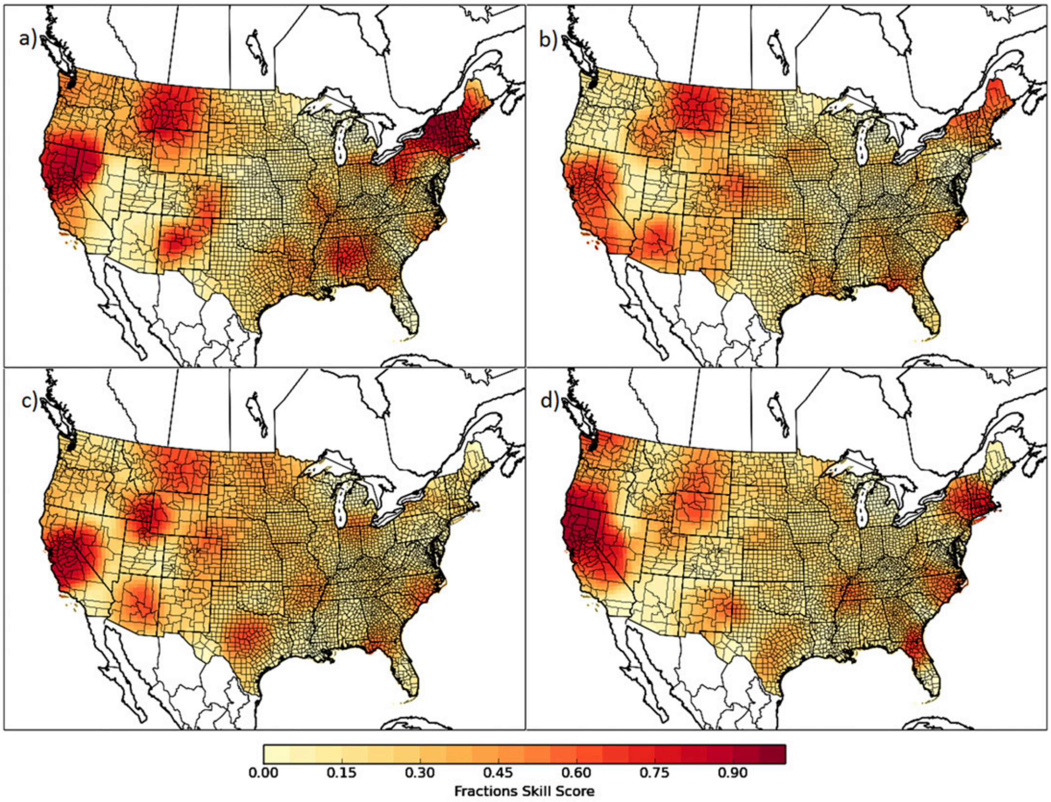

FIG. 16. Aggregated FSSs for the 0000 UTC initialization of the NSSL-WRF for 6-h accumulated precipitation forecasts verified for 2-yr return period thresholds. Verifications over forecast hours (a) 12–18 (1200–1800 UTC), (b) 18–24 (1800–0000 UTC), (c) 24–30 (0000–0600 UTC), and (d) 30–36 (0600–1200 UTC). Verification is performed over the 9 Jun 2009–30 Aug 2014 period. FSSs correspond to an evaluation radius of 40 grid boxes on the stage IV HRAP grid.

CONUS-wide RPT grids for 6- and 24-h AIs were assembled from existing observational estimates for RPs between 1 and 100 yr. Bulk and regional bias characteristics were assessed individually for each modeling system. Major differences were identified in the behavior of locally extreme precipitation production between models. The most recent (as of August 2014) NAM-NEST was found to have a strong positive frequency bias nationwide, forecasting many more events at all return periods than were actually observed. This effect was particularly evident in the southwestern states, where the NAM-NEST over a 1-yr period forecasted an order of magnitude more events than were actually identified via stage IV precipitation analysis. Of the four 6-h accumulation periods centered about 0000 UTC, the NAM-NEST was found to exhibit the largest bias during the convective initiation period of 1800–0000 UTC, and the least in the minimally convective 1200–1800 UTC period. Other CAMs—the NSSL-WRF and HRRR—exhibited similar tendencies, also tending to overforecast extreme events from 1- to 100-yr RPs, and tended to have the strongest tendency to overforecast in areas of the West and Southwest, but both demonstrated greatly reduced overall frequency biases when compared with the NAM-NEST, with the HRRR actually being negatively biased at times. The NAM-NEST and to a much lesser extent other CAMs tended to be slightly more biased for both high and low RPs during the warm season months. At the 24-h AI, the NSSL-WRF was found to exhibit rather similar extreme precipitation characteristics to those seen in the 6-h AI. The coarser GEFS/R, however, had much different characteristics than the CAMs, producing almost no events at the higher RPs outside of the cool season Pacific coast synoptic systems and tropical cyclones from the Atlantic basin, resulting in almost no very extreme events forecast by the GEFS/R over the Great Plains and Midwest. With regard to model skill, models are unsurprisingly found to make the most skillful predictions of the less extreme (lower RP) events, with the worst skill typically observed in forecasting for

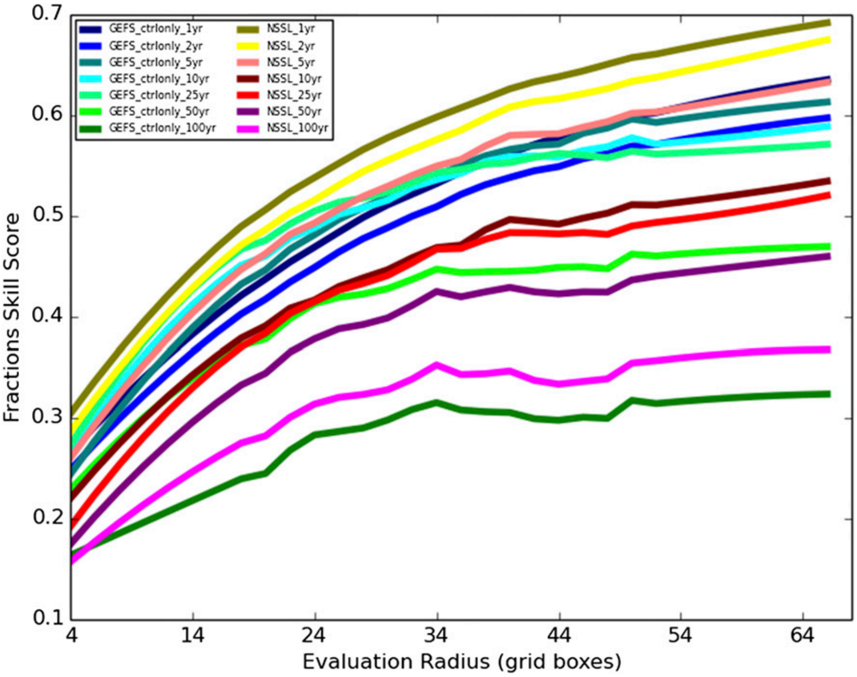

FIG. 17. Aggregated FSSs for GEFS/R and NSSL-WRF for the 24-h accumulation interval for the 1-, 2-, 5-, 10-, 25-, 50-, and 100-yr return periods. Verification is performed over the 9 Jun 2009–30 Aug 2014 period. Forecasts taken from the 12–36-h forecast of each model's 0000 UTC initialization. Results have been lightly smoothed with a Savitzky–Golay filter for clarity.

the rarest of event thresholds. For 6-h periods, the 1200–1800 UTC time frame stood out as the most skillful period for the NSSL-WRF, which is likely attributable to this period having the lowest proportion of low-predictability, small-scale convective events, and this period coinciding with the shortest forecast lead time of the four verification periods. Both the GEFS/R and NSSL-WRF were verified for the 24-h AI; at low RPs, locally higher skill was generally seen in both models in the West and the mid-Atlantic/New England, and the NSSL-WRF demonstrated superior forecast skill over most of the CONUS, particularly over the Great Plains and Midwest, and the Mississippi valley. At higher RPs, typically one event dominated the regional skill score, allowing comparison of model performance for individual recent extreme precipitation events but leaving insufficient data to robustly compare regionally compare model skill for highly extreme cases. Overall, the NSSL-WRF is found to demonstrate superior forecast skill at low and very high RPs, while the GEFS/R is competitive and occasionally outperforms the NSSL-WRF at moderate RPs.

These findings present several noteworthy implications to operations. First, care was taken in this study to ensure that the models evaluated were as consistent as possible during the analysis periods selected, and the precipitation and RPT estimates were also as consistent as possible within the context of the data sources available at the time this research was conducted. The findings illustrate the value of having a static, unchanged, model over a substantial period of time: one is able to note and correct for identified model biases based on the previous performance history of the model or models in a manner that would not be feasible if the model were routinely updated in ways that altered its bias characteristics. This can be done both through objective, quantitative means (e.g., Marsh et al. 2012; Scheuerer and Hamill 2015), but even absent that, operational forecasters frequently note a preference for inspecting output from more familiar models, where they are acquainted with its strengths and weaknesses and can subjectively identify when and where it may be biased or inaccurate. Here, forecasters can apply the findings of this research to downplay the importance of large QPFs portrayed by the NAM-NEST, and also downplay the importance of warm season QPFs in convective regions for low-resolution models such as the GEFS, in addition to making subtler regional or seasonal corrections based on the observed extreme

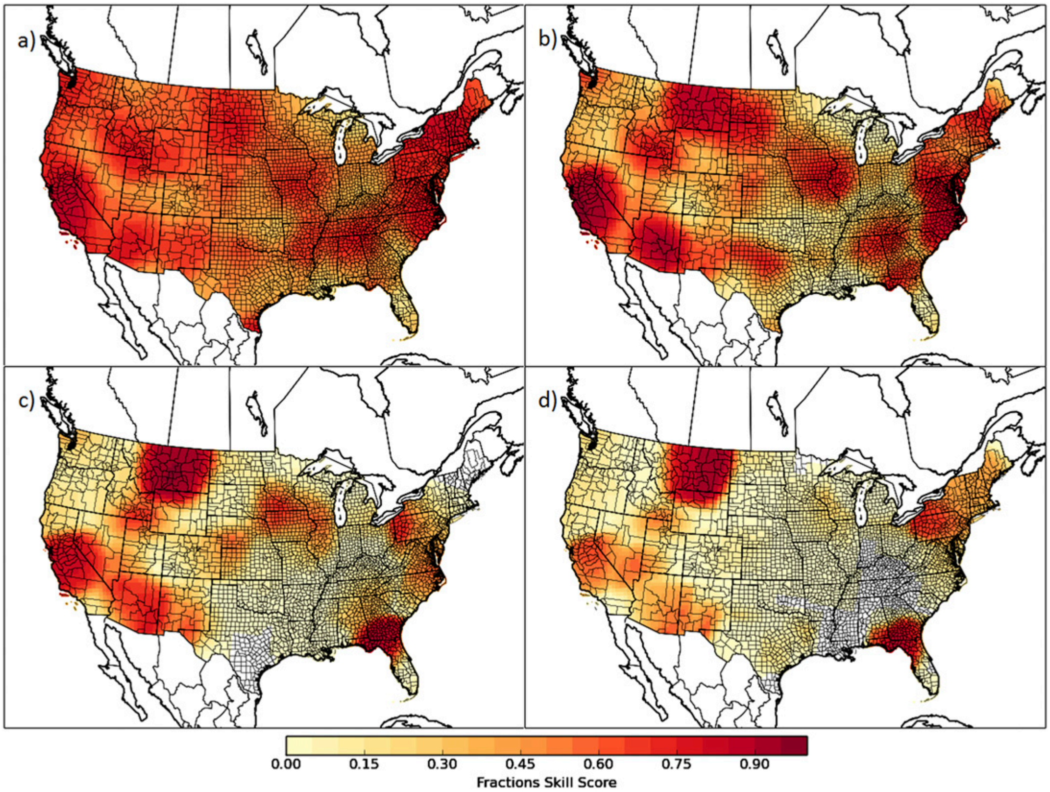

FIG. 18. Gridded aggregated FSSs for the NSSL-WRF for 24-h precipitation forecasts from the 12–36-h forecasts of the 0000 UTC model initialization. Verifications on the (a) 1-, (b) 5-, (c) 25-, and (d) 100-yr return periods. Verification is performed over the 9 Jun 2009–30 Aug 2014 period. FSSs correspond to an evaluation radius of 40 grid boxes on the stage IV HRAP grid.

threshold exceedance characteristics for the individual models. Second, changes to improve one aspect of a model's performance may often degrade another. For example, the August 2014 changes to the NAM-NEST have been noted to have produced more realistic-looking convective structures, particularly with regard to supercells (NWS 2014; Aligo et al. 2014), but as this study's analysis has illustrated, the same changes that improved the appearance of the convective structures resulted in a strong increase in QPF bias, at least at high thresholds. In the future, care should be taken when changing a model to clearly define the objectives of such changes and assure that the changes fulfill those goals. Third, at least within the framework of the verification presented herein, it appears that while increased model resolution is useful for improving verification of extreme precipitation caused by convection, it is not an all-around panacea; the coarse GEFS/R model performed as well or better than the high-resolution guidance at extreme QPFs for most other types of extreme-precipitation-producing systems.

Several notable insights may be gleaned regarding future verification and forecasting research as well. This work verifies the principal claims of Hamill and Juras (2006): verification across time and especially space within a fixed-threshold verification framework may often produce results that reflect the varying climatology, with bulk performance characteristics biased toward performance in climatologically favored times or regions. This may be desirable in some use cases, but when trying to diagnose general model properties applicable over the entire domain evaluated, based on the contrast between the results presented here and previous studies using fixed-threshold verification (e.g., Brooks and Stensrud 2000; Hitchens et al. 2013; Novak et al. 2014, among others), it is apparent that using a fixed-frequency framework such as that afforded through RPs or recurrence intervals is critical for a successful diagnosis. Development of consistent sets of climatology-aware thresholds will assist in the verification of models for future studies. Toward this end, the future completion of the NOAA Atlas 14 project will be of substantial benefit for future studies of extreme

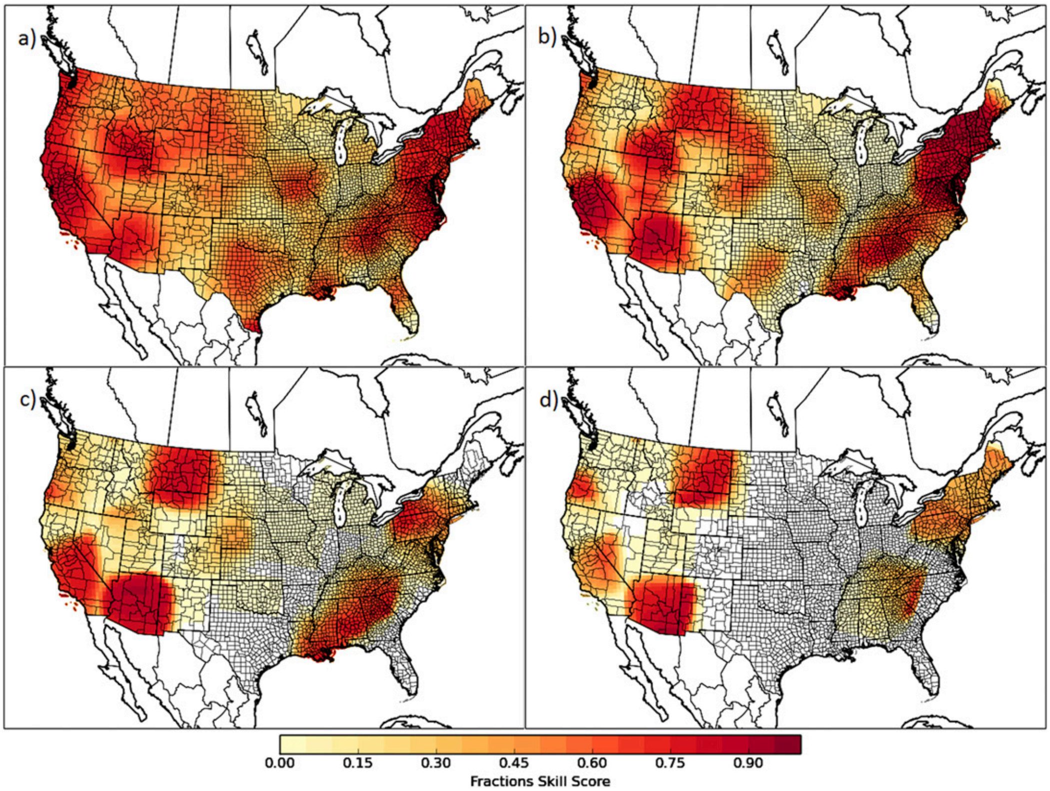

FIG. 19. As in Fig. 18, but for GEFS/R.

QPFs. Additionally, it should be noted that all of the analysis performed in this study examined only the raw QPFs (albeit regridded) from each of the models examined. The intent and scope of this research was to identify biases and evaluate the performance of the various modeling systems examined. Future work will examine the important question of how the raw NWP output used in this research can be optimally postprocessed by means of downscaling, quantile mapping, machine learning, probability calibration, and other methods to produce skillful deterministic and probabilistic predictions of RPT exceedances.

Acknowledgments. The authors thank Daniel Cooley, Elizabeth Barnes, Erik Nielsen, Stacey Hitchcock, Lynn McMurdie, and three anonymous reviewers for their helpful insights and input on this research. The authors also wish to thank Eric James for generously providing HRRR data used in this research, Adam Clark for providing NSSL-WRF data, and Tom Hamill for generating and providing GEFS/R data. Funding for this research was supported by NOAA Award NA14OAR4320125, Amendment 26; UCAR/COMET Award Z16-23465; National Science Foundation Grant ACI-1450089; and NASA Grant NNX15AD11G.

APPENDIX

Quality Control Details

Some points frequently and persistently report very large precipitation totals, usually because they are located in complex terrain and continuously report very large radar reflectivity from the nearest radar, resulting in unrealistically high associated automated accumulated precipitation estimates. Because of other priorities with the internal QC process used by RFCs, this is not always removed from the final stage IV product. Additionally, on some days, large regions of exceptionally high precipitation totals are reported and are not removed from the final product. A combination of automated and manual means is used to combat these issues. Exploiting that RPT exceedances, by definition, occur with some known specified frequency p, correlations of model QPF time series were performed to very crudely approximate e-folding times (in days) and distances (in grid points) beyond which one can assume independence of events. Given this, for any return period examined, one can readily formulate a forecast day or set of forecasts at a point as a series of independent Bernoulli trials in which the event occurs with probability p. Using

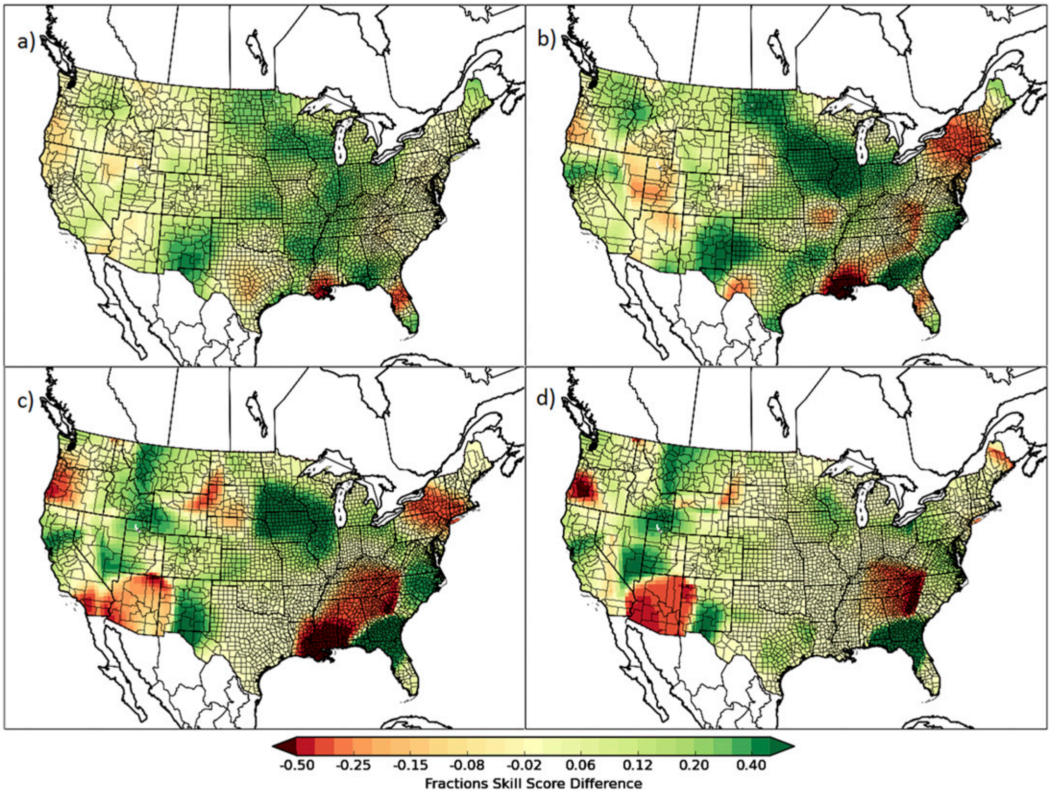

FIG. 20. As in Fig. 18, but instead showing the difference between NSSL and GEFS/R performance over the verification period. Greens indicate that the NSSL-WRF performed better over the region, while reds indicate that the GEFS/R performed better.

the binomial distribution, the a priori probability of experiencing at least *k* events may be readily tabulated. For each RP examined and for all days and points, any occurrence that exceeded the 99.99% percentile for the expected event count from the binomial distribution was flagged for removal from the dataset. These were subsequently manually perused to ascertain whether the rejection was legitimate, and if so, the recorded events from that day or location were removed from the dataset.

REFERENCES

Alexander, C., and Coauthors, 2015: Rapid Refresh (RAP) v3.0; High-Resolution Rapid Refresh (HRRR) v2.0. *WCOSS Science Quarterly*, NOAA, 27 pp. [Available online http://ruc.noaa.gov/pdf/RAPHRRR_WCOSS_2016Q1_Final-sb-12oct2015.pdf.]

Aligo, E., B. Ferrier, J. Carley, E. Rogers, M. Pyle, S. Weiss, and I. Jirak, 2014: Modified microphysics for use in high-resolution NAM forecasts. *Proc. 27th Conf. on Severe Local Storms*, Madison, WI, Amer. Meteor. Soc., 16A.1. [Available online at https://ams.confex.com/ams/27SLS/webprogram/Paper255732.html.]

Baxter, M. A., G. M. Lackmann, K. M. Mahoney, T. E. Workoff, and T. M. Hamill, 2014: Verification of quantitative precipitation reforecasts over the southeastern United States. *Wea. Forecasting*, **29**, 1199–1207, doi:10.1175/WAF-D-14-00055.1.

Benjamin, S. G., and Coauthors, 2016: A North American hourly assimilation and model forecast cycle: The Rapid Refresh. *Mon. Wea. Rev.*, **144**, 1669–1694, doi:10.1175/MWR-D-15-0242.1.

Blake, E. S., T. B. Kimberlain, R. J. Berg, J. P. Cangialosi, and J. L. Beven II, 2013: Tropical Cyclone report: Hurricane Sandy (AL182012) 22–29 October 2012. National Hurricane Center, 157 pp. [Available online at http://www.nhc.noaa.gov/data/tcr/AL182012_Sandy.pdf.]

Bonnin, G. M., D. Martin, B. Lin, T. Parzybok, M. Yekta, and D. Riley, 2004: *Precipitation-Frequency Atlas of the United States*. NOAA Atlas 14, Vol. 1, 271 pp. [Available online at http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume1.pdf.]

——, ——, ——, ——, ——, and ——, 2006: *Precipitation-Frequency Atlas of the United States*. NOAA Atlas 14, Vol. 2, 301 pp. [Available online at http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume2.pdf.]

Brooks, H. E., and D. J. Stensrud, 2000: Climatology of heavy rain events in the United States from hourly precipitation observations. *Mon. Wea. Rev.*, **128**, 1194–1201, doi:10.1175/1520-0493(2000)128<1194:COHREI>2.0.CO;2.

Clark, A. J., W. A. Gallus Jr., and M. L. Weisman, 2010: Neighborhood-based verification of precipitation forecasts from convection-allowing NCAR WRF model simulations and the operational NAM. *Wea. Forecasting*, **25**, 1495–1509, doi:10.1175/2010WAF2222404.1.

Doswell, C. A., III, H. E. Brooks, and R. A. Maddox, 1996: Flash flood forecasting: An ingredients-based methodology. *Wea. Forecasting*, **11**, 560–581, doi:10.1175/1520-0434(1996)011<0560:FFFAIB>2.0.CO;2.

Fritsch, J. M., and R. Carbone, 2004: Improving quantitative precipitation forecasts in the warm season: A USWRP research and development strategy. *Bull. Amer. Meteor. Soc.*, **85**, 955–965, doi:10.1175/BAMS-85-7-955.

Germann, U., I. Zawadzki, and B. Turner, 2006: Predictability of precipitation from continental radar images. Part IV: Limits to prediction. *J. Atmos. Sci.*, **63**, 2092–2108, doi:10.1175/JAS3735.1.

Gumbel, E. J., 1960: Bivariate exponential distributions. *J. Amer. Stat. Assoc.*, **55**, 698–707, doi:10.1080/01621459.1960.10483368.

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? *Quart. J. Roy. Meteor. Soc.*, **132**, 2905–2923, doi:10.1256/qj.06.25.

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

——, 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.

Hapuarachchi, H., Q. Wang, and T. Pagano, 2011: A review of advances in flash flood forecasting. *Hydrol. Processes*, **25**, 2771–2784, doi:10.1002/hyp.8040.

Herman, G. R., 2016: Model post-processing for the extremes: Improving forecasts of locally extreme rainfall. M.S. thesis, Dept. of Atmospheric Science, Colorado State University, 212 pp. [Available online at https://dspace.library.colostate.edu/bitstream/handle/10217/173452/Herman_colostate_0053N_13419.pdf?sequence=1&isAllowed=y.]

Hershfield, D. M., 1961: Rainfall frequency atlas of the United States: For durations from 30 minutes to 24 hours and return periods from 1 to 100 years. U.S. Weather Bureau Tech. Paper 40, 61 pp. [Available online at http://www.nws.noaa.gov/oh/hdsc/PF_documents/TechnicalPaper_No40.pdf.]

Hitchens, N. M., H. E. Brooks, and R. S. Schumacher, 2013: Spatial and temporal characteristics of heavy hourly rainfall in the United States. *Mon. Wea. Rev.*, **141**, 4564–4575, doi:10.1175/MWR-D-12-00297.1.

Kimberlain, T. B., 2013: Tropical Cyclone Report: Tropical Storm Debby (AL042012) 23–27 June 2012. National Hurricane Center, 51 pp. [Available online at http://www.nhc.noaa.gov/data/tcr/AL042012_Debby.pdf.]

Lin, Y., and K. E. Mitchell, 2005: The NCEP stage II/IV hourly precipitation analyses: Development and applications. Preprints, *19th Conf. on Hydrology*, San Diego, CA, Amer. Meteor. Soc., 1.2. [Available online at https://ams.confex.com/ams/pdfpapers/83847.pdf.]

Marsh, P. T., J. S. Kain, V. Lakshmanan, A. J. Clark, N. M. Hitchens, and J. Hardy, 2012: A method for calibrating deterministic forecasts of rare events. *Wea. Forecasting*, **27**, 531–538, doi:10.1175/WAF-D-11-00074.1.

Miller, J., R. Frederick, and R. Tracey, 1973: *Precipitation-Frequency Atlas of the Western United States*. NOAA Atlas 2, Vol. 3, 43 pp.

Moore, B. J., K. M. Mahoney, E. M. Sukovich, R. Cifelli, and T. M. Hamill, 2015: Climatology and environmental characteristics of extreme precipitation events in the southeastern United States. *Mon. Wea. Rev.*, **143**, 718–741, doi:10.1175/MWR-D-14-00065.1.

NCEI, 2015: Storm Events Database. National Centers for Environmental Information. [Available online at http://www.ncdc.noaa.gov/stormevents/eventdetails.jsp?id=602738.]

Neiman, P. J., F. M. Ralph, B. J. Moore, M. Hughes, K. M. Mahoney, J. M. Cordeira, and M. D. Dettinger, 2013: The landfall and inland penetration of a flood-producing atmospheric river in Arizona. Part I: Observed synoptic-scale, orographic, and hydrometeorological characteristics. *J. Hydrometeor.*, **14**, 460–484, doi:10.1175/JHM-D-12-0101.1.

Nelson, B. R., O. P. Prat, D.-J. Seo, and E. Habib, 2016: Assessment and implications of NCEP stage IV quantitative precipitation estimates for product intercomparisons. *Wea. Forecasting*, **31**, 371–394, doi:10.1175/WAF-D-14-00112.1.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Wea. Forecasting*, **29**, 489–504, doi:10.1175/WAF-D-13-00066.1.

NWS, 2002: Floods: The awesome power. National Weather Service, 14 pp. [Available online at http://www.floodsafety.noaa.gov/resources/FloodsTheAwesomePower_NSC.pdf.]

——, 2010: Southeast United States floods, September 18–23, 2009. National Weather Service Service Assessment, 35 pp. + appendixes. [Available online at http://www.nws.noaa.gov/om/assessments/pdfs/se_floods10.pdf.]

——, 2012a: Hurricane Irene, August 21–30, 2011. National Weather Service Service Assussement, 91 pp. + appendixes. [Available online at http://www.nws.noaa.gov/om/assessments/pdfs/Irene2012.pdf.]

——, 2012b: Remnants of Tropical Storm Lee and the Susquehanna River basin flooding of September 6–10, 2011. National Weather Service Regional Service Assessment, 45 pp. + appendixes. [Available online at http://www.nws.noaa.gov/om/assessments/pdfs/LeeSusquehanna12.pdf.]

——, 2013: Hurricane/post-Tropical Cyclone Sandy October 22–29, 2012. National Weather Service Service Assessment, 45 + appendixes. [Available online at http://www.nws.noaa.gov/os/assessments/pdfs/Sandy13.pdf.]

——, 2014: Technical Implementation Notice 14-29. Mesoscale Modeling Branch, National Centers for Environmental Prediction. [Available online at http://www.nws.noaa.gov/os/notification/tin14-29namcca.htm.]

——, 2015: Summary of natural hazard statistics for 2014 in the United States. Office of Climate, Weather, and Water Services, National Weather Service, 4 pp. [Available online at http://www.nws.noaa.gov/om/hazstats/sum14.pdf.]

Ortega, K. L., T. M. Smith, K. L. Manross, K. A. Scharfenberg, A. Witt, A. C. Kolodziej, and J. J. Gourley, 2009: The Severe Hazards Analysis and Verification Experiment. *Bull. Amer. Meteor. Soc.*, **90**, 1519–1530, doi:10.1175/2009BAMS2815.1.

Parzybok, T., B. Clarke, and D. M. Hultstrand, 2011: Average recurrence interval of extreme rainfall in real-time. *Earthzine*. [Available online at http://earthzine.org/2011/04/19/average-recurrence-interval-of-extreme-rainfall-in-real-time/.]

Perica, S., and Coauthors, 2011: *Precipitation-Frequency Atlas of the United States*. NOAA Atlas 14, Vol. 6, 241 pp. [Available online at http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume6.pdf.]

——, and Coauthors, 2013: *Precipitation-Frequency Atlas of the United States*. NOAA Atlas 14, Vol. 9, 171 pp. [Available online at http://www.nws.noaa.gov/oh/hdsc/PF_documents/Atlas14_Volume9.pdf.]

Reed, S., J. Schaake, and Z. Zhang, 2007: A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *J. Hydrol.*, **337**, 402–420, doi:10.1016/j.jhydrol.2007.02.015.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.*, **136**, 78–97, doi:10.1175/2007MWR2123.1.

Rossa, A., P. Nurmi, and E. Ebert, 2008: Overview of methods for the verification of quantitative precipitation forecasts. *Precipitation: Advances in Measurement, Estimation and Prediction*, S. C. Michaelides, Ed., Springer, 419–452.

Rutz, J. J., W. J. Steenburgh, and F. M. Ralph, 2014: Climatological characteristics of atmospheric rivers and their inland penetration over the western United States. *Mon. Wea. Rev.*, **142**, 905–921, doi:10.1175/MWR-D-13-00168.1.

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

Schmidt, J. A., A. Anderson, and J. Paul, 2007: Spatially-variable, physically-derived flash flood guidance. Preprints, *21st Conf. on Hydrology*, San Antonio, TX, Amer. Meteor. Soc., 6B.2. [Available online at https://ams.confex.com/ams/87ANNUAL/techprogram/paper_120022.htm.]

Schumacher, R. S., and R. H. Johnson, 2005: Organization and environmental properties of extreme-rain-producing mesoscale convective systems. *Mon. Wea. Rev.*, **133**, 961–976, doi:10.1175/MWR2899.1.

——, and ——, 2006: Characteristics of U.S. extreme rain events during 1999–2003. *Wea. Forecasting*, **21**, 69–85, doi:10.1175/WAF900.1.

Schwartz, C. S., and Coauthors, 2009: Next-day convection-allowing WRF model guidance: A second look at 2-km versus 4-km grid spacing. *Mon. Wea. Rev.*, **137**, 3351–3372, doi:10.1175/2009MWR2924.1.

Schwitalla, T., H.-S. Bauer, V. Wulfmeyer, and G. Zängl, 2008: Systematic errors of QPF in low-mountain regions as revealed by MM5 simulations. *Meteor. Z.*, **17**, 903–919, doi:10.1127/0941-2948/2008/0338.

Smith, P. L., and Coauthors, 2005: *Flash Flood Forecasting over Complex Terrain: With an Assessment of the Sulphur Mountain NEXRAD in Southern California.* National Academies Press, 206 pp.

Stevenson, S. N., and R. S. Schumacher, 2014: A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensor precipitation analyses. *Mon. Wea. Rev.*, **142**, 3147–3162, doi:10.1175/MWR-D-13-00345.1.

Sukovich, E. M., F. M. Ralph, F. E. Barthold, D. W. Reynolds, and D. R. Novak, 2014: Extreme quantitative precipitation forecast performance at the Weather Prediction Center from 2001 to 2011. *Wea. Forecasting*, **29**, 894–911, doi:10.1175/WAF-D-13-00061.1.

Tye, M. R., and D. Cooley, 2015: A spatial model to examine rainfall extremes in Colorado's Front Range. *J. Hydrol.*, **530**, 15–23, doi:10.1016/j.jhydrol.2015.09.023.

Villarini, G., 2016: On the seasonality of flooding across the continental United States. *Adv. Water Resour.*, **87**, 80–91, doi:10.1016/j.advwatres.2015.11.009.

Weckwerth, T. M., L. J. Bennett, L. J. Miller, J. Van Baelen, P. Di Girolamo, A. M. Blyth, and T. J. Hertneky, 2014: An observational and modeling study of the processes leading to deep, moist convection in complex terrain. *Mon. Wea. Rev.*, **142**, 2687–2708, doi:10.1175/MWR-D-13-00216.1.

Zhang, J., and Coauthors, 2011: National Mosaic and Multi-Sensor QPE (NMQ) system: Description, results, and future plans. *Bull. Amer. Meteor. Soc.*, **92**, 1321–1338, doi:10.1175/2011BAMS-D-11-00047.1.