A Method for Probabilistic Flash Flood Forecasting

Jill Hardy[a,*], Jonathan J. Gourley[b], Pierre-Emmanuel Kirstetter[c], Yang Hong[d], Fanyou Kong[e], Zachary L. Flamig[f]

[a] *School of Meteorology, University of Oklahoma, 120 David L. Boren Blvd., Norman, OK, USA 73072; jill.hardy@ou.edu*
[b] *NOAA/National Severe Storms Laboratory, 120 David L. Boren Blvd., Norman, OK, USA 73072; jj.gourley@noaa.gov*
[c] *Advanced Radar Research Center, University of Oklahoma, and NOAA/National Severe Storms Laboratory, 120 David L. Boren Blvd., Norman, Oklahoma USA 73072; pierre.kirstetter@noaa.gov*
[d] *School of Civil Engineering and Environmental Science, and Advanced Radar Research Center, University of Oklahoma, 120 David L. Boren Blvd., Norman, OK, USA 73072; yanghong@ou.edu*
[e] *Center for Analysis and Prediction of Storms, University of Oklahoma, 120 David L. Boren Blvd., Norman, OK, USA 73072; fkong@ou.edu*
[f] *Cooperative Institute for Mesoscale Meteorological Studies, and Advanced Radar Research Center, University of Oklahoma, and NOAA/National Severe Storms Laboratory, 120 David L. Boren Blvd., Norman, OK, USA 73072; zac.flamig@noaa.gov*

Abstract

Flash flooding is one of the most costly and deadly natural hazards in the United States and across the globe. This study advances the use of high-resolution quantitative precipitation forecasts (QPFs) for flash flood forecasting. The QPFs are derived from a stormscale ensemble prediction system, and used within a distributed hydrological model framework to yield basin-specific, probabilistic flash flood forecasts (PFFFs). Before creating the PFFFs, it is important to characterize QPF uncertainty, particularly in terms of location which is the most problematic for hydrological use of QPFs. The SAL methodology (Wernli et al., 2008), which stands for structure, amplitude, and location, is used for this error quantification, with a focus on location. Finally, the PFFF methodology is proposed that produces probabilistic hydrological forecasts. The main advantages of this method are: 1) identifying specific basin scales that are forecast to

*Corresponding author
Email address: jill.hardy@ou.edu (Jill Hardy)

be impacted by flash flooding; 2) yielding probabilistic information about the forecast hydrologic response that accounts for the locational uncertainties of the QPFs; 3) improving lead time by using stormscale NWP ensemble forecasts; and 4) not requiring multiple simulations, which are computationally demanding.

Keywords: flash flood, probabilistic, NWP, distributed modeling

1. Introduction

According to the U.S. Natural Hazard Statistics, flooding is the number one weather-related killer over a 30-year average (National Weather Service, 2014). In particular, flash flooding can be very dangerous due to its short timescales. Generally, flash floods are defined as flooding that occurs within six hours of a causative event (Hapuarachchi et al., 2011). They tend to occur in small headwater catchments, less than a few hundred square kilometers, due in part because these basins respond quickly to excessive rainfall amounts that fall in the short time periods characterized by flash flood-producing events (Kelsch, 2001). Unfortunately, these small basins can also be located in urban areas where the effects of flash flooding on society can be substantial.

In the simplest sense, as described by Doswell III et al. (1996), "a flash flood event is the concatenation of a meteorological event with a particular hydrological situation." Meteorologically, it is crucial to properly predict not only the occurrence of a rain event, but more importantly, the intensity and movement of the rainfall to accurately depict the conditions of a flash flood event. However, the meteorological component is only half of the problem. Hydrologically, it is necessary to understand the antecedent soil conditions, land and soil characteristics, topography, and basin size to know how the rainfall will impact the basin response (Davis, 2001).

Therefore, this study focused on both sides of the problem: inputting high-resolution quantitative precipitation forecasts (QPFs), that attempt to capture the dynamics of heavy rainfall (e.g. cell motion, development, intensity, duration) into a distributed hydrological model, that will take into account the necessary hydrological factors. It should be noted that the focus of this paper will be on the meteorological component and its application in a hydrological framework.

In regards to the meteorological component, several studies have examined the accuracy of high-resolution, convection-allowing numerical weather prediction (NWP) models. Simply considering resolution, Roberts (2005) showed that higher resolution NWP models (1- or 4-km) have more reliable forecasts of flood-producing rainfall (up to 7 hours ahead) as compared to lower resolution (12- or 60-km) models. Schwartz et al. (2009) delved into the issue of convection-allowing versus convection-parameterizing models; the difference being that convection-allowing models can generate and resolve convection, while the parameterizing models represent convective processes that occur at sub-pixel resolution using a statistical approach. They found that higher resolution (2- and 4-km), convection-allowing models were more skillful at predicting amplitude and location of heavy rainfall as compared to the 12-km, convection-parameterizing model. Furthermore, Clark et al. (2009) compared a high-resolution, convection-allowing ensemble with a coarser, parameterized-convection model. They found the ensemble to produce more skillful precipitation forecasts, even with a small number of members, thus showing the promise of such ensembles.

Particular to the use of high-resolution QPFs comes the issue of displacement errors of finescale features (Ebert, 2008). These small errors can have significant effects on flash flood prediction since flash flooding is very location-dependent. The smallest offset of heavy rainfall can make the difference between an event and non-event because basins prone to flash flooding are commonly quite small

(Vincendon et al., 2011). Probabilistic forecasting offers the potential to quantify this locational uncertainty, thus it is the focus of our study.

In regards to the hydrological component, the use of hydrological models for flood forecasting has been commonplace for many years (Singh et al., 1995). However, their use for flash flood forecasting is at a relative infancy (Reed et al., 2007). More and more operational hydrological models incorporate radar-derived estimates of rainfall as their main precipitation input. These estimates can have resolutions as high as 1-km with a 2-min update cycle, and once input into the model, provide a good depiction of the present state of the hydrologic cycle. However, the radar estimates are also subject to uncertainties (Zhang et al., 2015), but more importantly, only allow for hydrological modeling once the water is already hitting the ground. The time interval between heavy rainfall observations and flash flooding can be on the order of minutes, especially for small (sometimes, urban) basins. This short lead time makes it imperative to receive information prior to radar measurements of rainfall.

Increasing the lead time for these events is necessary in order to better protect life and property (Stensrud et al., 2009; Hapuarachchi et al., 2011; Vincendon et al., 2011). The best way to do this is by improving guidance to hydrological models via inputting quantitative precipitation forecasts, derived from numerical weather prediction models, into the models (Collier, 2007). Fritsch & Carbone (2004) discussed the need to focus on warm-season QPF improvement, with one of the main purposes being the application to hydrological forecasting. They argued that a major research area needs to be determining whether QPFs are valuable to hydrological prediction, especially since hydrological predictions "are among the principal societal payoffs resulting from warm-season QPF improvement...". Our study assumes that QPFs on their own give an estimate of the relative location and intensity of future rainfall, however, giving them a hydrologic relevance is the only way they will be useful for flash flood forecasting.

In particular, the desire for ensembles of QPFs (no matter the resolution) as inputs for hydrological models is apparent in the field of flash flood forecasting (Cloke & Pappenberger, 2009). The methods thus far have been to: 1) input individual members of a QPF ensemble directly into a hydrological model to create an ensemble of hydrologic forecasts (Zappa et al., 2008; Verbunt et al., 2007), 2) perturb one deterministic QPF to create an ensemble for input into the hydrological model (Vincendon et al., 2011), or 3) perturb ensemble members and hydrologic model parameters. Our study is unique in that it creates a high-resolution deterministic representative of all ensemble members (via probability matching) for input into the hydrological model. This method cuts back the computational expense (compared to running multiple simulations), while still accounting for the optimal location defined by the ensemble mean, and the rainfall intensity represented by the entire QPF ensemble.

With such ensemble hydrologic outputs, probabilistic flash flood forecasting has been discussed in the above studies, and others (Krzysztofowicz, 2001; Drobinski et al., 2014). This study's method is novel in that it creates a final probabilistic product not from considering the fraction of hydrologic output members that exceeds a certain discharge threshold, but rather from the multiplication of meteorological and hydrological probabilistic products. In brief, the ultimate goal of this study is to derive basin-specific probabilistic flash flood forecasts (PFFFs) using an ensemble of forecast members (QPFs), combined with simulated basin responses (derived from a distributed hydrological model), in order to identify basin scales and lead times for flash flood prediction. It is noted that the proposed method deals with locational uncertainties in QPFs alone. Future methods should also consider additional errors in timing, storm structure, and amplitude. The rest of this paper is outlined as follows: Section 2 describes the two precipitation datasets and the distributed hydrological model used in this study; Section 3 explains the error quantification procedure that was done to find the biases related to the QPFs; Section 4 details the methodology conducted to create the PFFFs, and is followed by Section 5 discussing the results from the case study; and finally, Section 6 summarizes the conclusions from the study.

2. Datasets

2.1. Forecast Rainfall

This study relies on the use of a NWP model that is capable of producing stormscale QPFs. These QPFs serve as the input precipitation field for the hydrological model. As part of the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Testbed (HWT) Spring Experiment, the Center for Analysis and Prediction of Storms (CAPS) at the University of Oklahoma (OU) has developed a multi-model storm-scale ensemble forecast (SSEF) in real-time (Kong et al., 2011). Since the 2007 Spring Experiment, CAPS has been improving the SSEF each year to include such items as radar data assimilation, more members, larger domains, post-processed products, and longer forecasts.

QPFs produced during the 2010-2012 NOAA HWT Spring Experiments have a 4-km resolution, are produced hourly, and cover the entire continental U.S. (CONUS). Only ensemble members that included assimilated radar data into their initial conditions were used, since this information is useful in adjusting initial model states with the aim of improving rainfall forecasts. All members were initialized at 00Z and produced hourly QPFs up to 36 hours ahead. Table 1 shows the overall details of each year's ensemble used in this study, including the number of members, number of days the ensemble was run, and number of forecast hours. The annual CAPS Spring Experiment Program Plans (`http://hwt.nssl.noaa.gov/efp/`) provide detailed information about each member, including their initial and boundary conditions, microphysics scheme, land surface model, and planetary boundary layer (PBL) scheme. The case study used to introduce our new methodology, as described in later sections, occurs during 2010. The next paragraph will go into more depth about the ensemble members of this particular year, but the interested reader is referred to the program website provided above for even more information about the members, as well as the experimental design for 2011 and 2012.

For 2010, the CAPS ensemble had 24 members: 18 members that were produced using the Weather Research and Forecast (WRF) Advanced Research WRF core (ARW; Skamarock et al. 2005), four members produced using the WRF Nonhydrostatic Mesoscale Model (NMM; Janjic 2003), one member produced using the CAPS Advanced Regional Prediction System (ARPS; Xue et al. 2003), and one produced at the Storm Prediction Center (SPC). The WRF model cores used V3.1.1 (Kong et al., 2011), which included new microphysics and PBL schemes. All members were initialized with the NAM 12-km 00Z analyses as background. A subset of the members had initial condition perturbations, which were obtained from the National Center for Environmental Prediction (NCEP) Short-Range Ensemble Forecast (SREF; Du et al. 2006). Another subset tested physics perturbations only to assess the impact of microphysics and PBL schemes. The ARPS three-dimensional variational data assimilation (3DVAR) and Cloud Analysis package (Gao et al., 2004) was used to assimilate Doppler radar radial wind and reflectivity data from the national network of WSR-88Ds. Since the NMM has a different horizontal grid compared to both the ARW and ARPS (E-grid vs. C-grid), the NMM forecasts were converted onto the same grid as the other two. More information is available in Xue et al. (2010).

Year	**Number of Members**	**Number of Analysis Days**	**Number of Forecast Hours**
2010	24	36	30
2011	45	35	36
2012	24	35	36

Table 1: Details of the CAPS SSEF for the years 2010, 2011, and 2012.

2.2. Observed Rainfall

To complete the error analysis, the hourly QPFs were compared to hourly quantitative precipitation estimates (QPEs) of the same resolution, 4-km. The selected QPEs were the NCEP Stage IV QPEs (Lin & Mitchell, 2005), which are derived from mosaicking hourly precipitation analyses from each of the 12 River Forecast Centers (RFCs) across the CONUS (see Figure 1). These QPEs are considered multi-sensor products, meaning they are derived from both radar and gauge observations that are manually quality-controlled at the individual RFCs, then sent to NCEP for mosaicking.

2.3. The CREST Distributed Hydrological Model

The Coupled Routing and Excess STorage (CREST; Wang et al. 2011) model is a distributed hydrological model capable of simulating both spatial and temporal variations in surface and subsurface water fluxes, as well as cell-to-cell water storage. It was jointly created by the University of Oklahoma (`http://hydro.ou.edu/research/crest/`) and the National Aeronautics and Space Administration (NASA) SERVIR project. Physically-based spatially-distributed models, such as CREST, are the most capable for use in flash flood prediction because they provide streamflow estimates at any location within the basin (not just at the outlet). Additionally, CREST allows for upstream routed water to reenter the soil moisture reservoirs of downstream cells, which is what happens naturally in losing and gaining streams. The CREST model has user-defined grid cell resolution, which allows for its application at any scale from basin-specific to global. The CREST model will be one of the core hydrologic models in the Flooded Locations And Simulated Hydrographs (FLASH) project for improving flash flood prediction in the U.S. National Weather Service. The model can be driven by satellite-based products, rain gauge observations, radar-derived precipitation estimates, or NWP QPFs. The reader is pointed to Wang et al. (2011) for the detailed physical equations in the CREST model.

For this study, the CREST model was forced with the CAPS QPFs, and run without calibration. These forecasts were later compared to CREST outputs forced by the Stage IV QPEs. While this study's final product is a probabilistic one, the CREST model outputs discharge in m^3/s. Thus, discharge was converted to return period via a reanalysis process described in Section 4.2 in order to complete this study. The soil information comes from a multi-layer, CONUS-wide dataset developed at the Pennsylvania State University College of Earth and Mineral Sciences (`http://www.soilinfo.psu.edu/`). CONUS-SOIL is a soil characteristics dataset that provides geographic layers of many different hydrological parameters (e.g. porosity and available water capacity), derived from the U.S. Department of Agriculture's State Soil Geographic Database (STATSGO).

3. QPF Error Quantification

Before using the CAPS SSEF of QPFs within the hydrological framework, it is important to understand the error characteristics of the individual members in order to allow for better interpretation of the final results. Because location error is important for flash flood forecasting, using evaluation metrics that incorporate location errors is vital (Gilleland et al., 2009).

Additionally, with the use of high-resolution QPFs, the evaluation metrics must overcome the issue of displacement errors of finescale features (Ebert, 2008). One of the main obstacles of these displacement errors is the occurrence of "double penalties," or rather, when the forecast is penalized twice for appearing to have missed the observed event when it is really just offset slightly (Clark et al., 2010). With such a penalization, it becomes necessary to verify the forecasts with techniques that can quantify the displacements. Traditional metrics for evaluating forecasts (e.g. root mean square error, RMSE) do not capture the true quality associated with the forecasts (Kain et al., 2003), so new verification metrics must be used (Gilleland et al., 2009).

Wernli et al. (2008) (hereafter, SAL2008) developed a technique called SAL which stands for structure, amplitude, and location, the three characteristics evaluated for the precipitation forecasts. It is an object-based verification technique, meaning it requires criteria to define an enclosed precipitation object in order to conduct the analyses. The forecast and observed objects are compared in terms of structure (S), amplitude (A), and location (L) in order to explain the errors associated with the forecast member.

Wernli et al. (2009) (hereafter, SAL2009) proved that the SAL method is a reliable error evaluation for stormscale, convective objects by comparing it with classical error metrics (like RMSE) that are gridpoint-based, as opposed to object-based. In particular, SAL2009 study was similar to this study because it compared high-resolution, hourly QPFs. They found that the SAL technique does provide useful guidance for QPFs, which is sometimes more meaningful than the gridpoint-based measures (in particular, for intense events). Additionally, Vincendon et al. (2011) used SAL to create error statistics used in their perturbation method, and found that location perturbations of the rainfall objects had the strongest impact on the skill of the ensemble. Further, Zimmer & Wernli (2011) extended the SAL method by introducing a fuzzy approach to account for timing errors in QPFs, which can also cause "double penalties". They found that both spatial and temporal errors are accounted for when using the SAL method, and timing errors can be quantified with the fuzzy approach. This is an important finding because it further explains QPF errors characteristics, and provides more explicit verification results. Future work with SAL should consider this extension to help better explain locational uncertainties.

The SAL method is an object-based verification technique, meaning it requires criteria to define an enclosed precipitation object in order to conduct the analyses. The forecast and observed objects are compared in terms of structure (S), amplitude (A), and location (L) in order to explain the errors associated with the forecast member.

Because this technique is what Gilleland et al. (2009) defines as "features-based," the objects must be identified in both the forecast (CAPS) and observed (Stage IV) fields, as well as the domain in which they will be compared. SAL2009 advocates for smaller domain sizes for doing the verification. They argue that large domains may contain different meteorological regimes that can affect the results (e.g. the convective QPF did well, but the stratiform rainband may not). Thus, it is necessary to limit the domain to be small enough that QPF regimes are isolated.

With this in mind, the domains were limited to areas the U.S. Geological Survey (USGS) calls hydrologic units. Each unit is defined by the drainage area of a major river, or a combination of drainage areas from several major rivers, within the United States (Seaber et al., 1987). There are 21 regions within the U.S., but only 18 that make up the CONUS (see Figure 2). While these regions do not denote different meteorological regimes (as recommended by SAL2009), they are small enough to suffice the SAL requirements. But more importantly, using hydrologic unit regions (hereby, HUC regions) for this study's domains gives the analysis some hydrologic relevance by allowing for the comparison of rainfall objects that will ultimately have similar drainage paths, which is the primary goal of properly predicting the location of flash flooding.

Within each of the 18 HUC regions, only the largest, spatially continuous object for each the forecast and observed fields was compared at each hour. This step deviates from the original SAL2008 procedure, which used all objects in the domain for comparison. This decision was made because, for high-resolution hourly rainfall, there can be many discrete objects. However, many may be small, spurious objects (e.g. trailing stratiform) that are secondary to a larger, more mature object (e.g. mesoscale convective system). Thus, only comparing the largest forecast and observed object in each HUC region (i.e. up to 18 analyses per hour) should help reduce biased SAL errors. With this change, the descriptions below for each of the SAL components are based on our one-to-one object comparison, and may differ from the original SAL2008 definitions.

The structure (S) element corresponds to the normalized difference in volume between the forecast and observed objects. The idea is to compare the size and shape of the objects, and leave the amplitude element to compare the intensity of them. S is bounded by [-2, 2], with zero being a perfect score. Positive values denote that the forecast field is too large and/or too flat, whereas negative values mean the forecast object is too small and/or too peaked. For example, from an applied standpoint, large values of S may show the model predicted widespread, stratiform precipitation when there was actually a small, convective event. This information clearly explains the importance of S being based on volumes, as it gives information about the areal coverage, as well as peakedness of an event.

The amplitude (A) element signifies the normalized difference of the object-averaged precipitation values. Thus, A is meant to compare the average amount of precipitation per grid point in each the forecast and observed objects, hence it being the element that describes the overall intensity of the objects. A is bounded by [-2, 2], with zero being a perfect score. A positive value of A denotes that the forecast object overestimated the object-averaged rainfall, while a negative value means the forecast object underestimated it. $A = +2$ means that the CAPS forecast produced a precipitation object, while the Stage IV product did not (i.e. a false alarm). Similarly, for A = -2, the Stage IV observation has an object when the CAPS forecast does not (i.e. a missed event). The amplitude element is different from the structure element because S gives information on the size and shape of the object, while A tells about the overall magnitude of the object's rainfall amount.

Finally, the location (L) element corresponds to the normalized distance between the centers of mass of the modeled and observed precipitation objects. L is within [0, 1], with zero being a perfect forecast of identical centers of mass. It is noted here and in SAL2008 that one caveat of this method is that L is not sensitive to rotation about the center of mass.

3.1. SAL Results

The SAL analyses were completed for all three years of data, for every available CAPS member at every available forecast hour over the conterminous U.S. Using Table 1, the total sample size is approximately 112, 860. Some members occasionally had missing data, but no member had enough missing to bias its results. The values of S, A, and L were analyzed as a function of forecast hour. The reader is reminded that the SAL methodology creates a perfect verification score when S, A, and L all have values of zero.

Figure 3 is a time series of average SAL values at every forecast hour (f00 to f36 hours) represented in this study. To create this plot, first, the structure, amplitude, and location values were each averaged over all members per hour, for each day of each year. This gives one value of S, A, and L for every hour of the entire three-year analysis. Then, all of the days (from all three years) were averaged to give one value at every forecast hour. Therefore, there was equal weighting between all days (i.e. no dependence on the number of analysis days in a year). It should be noted that after hour 30, the results only show information from 2011 and 2012, when the CAPS lengthened its forecast period to 36 hours.

The top panel shows that S and A have a diurnal pattern, with quick spikes in positive errors within the first few forecast hours, only to slowly decrease towards zero during f03 to f12. Then, the values gradually increase until f24-f26, when they begin to decrease again. S slightly increases from f30 through f36, while A continues to diminish through the end of the forecast period. A potential explanation for this pattern is that the CAPS ensemble is initialized at 0000 UTC, which is generally when there is convection in the late afternoon hours. High-resolution NWP models have notorious difficulty capturing warm-season convective initiation (Fritsch & Carbone, 2004), so the high S and A errors at the beginning of the forecast period and 24 hours later could be linked to this NWP problem. In terms of structure and amplitude, the CAPS ensemble predicts objects that are too large and/or too flat while overestimating the object-averaged precipitation. Physically speaking, one interpretation is that the CAPS ensemble generally predicts widespread heavy precipitation when there is actually a weak convective event.

However, the bottom panel of Figure 3 (with scale modified to highlight L errors) shows the L element steadily increases throughout the entire time period, and has no diurnal dependence. This is an important result because it shows that location errors in the CAPS high-resolution QPFs increase the further you get from model initialization. Thus, it gets progressively harder for the model to properly predict the location of heavy rainfall objects. This finding justifies this study's methodology of using probabilistic forecasts by exhibiting the need for quantifying location uncertainty associated with the QPFs. This is especially true for flash-flood forecasting which heavily relies on locational accuracy.

4. PFFF Methodology

Now that we have some understanding of the errors associated with the ensemble of QPFs, they can be utilized within a hydrologic framework. As stated previously, location is the primary component for accurately forecasting flash flooding. The hydrologic characteristics (e.g. basin size, slope, soil type and saturation) and meteorological factors (e.g. rainfall intensity and duration) dictate the *potential* for flash flooding, but actually predicting the location of the rainfall dictates the *susceptibility* for flash flooding. This susceptibility is the hardest to forecast at reasonable time scales, yet is the most important question in terms of societal impacts. For the reasons mentioned in Section 1, this study works towards this challenge by creating PFFFs that identify basin scales and lead times for flash flood prediction. The process for creating these PFFFs is the novel portion of this work, and therefore, will be the focus of the rest of this paper.

4.1. The Case Study

On June 14, 2010, Oklahoma City, Oklahoma, USA experienced significant flash flooding during the morning hours. Before the rainfall arrived in the metropolitan, activity began late the night before in northwest Oklahoma. A slow-moving cold front existed from the northern Texas panhandle up through southwest Kansas, with an associated outflow boundary moving eastward. Thunderstorms initiated ahead of this cold front, and multiplied as they moved eastward along the outflow boundary. In the early morning hours of the 14th, the outflow boundary shifted toward central Oklahoma, and was met by a southwesterly moist low-level jet. This encounter further strengthened the development of thunderstorms near the boundary, dumping anomalous amounts of rainfall over Oklahoma City.

Figure 4 shows the 24-hour radar-estimated rainfall totals over the Oklahoma City metropolitan from 2000 UTC on June 13th through 2000 UTC on June 14th. The highest total was in the northern extent of the area, with a gauge-measured amount of 313 mm (National Weather Service, 2010). This event recorded the highest all-time (1880 to present) daily precipitation for Oklahoma City at 194 mm. Luckily, no lives were lost, but significant damage to property took many days to clean up.

The methodology for creating the PFFFs requires many steps that are not always linear. Because of this, a schematic was created to illustrate the process (Figure 5), and the rest of this section will be devoted to walking the reader through this schematic. As a reminder, the CAPS SSEF was initiated at 0000 UTC. Therefore, because this event affected the Oklahoma City metro mainly during the morning and early afternoon hours, all of the steps in the methodology that require hourly plots were only calculated within the hours of 1300 UTC through 1800 UTC (i.e., 0800 to 1300 local time), of which only a subset of those hours will be shown.

4.2. Calculating Probabilities of Exceedance

The first step to calculating a probability of exceedance (POE) product is to prepare a QPF input for the hydrological model. Because it is computationally expensive to use every member's QPF at every hour, an alternative is to calculate a single QPF product that adequately describes the ensemble. For this study, the probability-matched mean (PM) was chosen to represent the ensemble as a single input into the hydrological model.

As defined by Ebert (2001), the probability-matched mean has the same *spatial pattern* as the ensemble mean, and the same *frequency distribution* of rain rates as the ensemble of QPFs. The ensemble mean generally captures the location of the rain center, but "smears" the rain rates, such that it reduces the maximum and increases the minimum (as would be expected by a mean). Individual members do not always capture the right location, but they maintain the extreme rain rates that are important for predicting flash flooding. Thus, the PM uses the strengths of each component, while improving upon their weaknesses.

To calculate it, two sorted lists are made: 1) the ensemble mean values, and 2) all member values ($n = 24$, with n being the number of members in the 2010 CAPS ensemble). In order to sample both distributions entirely, a matching is done such that every nth value from the member list is matched to the ensemble list. This makes the lists the same size. Finally, the member values replace the mean values, but in the geographic locations of the mean value. Thus, the final product is a grid where the highest member value is in the location of the highest mean value, and the nth member value is in the second highest mean value, and so on.

Clark et al. (2012), Kong et al. (2011), and Xue et al. (2010) each calculated the PM for the 2010 CAPS SSEF, the same as used here. Figure 6, from Xue et al. (2010), shows that the SSEF PM outperforms its own ensemble mean, the NCEP short-range ensemble forecast (SREF) 32-km ensemble mean and PM, and the 12-km NCEP North American Model (NAM) for a 6-hour accumulation period during this event. These findings validate the use of the

4-km CAPS SSEF PM as a viable representation of the ensemble for input into the hydrological model.

Before using the PM to force the CREST hydrological model, the next step is to conduct a reanalysis of observed data in order to have a reference of historical streamflow. The purpose of this reanalysis is to create a reference from simulated data so that the rarity or severity of the forecasts can be evaluated at grid points where there are no stream gauge measurements. This creates a longer-term description of the considered basin that better explains the nature of its streamflow. To complete the reanalysis, the CREST model was run using a priori parameters from 2002 through 2012 (for 11 total years of data). This time period was chosen because it was the longest period of record for the Stage IV product. Hourly NCEP Stage IV QPEs served as the precipitation input, and three-hourly potential evapotranspiration data came from the NCEP North American Regional Reanalysis (NARR) project. The NARR project uses the high-resolution NCEP Eta model and its three-dimensional data assimilation (3DVAR) technique to assimilate observations of many atmospheric variables, including potential evapotranspiration, into the analyses (Mesinger et al., 2006).

Because this process requires a long period of data, and thus, a lot of computing time, the reanalysis was only completed over the Deep Fork river basin. In particular, USGS gauge 07242380 was modeled as the basin outlet. This basin was known to have experienced flooding during this case, so it was a good choice for the analyses. The reanalysis saved a maximum discharge value for every grid cell in the basin for every year. These stored discharge values are then used as the annual peaks for computing the Log-Pearson III flood frequency relationship that converts between discharge (in cms) and return period (in years). Thus, the final product of the reanalysis was a probability density function (PDF) of historically simulated streamflow at each grid cell within the basin that can then be used as a reference to estimate the yearly recurrence of streamflow of specific magnitudes at each grid cell in the basin.

After return periods were established, the PM was input into the CREST model at each hour during the event (1300 through 1800 UTC). The output was simulated streamflow (in cms) that was converted to estimated return period (in years) at each grid cell, based on the reanalysis. The estimated return period of the grid cell was then compared to the cell's flow accumulation (i.e. catchment area or basin scale), and plotted. Figure 7 shows these plots for 1400 and 1700 UTC during the event, with each point representing a grid cell within the basin that registered above a one-year return period during that hour. The axes were converted from the log values to the true values of each variable so to better interpret the data.

Because the true return period at each cell within the basin is unknown, the distribution of the return period as a function of the flow accumulation is considered. It is expected that return period will depend on catchment size at each grid cell, since the amount of water a basin can convey is strongly dependent on its size. As can be seen in Figure 7, the data sample alone may not be sufficient to describe the distribution of return period for all flow accumulation values. For example, at 1400 UTC in Figure 7, there are only a few points that provide information for basin scales greater than 55 km^2. Thus, the distributions of return period conditioned on the basin scale were modeled with the Generalized Additive Models for Location, Scale, and Shape (GAMLSS; Stasinopoulos & Rigby 2007). GAMLSS is a semi-parametric approach which aims at modeling a response variable's distribution (here, the return period). Two main assumptions are made: (i) the response variable is a random variable following a known parametric distribution with density conditional on parameters, μ and σ, and (ii) the response variables are mutually independent. Each parameter is modeled as a function of the flow accumulation (i.e. the explanatory variable) using smooth link functions (i.e. locally weighted scatterplot smoothing, LOESS).

For the sake of simplicity, several two-parameter density functions (with the first two moments: location μ and scale σ) were tested for this basin, and the Gamma distribution best fit the data. Moving across different basin scales, conditional distributions of return period were interpolated across scales not well represented in this particular basin. For example, if a basin has no interior grid cells with a catchment size of 10 km^2, then the semi-parametric model would interpolate data from the surrounding catchment sizes in order to provide information for a 10-km^2 size. The final result (as seen in Figure 7) is a smooth quantile map of return periods for all scales. Here, the conditional median is yellow, meaning that 50% of catchments are expected to exceed a given return period. For instance, for 1700 UTC in Figure 7, at a 20-km^2 catchment size, 50% of catchments of that size would be expected to exceed a 7-year return period. Using the quantiles, instead of the individual points of these plots, to explain the return period patterns is better because the method is not limited to small sample sizes or lack of data at certain flow accumulations.

These plots also show that, at the start of the period (1400 UTC), return periods are large in cells with small flow accumulations (i.e. small catchments). This is indicative of flash flood prone cells, as they are more affected by rapid onsets of heavy rainfall and concomitant flash flooding over short amounts of time. At 1400 UTC, there are very few cells with large catchment areas that have high return periods. However, at 1700 UTC, the quantile peaks shift to the right, demonstrating that water is flowing downstream into the larger catchments. Return periods remain high for the small catchments because this was an event where the storms generated in a “train effect” and continued to impact the study region at small basin scales. Even with the increased quantile values in the medium-sized catchments, the largest catchments still had relatively low values throughout the period. These large cells are more capable of handling large amounts of rainfall, thus, taking far more water over longer durations to flood. This ”wave” effect with time was expected in order to make sure the CREST model was properly simulating the progression of such a flash flood event.

Using the flow accumulation versus PM return period plots, the final step is to calculate probabilities of exceedance at various return periods. Getting these POEs is the last step in the hydrologically-relevant component of the PFFF, as denoted by the right side of the schematic in Figure 5. The POEs are based on the quantiles described above. For a given return period, the *quantiles* that exceed the return period determine the POE. Remember, these quantiles were interpolated across all catchment sizes; therefore, the POE for a particular return period (at a particular time) changes as a function of catchment size.

For example, consider the probability of exceeding a 20-year return period at 1700 UTC in Figure 7. We have annotated a red line across the plot at that return period. At a 1-km^2 catchment size, all quantiles above 0.4% exceed the return period. Therefore, the POE is equal to 99.6%, at this basin scale. Moving across the 20-year red line to the 20-km^2 catchment size, the 75th percentile is located here. Thus, the POE equals 25%. Finally, if you continue to follow the 20-year return period out to a 148-km^2 catchment, the line is between the 98th and 99.6th percentiles. Even though there are no grid cells that exceed this return period at this size, the GAMLSS model fit the distribution to account for the small sample size of the basin. Therefore, the POE(RP $\geq$ 20 yrs) $\approx$ 1% at this flow accumulation (since it is between the two quantiles). Using the method of following a particular return period horizontally across all basin scales, it is possible to know the POE for every grid cell in the basin. This study considered the 5-year and 50-year return periods for comparison, and calculated their POEs for each forecast hour of the event.

4.3. Creating a Probabilistic QPF Field

The meteorological component of calculating the PFFFs is the probabilistic QPF (PQPF) field, as shown by the left side of Figure 5. This step can be done in parallel, since it does not directly rely on anything but the original CAPS QPFs. The purpose of incorporating a PQPF field, in addition to the POE field, is to account for the spatial uncertainty related to the ensemble of QPFs. Otherwise, the process would rely solely on the PM field, whose location is entirely dependent on that of the ensemble mean. Even though the ensemble mean is a good deterministic predictor of the rain center (as previously described), a PQPF field adds value to the final product by allowing for error in the spatial extent, in terms of providing a weighted probability map.

To calculate a PQPF field, the first step is to create a cumulative precipitation map for each member over the entire forecast period. Because this case study was in 2010, the CAPS ensemble had QPFs up to 30 hours in advance. Next, a rainfall threshold is determined, where a grid cell receives a 1 for exceeding, or a 0 for not exceeding. This threshold ideally highlights the event being considered, without being too focused or too broad. The subjective choice for this study was a threshold of greater than or equal to 100 mm in 30 hours. Further work should be conducted to automate this threshold selection, such as considering the average recurrence interval of forecast rainfall (e.g., 25-yr rainfall event).

These member binary grids are then used to create a neighborhood probability map of exceeding the threshold. At each grid cell, a 40-km radius is sampled (as done by Schwartz et al. (2010)), and if any surrounding cells have a one (i.e. exceed the threshold), then that cell is given a 1. This introduces some spatial uncertainty by allowing neighboring cells to affect the value of the given cell. The schematic in Figure 8 can be used to understand the process of creating a neighborhood binary grid for one member. Once this neighborhood binary grid is calculated for every member, the member grids are summed together, and then divided by the total number of members. Thus, the neighborhood probability map can be explained as the confidence the members have in exceeding the threshold within 40-km of a point. For instance, a cell with 100% denotes every member had a cell within 40-km exceeding the threshold.

To further account for the spatial uncertainty, a 2-D Gaussian smoother is applied to the neighborhood probability map, similar to the "practically perfect" forecast created in Hitchens et al. (2013). The smoother, which is applied for every grid cell in the domain, takes into consideration the surrounding grid cells' values when calculating that cell's final value. This is done such that the weights of the surrounding grid cells (within a radius of influence, here 40-km) drop off with increasing distance from the considered cell. Kong et al. (2011) completed this neighborhood smoothing with the same 2010 CAPS dataset, and found that the probabilistic skill scores increased compared to no smoothing.

Once this process has been completed throughout the entire domain, the final product is a Gaussian-smoothed, 40-km neighborhood PQPF map of exceeding the threshold of 100mm/30hr. Figure 9 shows this PQPF field for the Oklahoma City event. Notice how the map focuses the threat ($\geq$ 80%) over Oklahoma, extending into eastern Kansas and western Missouri.

4.4. Creating the Probabilistic Flash Flood Forecast

The final step in the process to create a PFFF is to multiply the hydrologic model output (i.e. the POEs) by the meteorological ensemble information (i.e. PQPFs). To do this, first, the fields need to be at the same grid size. Using the GAMLSS interpolation of return periods across basin scales (see Section 4.2), POEs were calculated based on their catchment area over the CONUS-wide flow accumulation grid. This is a fundamental step in our method's process because it creates return period estimates across all basin scales. With the two maps on the same grid size, they can be multiplied, and the final product is a CONUS-wide PFFF. Since the POE map is dependent on a chosen return period, so is the PFFF. Thus, each PFFF can be described as a map of basin-scale susceptibility for the given return period (e.g. "probabilistic flash flood forecast for exceeding a five-year return period").

5. Results

Figure 10 shows the PFFF for exceeding a five-year return period at 1400 and 1700 UTC during the June 14, 2010 Oklahoma City flash flood event. To properly interpret these maps, the reader should focus on the stream and river pixels. When the POE field is multiplied by the PQPF field, the overland grid cells take on the PQPF values entirely, due to being multiplied by approximately one. Hence, when looking at these overland areas, the high probabilities that existed in the PQPF map are clearly outlined in the PFFF.

Beginning at 1400 UTC (Figure 10, left), probabilities increase significantly over the Oklahoma City metro, with some of the very small headwater basins having probabilities near 100% within the main threat area (as outlined by the overland cells). As time progresses, the downstream basins start having the higher (warmer-colored) probabilities. The smaller, upstream basins maintain high values, as well, because this was a training echo event with high rain rates for multiple hours. This example, using the five-year return period, displays the value of this product. It provides an understandable quantity (via probabilities) that explains the threat of flash flooding across scales, using high resolution information which is necessary to capture flash-flood producing rainfall.

It is interesting to compare the PFFFs for a 5-year and a 50-year return period (Figure 11), to further understand how the threat changes based on the likelihood of the event. For both return periods, the PQPF-influenced overland grid cells highlight the threat area in the background overland cells. Overall, the 50-year return period plot (Figure 11, right) does not have as high PFFFs as the 5-year plot (Figure 11, left). Another interesting result of this comparison is that the 5-year plot appears to be smoother than the 50-year plot, even though both have the same resolution. This is caused simply by the difference in threshold. Since the 5-year return period is a lower threshold, the pixels in this plot are more susceptible to exceeding it. This leads to high PFFF values in the small, upstream basins, which tend to blend in with the background pixels.

5.1. Comparing PFFF Forcings

To test the validity of the PFFF results, the whole process was completed using Stage IV QPEs, rather than the CAPS ensemble of QPFs. Since these QPEs are quality-controlled estimates of gridded rainfall observations, using them to force the hydrologic model and define the PQPF field should provide a reasonable comparison with the QPF-driven PFFF product. However, the PFFF methodology was created to be used with an ensemble, which requires multiple rainfall fields. Thus, the methodology was altered slightly to fit the use of a deterministic Stage IV field at each hour.

First, instead of forcing the hydrologic model with the PM, it was forced simply with the Stage IV product at each hour. From there, the process is the same to get the POE fields. However, a little more has to be done to emulate the PQPF field since this field is based upon probabilities derived from averaging individual members. To begin, similar to the CAPS members, the Stage IV grids are summed over the 30-hour period. A binary grid is created where a 1 is given to a cell that exceeds the 100mm/30hr threshold. The 40-

km neighborhood binary map is made, thus introducing some artificial spatial uncertainty to the Stage IV product. At this point, since there is only one field, no neighborhood probability map is created. Instead, the Gaussian smoother is applied directly to the neighborhood binary map to scale it. Here, a point where all of the surrounding cells within a 40-km radius exceed the threshold, will still receive a 100%. The final grid is a "probabilistic" QPE (PQPE) field that is then multiplied by the POE field to create the PFFF map.

Figure 12 shows a side-by-side comparison of the QPE-forced (left) and QPF-forced (right) PFFFs for a 5-year return period at 1700 UTC. The reader is reminded to focus on the stream and river pixels, not the background overland pixels. Immediately, it is noticeable that the Stage-IV-forced PFFFs are much more focused, while the CAPS-PM-forced PFFFs show a broader stream and river threat. This is to be expected since the CAPS QPFs contain more spatial uncertainty, while the Stage IV QPEs show that the actual event was much smaller in spatial scale than was predicted. The display of this information is further confounded by a possible nuance related to the GAMLSS model's ability to fit a distribution to the return period data. For the QPE-forced return periods, it is possible that the Gamma distribution did a better job fitting to the data; whereas, the QPF values were overestimated. In this sense, the ideal outcome would be a CAPS PFFF map that would have a better fit to the chosen distribution, and thus have a similar appearance to the Stage IV map (i.e. muted background pixels). Future work should involve looking into other statistical distributions with which to fit the data.

The hourly Stage IV PFFFs have a similar temporal trend as the CAPS PFFFs (not shown). With time, the flash-flooding threat shifts from the smaller basins to the larger basins, with the Stage IV PFFFs ultimately highlighting one major river basin with high probabilities (Figure 12, left). This trend was seen in the CAPS PFFF maps, just over a broader area. An interesting difference in the hourly analyses is that, overall, the Stage IV-forced simulations targeted larger-scale basins compared to the CAPS QPFs. This, too, could be explained by the GAMLSS model's distribution choice, but further work should be done to understand this result.

Finally, areas that were the most threatened in the Stage IV analysis were encompassed by the main threat area of the CAPS maps. This is a promising result of the methodology, because it shows that a forecast up to 18 hours in advance of an event can predict the area that it will occur, with more hydrological relevance than just using QPF analyses. However, a downfall of the current method is that this QPF-forced high-risk area is rather large and overestimated, and may cause high false alarm rates. As described in the previous paragraph, working with the statistical model to better depict this threat is of utmost importance for future work.

6. Conclusions

This study offers a method for creating probabilistic flash flood forecasts (PFFFs) using an ensemble of high-resolution quantitative precipitation forecasts (QPFs). Firstly, the errors of the individual CAPS members were quantified in order to understand the biases associated with the ensemble before applying them to a hydrological framework. The SAL technique (Wernli et al., 2008) was chosen for this analysis because it builds on traditional metrics by providing information about the structure, amplitude, and, particularly, the location errors of forecast objects. Location errors are of utmost importance when forecasting flash floods, because the smallest offset in heavy rainfall can define whether the event is captured.

Results from the SAL analysis found that structure and amplitude errors had a distinct diurnal cycle, with positive errors in both components occurring during typical hours of afternoon and evening convection. This is indicative of widespread rainfall being forecast, but smaller-scale, isolated convective events occur. These errors dropped overnight and into the morning, when NWP models are traditionally better at forecasting the behavior of the atmosphere. Future work in the NWP community should focus on reducing these errors in structure and amplitude, while future work in the hydrological community should incorporate them into other forecast modeling approaches. On the other hand, the location errors of the QPF members steadily increased with forecast hour. This result is important for flash-flood forecasting because it shows that the model has difficulty pinpointing the location of heavy rainfall as time moves farther from model initialization. With a need for better lead times of flash floods, improving how to quantify this spatial uncertainty in a hydrologically- and meteorologically-driven product was the driving force behind our methodology.

The rest of the paper discussed this new methodology, and how the PFFFs account for locational uncertainties. The probabilities of exceedance (i.e., the hydrological component) were calculated by first inputting the probability-matched mean into the CREST hydrological model. The simulated streamflow was then converted to return periods using information from the model reanalysis, and plotted against basin size. Finally, statistical modeling was performed to interpolate probabilities of exceeding various return periods across all basin scales. Next, the probabilistic QPF field (i.e., the meteorological component) was calculated to account for the spatial uncertainty of the QPFs. For each QPF member, a binary field was created to determine if a particular rainfall threshold was exceeded over the forecast period. A 40-km radius of influence was then applied to each field to introduce spatial uncertainty. The final PQPF field became a smoothed map of the fraction of members that exceeded the threshold within a 40-km radius. The POE and PQPF were then multiplied to create the final PFFF, which is dependent on the return period selected for the POE.

If just the probabilistic QPF map was used to define a risk area, the contours would be too smooth to provide any valuable flash flood information. More importantly, the QPFs must be given a hydrological context in order to get susceptible scales. On the other hand, if just the probabilities of exceedance were used, it would provide detailed basin information; however, it would not focus the threat due to its interpolation across all basin sizes. The scaling provided by multiplying it with the PQPF field creates higher PFFFs where the QPFs are most confident of heavy rainfall occurrence. Results showed that the final probabilistic output compared positively with its QPE counterpart. The right location was highlighted by the PFFFs up to 18 hours in advance of the event, which is the crucial factor for completing the analyses this way. However, this location was quite broad, which could lead to high false alarm rates. Caveats of this method include having several steps that require subjective choices, as well as needing a better fit to a chosen statistical distribution. Furthermore, the developed method only deals with locational uncertainties in QPFs. Future studies should also address errors in timing, storm structure, and amplitude. Operational considerations need to first include sensitivity tests of the QPF thresholds related to intensity and spatial scale. Presently, the method requires the selection of a candidate basin within the high QPF region, which may prove to be difficult to automate. As such, the method represents an approach that deals with locational uncertainties in QPF inputs that will motivate future developments in their hydrological applications.

Probabilistic forecasts have been growing in popularity over the last several years because of their ability to quantify uncertainty information. Additionally, use of QPS ensembles has increased because of the computational advances that allow for high-enough resolution to begin to describe storm-scale, convective events without needing parameterization of deep convection. Finally, distributed hydrological models have progressed to be more realistic representations of streamflow at smaller, flash-flood scales. Combined, this methodology is at the forefront of utilizing the most advanced capabilities in all of these fields, with promising implications to the field of flash flood forecasting. However, future work should build upon these techniques to further prove operational capabilities.

Acknowledgments

The authors acknowledge the financial support of the National Science Foundation (NSF) Graduate Research Fellowship under Grant no. DGE-1102691. Partial funding was also provided by NOAA/Office of Oceanic and Atmospheric Research under the NOAA-University of Oklahoma Cooperative Agreement #NA11OAR4320072, U.S. Department of Commerce. Thank you to the University of Oklahoma (OU) and NOAA/National Severe Storms Laboratory (NOAA/NSSL) for providing the facilities, technical, and data support necessary to complete this work. We also thank OU for providing the CAPS ensemble dataset, the National Centers for Environmental Prediction for the Stage-IV observational precipitation dataset, and Pennsylvania State University for the soil dataset. Finally, thank you to the anonymous reviewers for their insightful comments, as well as the editor, for helping to improve this manuscript.

References

Clark, A. J., Gallus Jr, W. A., & Weisman, M. L. (2010). Neighborhood-based verification of precipitation forecasts from convection-allowing ncar wrf model simulations and the operational nam. *Weather and Forecasting*, **25**, 1495–1509.

Clark, A. J., Gallus Jr, W. A., Xue, M., & Kong, F. (2009). A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Weather and Forecasting*, **24**, 1121–1140.

Clark, A. J., Weiss, S. J., Kain, J. S., Jirak, I. L., Coniglio, M., Melick, C. J., Siewert, C., Sobash, R. A., Marsh, P. T., Dean, A. R. et al. (2012). An overview of the 2010 hazardous weather testbed experimental forecast program spring experiment. *Bulletin of the American Meteorological Society*, **93**.

Cloke, H., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, **375**, 613–626.

Collier, C. (2007). Flash flood forecasting: What are the limits of predictability? *Quarterly Journal of the Royal Meteorological Society*, **133**, 3–23.

Davis, R. S. (2001). Flash flood forecast and detection methods. *Meteorological Monographs*, **28**, 481–526.

Doswell III, C. A., Brooks, H. E., & Maddox, R. A. (1996). Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting*, **11**, 560–581.

Drobinski, P., Ducrocq, V., Alpert, P., Anagnostou, E., Béranger, K., Borga, M., Braud, I., Chanzy, A., Davolio, S., Delrieu, G. et al. (2014). Hymex: a

10-year multidisciplinary program on the mediterranean water cycle. *Bulletin of the American Meteorological Society*, **95**, 1063–1082.

Du, J., McQueen, J., DiMego, G., Toth, Z., Jovic, D., Zhou, B., & Chuang, H. (2006). New dimension of ncep short-range ensemble forecasting (sref) system: Inclusion of wrf members. In *Preprint, WMO Expert Team Meeting on Ensemble Prediction System.*

Ebert, E. E. (2001). Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Monthly Weather Review*, **129**, 2461–2480.

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, **15**, 51–64.

Fritsch, J. M., & Carbone, R. (2004). Improving quantitative precipitation forecasts in the warm season: A uswrp research and development strategy. *Bulletin of the American Meteorological Society*, **85**, 955–965.

Gao, J., Xue, M., Brewster, K., & Droegemeier, K. K. (2004). A three-dimensional variational data analysis method with recursive filter for doppler radars. *Journal of Atmospheric and Oceanic Technology*, **21**, 457–469.

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., & Ebert, E. E. (2009). Intercomparison of spatial forecast verification methods. *Weather and Forecasting*, **24**, 1416–1430.

Hapuarachchi, H., Wang, Q., & Pagano, T. (2011). A review of advances in flash flood forecasting. *Hydrological Processes*, **25**, 2771–2784.

Hitchens, N. M., Brooks, H. E., & Kay, M. P. (2013). Objective limits on forecasting skill of rare events. *Weather and Forecasting*, **28**, 525–534.

Janjic, Z. (2003). A nonhydrostatic model based on a new approach. *Meteorology and Atmospheric Physics*, **82**, 271–285.

Kain, J. S., Baldwin, M. E., Janish, P. R., Weiss, S. J., Kay, M. P., & Carbin, G. W. (2003). Subjective verification of numerical models as a component of a broader interaction between research and operations. *Weather and Forecasting*, **18**, 847–860.

Kelsch, M. (2001). Hydrometeorological characteristics of flash floods. In *Coping with Flash Floods* (pp. 181–193). Springer.

Kong, F., Xue, M., Thomas, K. W., Wang, Y., Brewster, K., Wang, X., Gao, J., Weiss, S. J., Clark, A., Kain, J. et al. (2011). Evaluation of caps multimodel storm-scale ensemble forecast for the noaa hwt 2010 spring experiment. 24th Conference on Weather and Forecasting/20th Conference on Numerical Weather Prediction. AMS.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, **249**, 2–9.

Lin, Y., & Mitchell, K. E. (2005). 1.2 the ncep stage ii/iv hourly precipitation analyses: Development and applications. In *19th Conf. Hydrology, American Meteorological Society, San Diego, CA, USA*. Citeseer.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P. C., Ebisuzaki, W., Jovic, D., Woollen, J., Rogers, E., Berbery, E. H. et al. (2006). North american regional reanalysis. *Bulletin of the American Meteorological Society*, **87**, 343–360.

National Weather Service (2010). Record setting rainfall and significant flooding over oklahoma. URL: `http://www.srh.noaa.gov/oun/?n=events-20100614`.

National Weather Service (2014). Weather fatalities. URL: `http://www.nws.noaa.gov/om/hazstats.shtml`.

Reed, S., Schaake, J., & Zhang, Z. (2007). A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. *Journal of Hydrology*, **337**, 402–420.

Roberts, N. (2005). *An investigation of the ability of a storm scale configuration of the Met Office NWP model to predict flood-producing rainfall.*. Technical Report 455 Met Office.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., & Coniglio, M. C. (2009). Next-day convection-allowing wrf model guidance: A second look at 2-km versus 4-km grid spacing. *Monthly Weather Review*, **137**, 3351–3372.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., Thomas, K. W., Levit, J. J., Coniglio, M. C., & Wandishin, M. S. (2010). Toward improved convection-allowing ensembles: Model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather and Forecasting*, **25**, 263–280.

Seaber, P. R., Kapinos, F. P., & Knapp, G. L. (1987). *Hydrologic unit maps*. US Government Printing Office.

Singh, V. P. et al. (1995). *Computer models of watershed hydrology.*. Water Resources Publications.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M., Wang, W., & Powers, J. G. (2005). *A description of the advanced research WRF version 2*. Technical Report DTIC Document.

Stasinopoulos, D. M., & Rigby, R. A. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*, **23**, 1–46.

Stensrud, D. J., Wicker, L. J., Kelleher, K. E., Xue, M., Foster, M. P., Schaefer, J. T., Schneider, R. S., Benjamin, S. G., Weygandt, S. S., Ferree, J. T. et al. (2009). Convective-scale warn-on-forecast system: A vision for 2020. *Bulletin of the American Meteorological Society*, **90**, 1487–1499.

Verbunt, M., Walser, A., Gurtz, J., Montani, A., & Schär, C. (2007). Probabilistic flood forecasting with a limited-area ensemble prediction system: Selected case studies. *Journal of Hydrometeorology*, **8**, 897–909.

Vincendon, B., Ducrocq, V., Nuissier, O., & Vié, B. (2011). Perturbation of convection-permitting nwp forecasts for flash-flood ensemble forecasting. *Natural Hazards and Earth System Science*, **11**, 1529–1544.

Wang, J., Hong, Y., Li, L., Gourley, J. J., Khan, S. I., Yilmaz, K. K., Adler, R. F., Policelli, F. S., Habib, S., Irwn, D. et al. (2011). The coupled routing and excess storage (crest) distributed hydrological model. *Hydrological Sciences Journal*, **56**, 84–98.

Wernli, H., Hofmann, C., & Zimmer, M. (2009). Spatial forecast verification methods intercomparison project: Application of the sal technique. *Weather and Forecasting*, **24**, 1472–1484.

Wernli, H., Paulat, M., Hagen, M., & Frei, C. (2008). Sal-a novel quality measure for the verification of quantitative precipitation forecasts. *Monthly Weather Review*, **136**, 4470–4487.

Xue, M., Kong, F., Thomas, K., Wang, Y., Brewster, K., Gao, J., Wang, X., Weiss, S., Clark, A., Kain, J. et al. (2010). Caps realtime storm scale ensemble and high resolution forecasts for the noaa hazardous weather testbed 2010 spring experiment. In *25th Conf. Severe Local Storms*.

Xue, M., Wang, D., Gao, J., Brewster, K., & Droegemeier, K. K. (2003). The advanced regional prediction system (arps), storm-scale numerical weather prediction and data assimilation. *Meteorology and Atmospheric Physics*, **82**, 139–170.

Zappa, M., Rotach, M. W., Arpagaus, M., Dorninger, M., Hegg, C., Montani, A., Ranzi, R., Ament, F., Germann, U., Grossi, G. et al. (2008). Map d-phase: Real-time demonstration of hydrological ensemble prediction systems. *Atmospheric Science Letters*, **9**, 80–87.

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cocks, S., Martinaitis, S. et al. (2015). Multi-radar multi-sensor (mrms) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, .

Zimmer, M., & Wernli, H. (2011). Verification of quantitative precipitation forecasts on short time-scales: A fuzzy approach to handle timing errors with sal. *Meteorologische Zeitschrift*, **20**, 95–105.

List of Figures

Figure 7: **Flow accumulation (km^2) versus return period (years) for Oklahoma City event.** Flow accumulation (km^2) versus return period (years) from 14Z (left) and 17Z (right) on June 14, 2010. The relationship between streamflow and estimated return period was established by a reanalysis of long-term CREST simulations with Stage IV forcing. Each blue cross represents a grid cell within the modeled basin, with a particular flow accumulation and estimated return period response (converted from streamflow) from forcing the CREST model with the CAPS SSEF PM. Modeled conditional quantiles of return periods are denoted by the colored lines, interpreted as probabilities of exceedance (POE) at certain return periods. The red line annotated on the 17Z plot is an example of how to interpret the POE for a 20-year return period across all basin scales. This is done by looking at which quantile lines have exceeded the red line at each flow accumulation.

Figure 8: **Schematic of how to create the neighborhood binary grid for one member.** Schematic of how to create the neighborhood binary grid for one member, which is then summed with other member's grids to create the final neighborhood probability map.

Figure 9: **Probabilistic QPF field for the Oklahoma City event.** Gaussian-weighted, 40-km neighborhood probabilistic QPF field for exceeding the threshold of 100mm/30hr during the 30-hour period beginning at 00Z on June 14, 2010.

Figure 10: **Probabilistic Flash Flood Forecasts for exceeding a return period of 5 years during the Oklahoma City event.** Probabilistic Flash Flood Forecasts for exceeding a return period of 5 years, for the June 14, 2010 Oklahoma City event at (left) 14Z and (right) 17Z.

Figure 11: **PFFF comparison for exceeding 5 years and 50 years** Probabilistic Flash Flood Forecasts for exceeding a return period of: (left) 5 years, and (right) 50 years, at 15Z on June 14, 2010. The images are zoomed in, focusing on Oklahoma and Cleveland Counties.

Figure 12: **PFFF comparison for the PFFF forced using the Stage IV QPE and forced using the CAPS ensemble.** Probabilistic Flash Flood Forecasts for exceeding a return period of 5 years, at 17Z during the June 14, 2010 Oklahoma City event. (left) PFFF forced using the Stage IV QPE, and (right) PFFF forced using the CAPS ensemble.

Figure 1

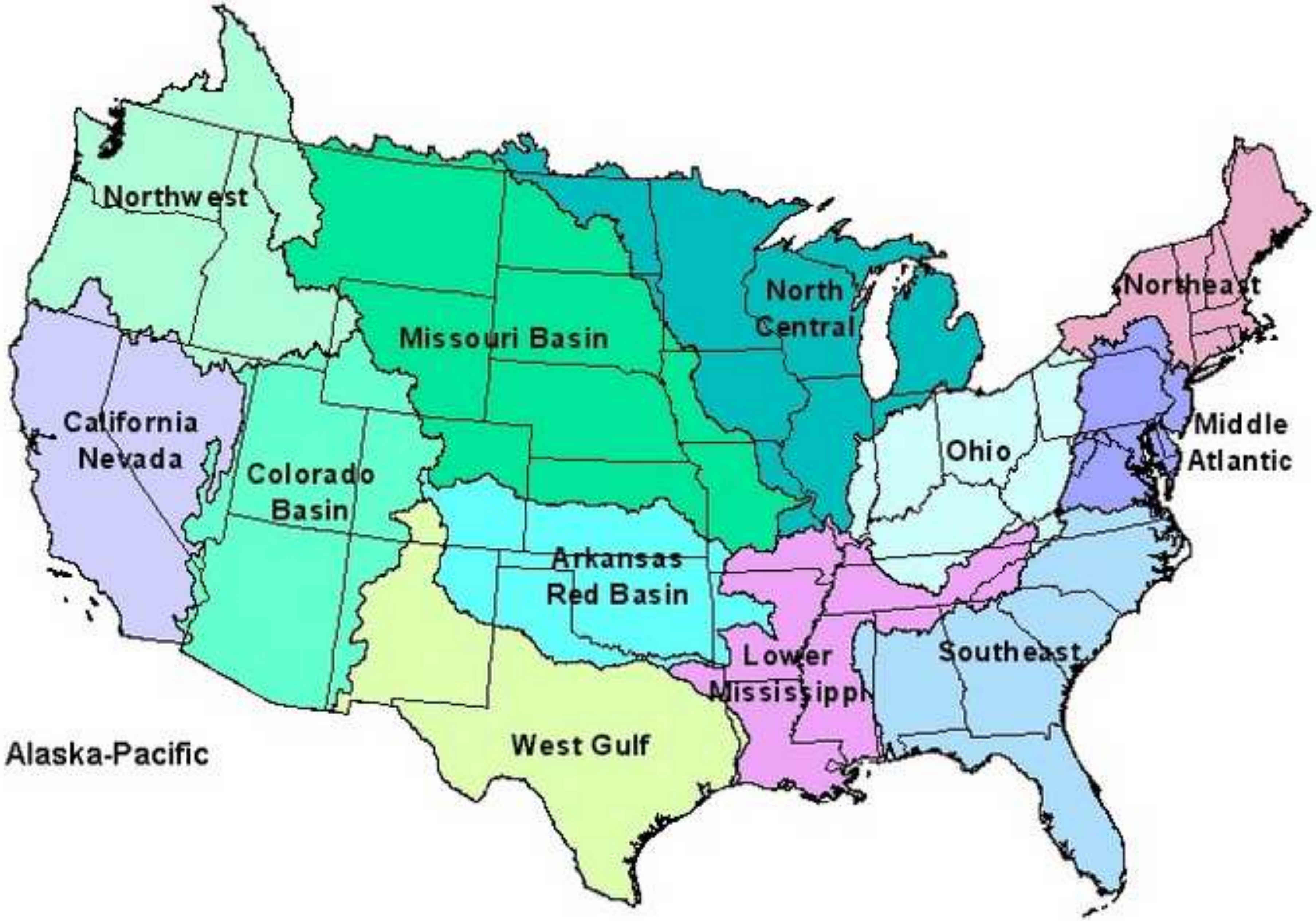

Northwest
Missouri Basin
North Central
Northeast
California Nevada
Colorado Basin
Ohio
Middle Atlantic
Arkansas Red Basin
Lower Mississippi
Southeast
West Gulf
Alaska-Pacific

Figure 2

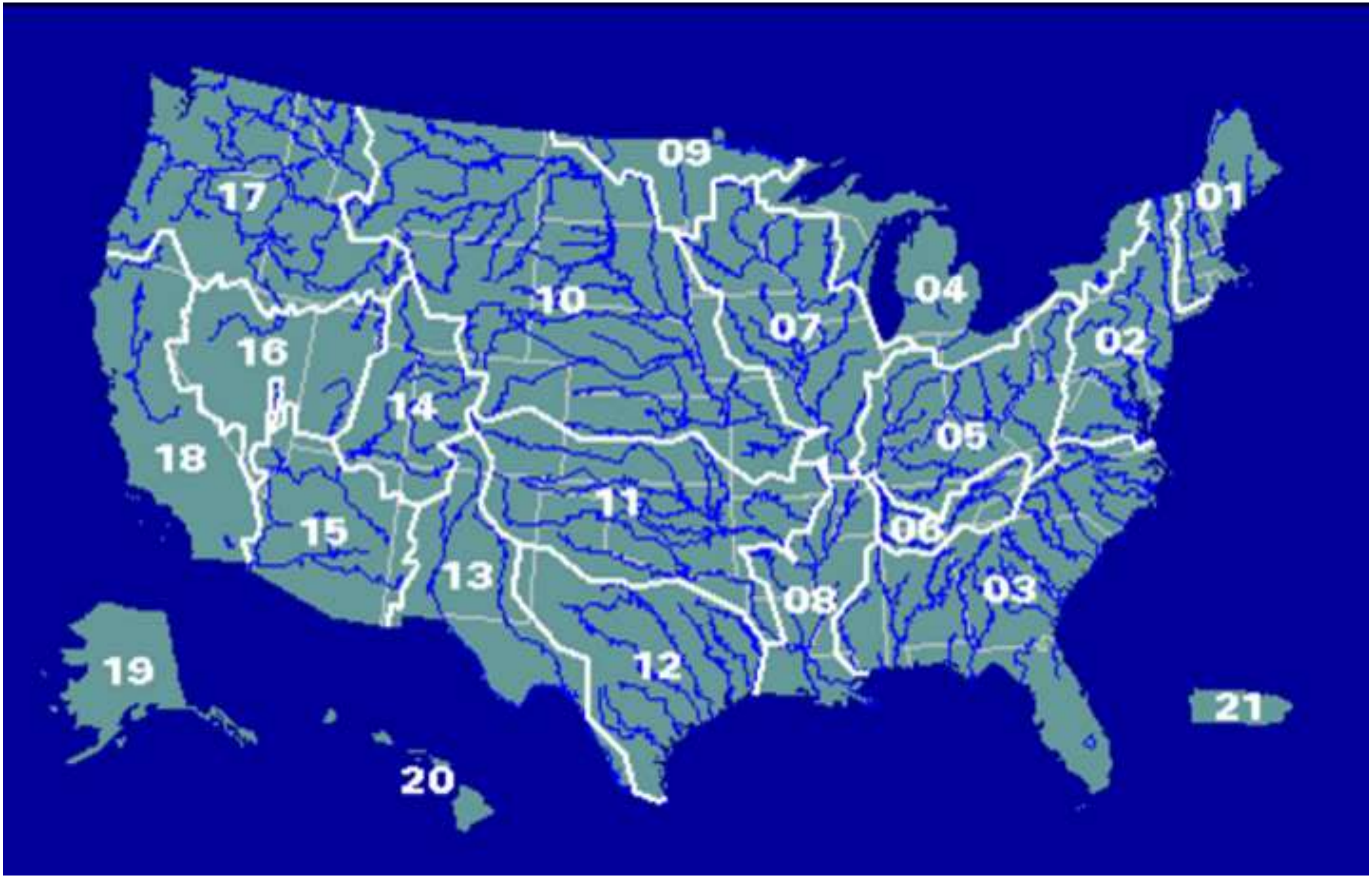
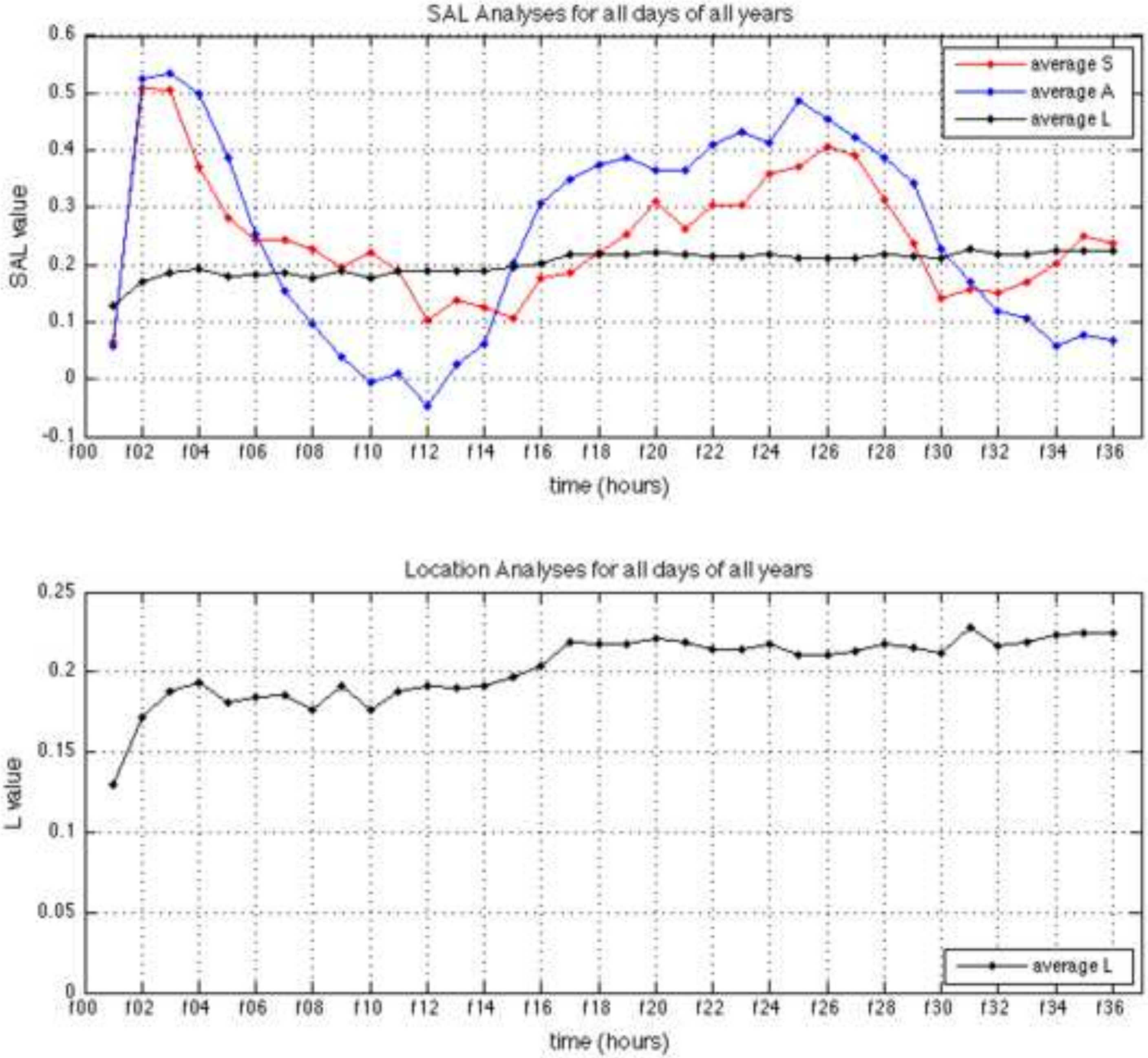


Figure 3



SAL Analyses for all days of all years

Location Analyses for all days of all years

Figure 4

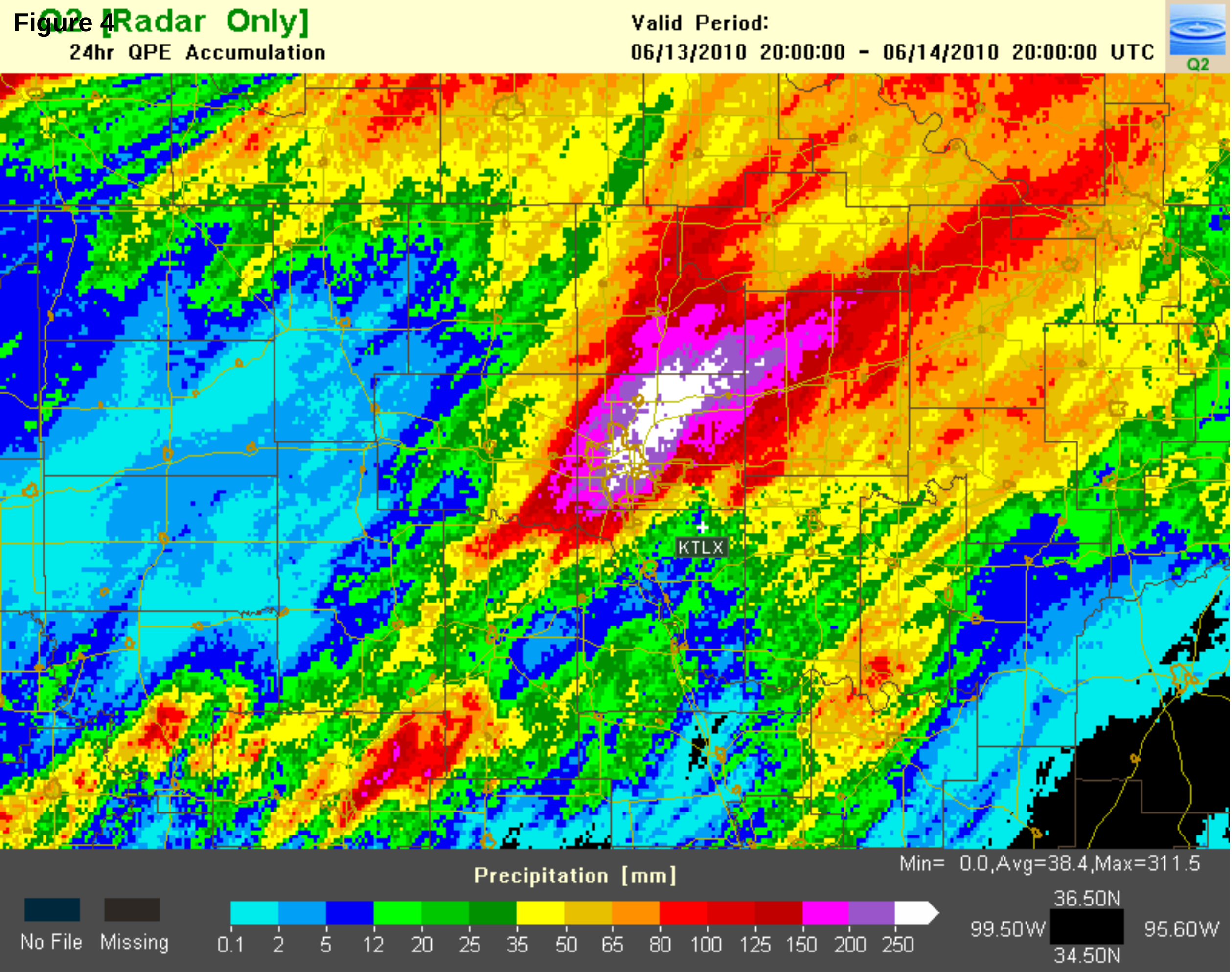
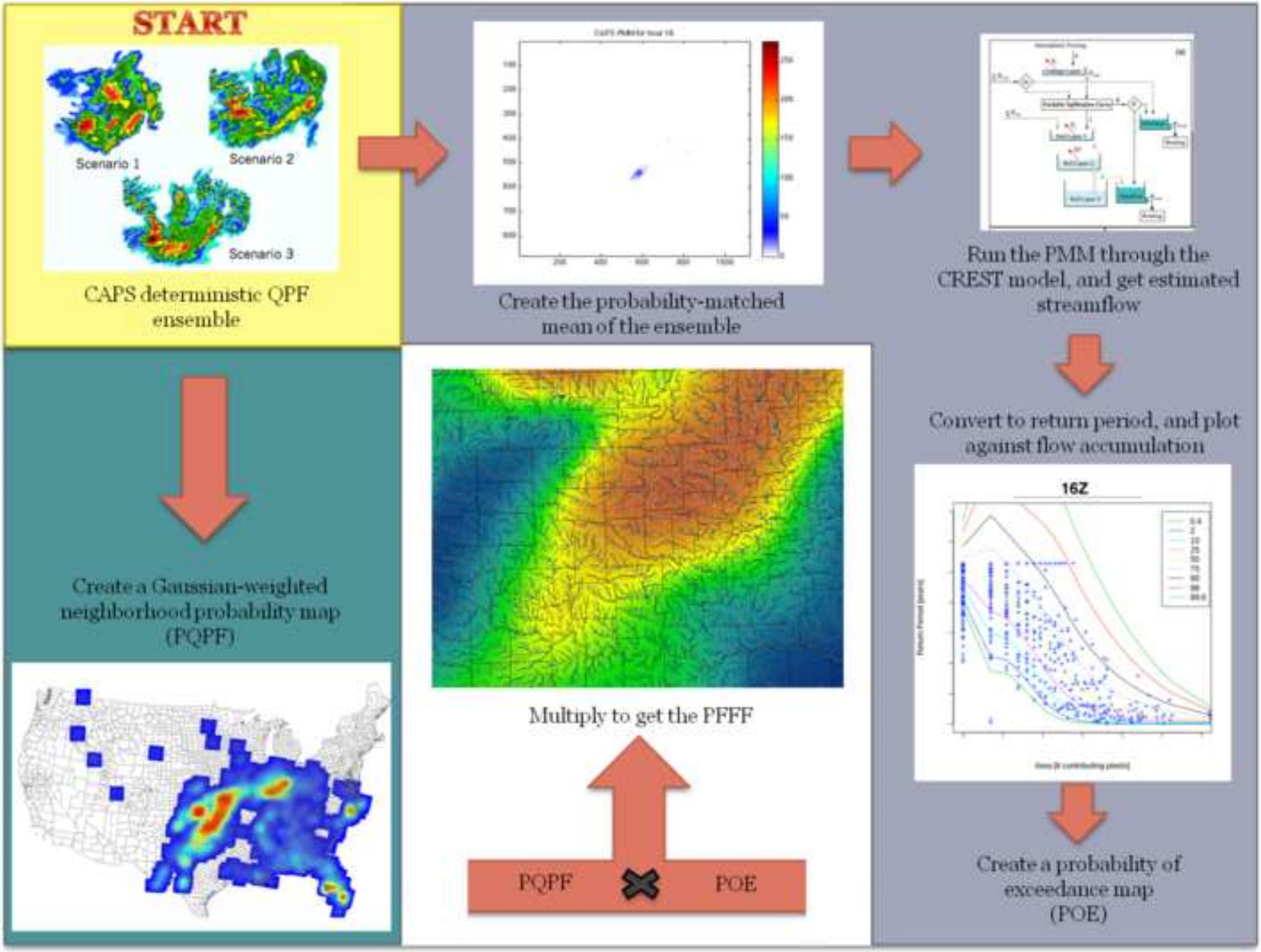

Q2 [Radar Only]
24hr QPE Accumulation
Valid Period:
06/13/2010 20:00:00 - 06/14/2010 20:00:00 UTC
Q2
KTLX
Precipitation [mm]
Min= 0.0,Avg=38.4,Max=311.5
No File
Missing
0.1
2
5
12
20
25
35
50
65
80
100
125
150
200
250
36.50N
99.50W
95.60W
34.50N

Figure 5



START

CAPS deterministic QPF ensemble

Create the probability-matched mean of the ensemble

Run the PMM through the CREST model, and get estimated streamflow

Convert to return period, and plot against flow accumulation

Create a probability of exceedance map (POE)

Create a Gaussian-weighted neighborhood probability map (PQPF)

Multiply to get the PFFF

Figure 6

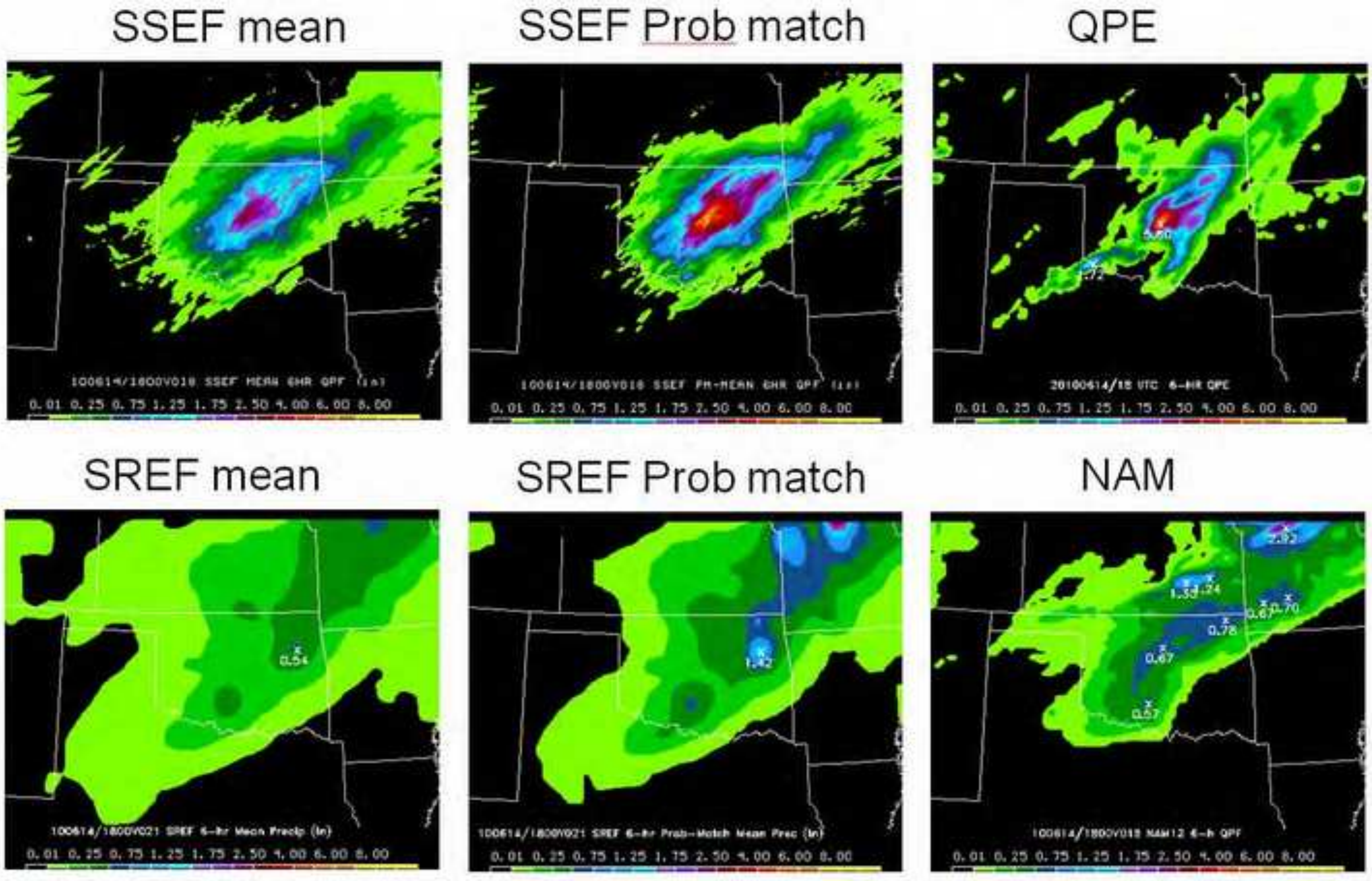

SSEF mean
SSEF Prob match
QPE
100614/1800V018 SSEF MEAN 6HR QPF (in)
100614/1800V018 SSEF PM-MEAN 6HR QPF (in)
20100614/18 UTC 6-HR QPE
5.60
1.72
0.01 0.25 0.75 1.25 1.75 2.50 4.00 6.00 8.00
SREF mean
SREF Prob match
NAM
0.54
1.42
2.92
1.33
1.24
0.67
0.70
0.78
0.67
0.57
100614/1800V021 SREF 6-hr Mean Precip (in)
100614/1800V021 SREF 6-hr Prob-Match Mean Prec (in)
100614/1800V018 NAM12 6-h QPF
0.01 0.25 0.75 1.25 1.75 2.50 4.00 6.00 8.00

Figure 7

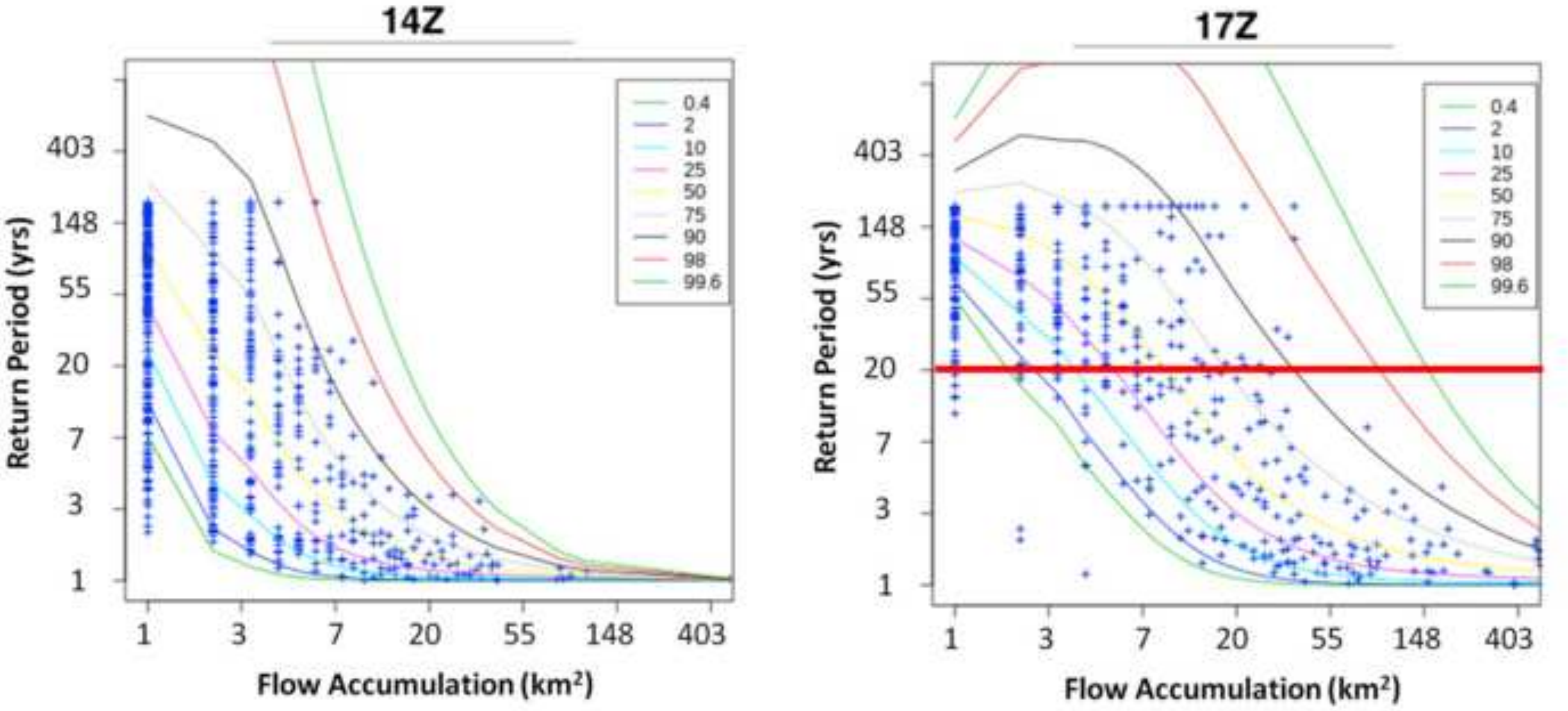
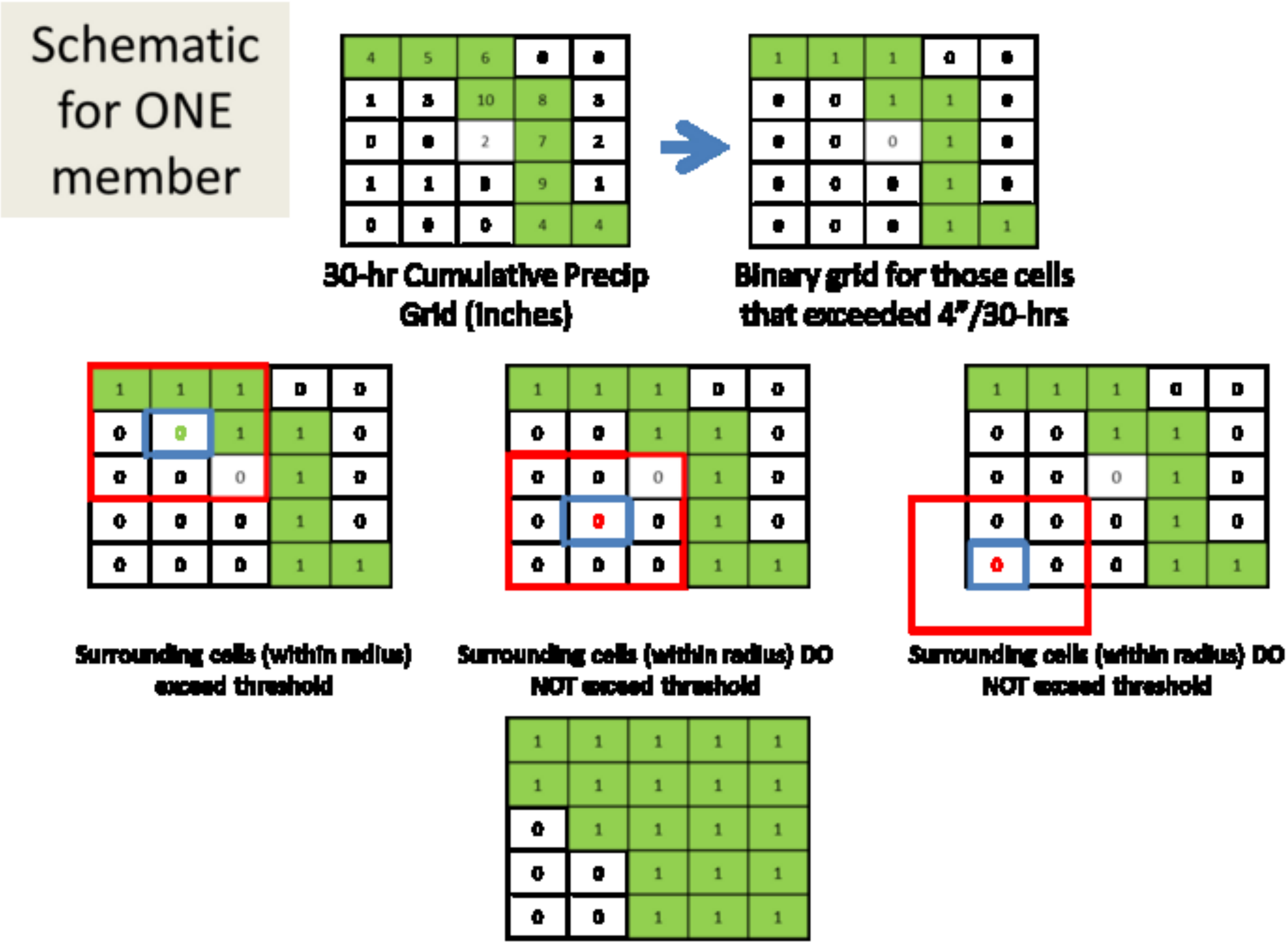

14Z
17Z
Return Period (yrs)
403
148
55
20
7
3
1
Flow Accumulation (km²)
1
3
7
20
55
148
403
0.4
2
10
25
50
75
90
98
99.6

Figure 8



Schematic for ONE member

4	5	6	0	0
1	3	10	8	3
0	0	2	7	2
1	1	0	9	1
0	0	0	4	4

30-hr Cumulative Precip Grid (Inches)

1	1	1	0	0
0	0	1	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1

Binary grid for those cells that exceeded 4"/30-hrs

1	1	1	0	0
0	0	1	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1

Surrounding cells (within radius) exceed threshold

1	1	1	0	0
0	0	1	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1

Surrounding cells (within radius) DO NOT exceed threshold

1	1	1	0	0
0	0	1	1	0
0	0	0	1	0
0	0	0	1	0
0	0	0	1	1

Surrounding cells (within radius) DO NOT exceed threshold

1	1	1	1	1
1	1	1	1	1
0	1	1	1	1
0	0	1	1	1
0	0	1	1	1

Final neighborhood binary grid

Figure 9

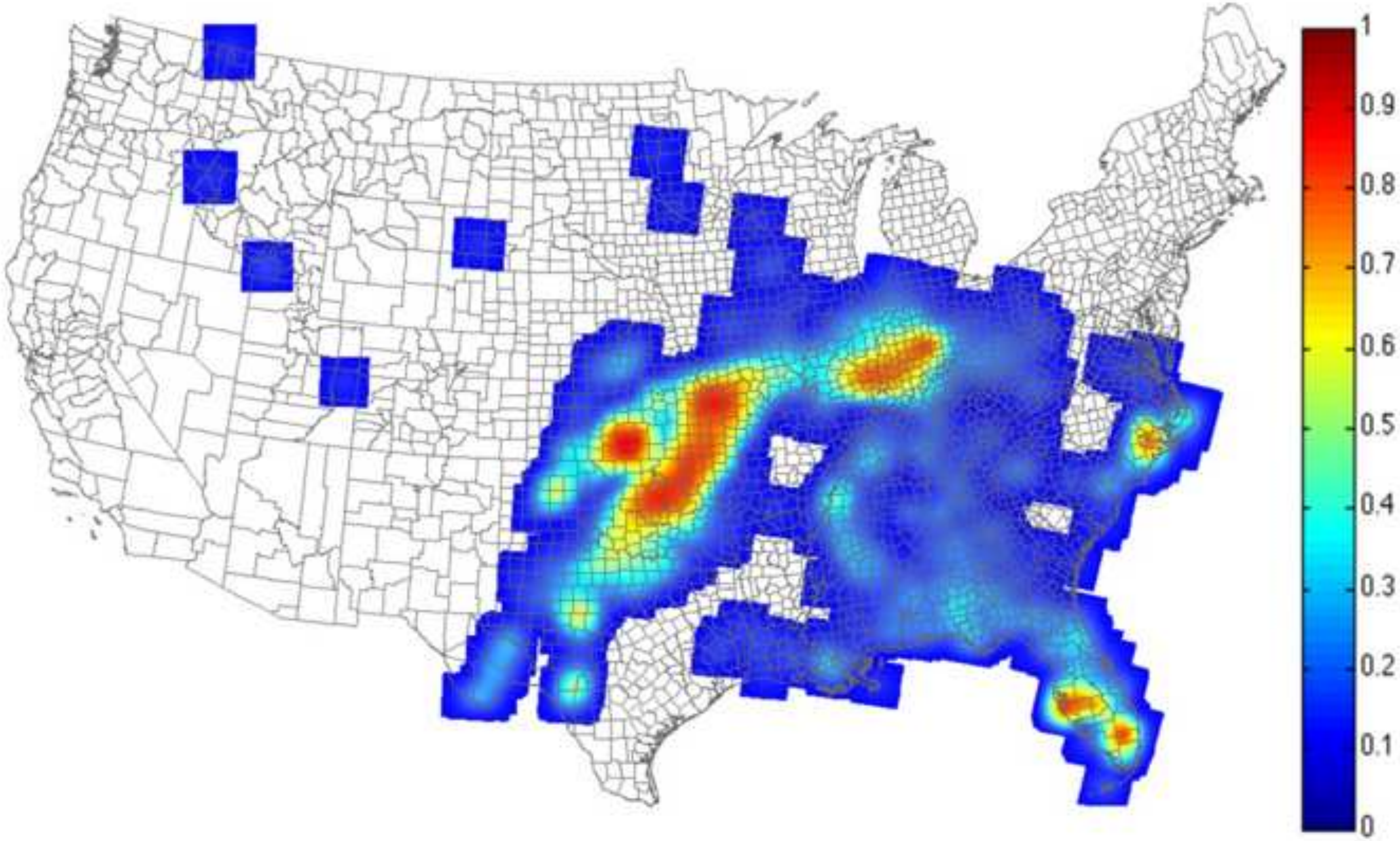


Figure 10

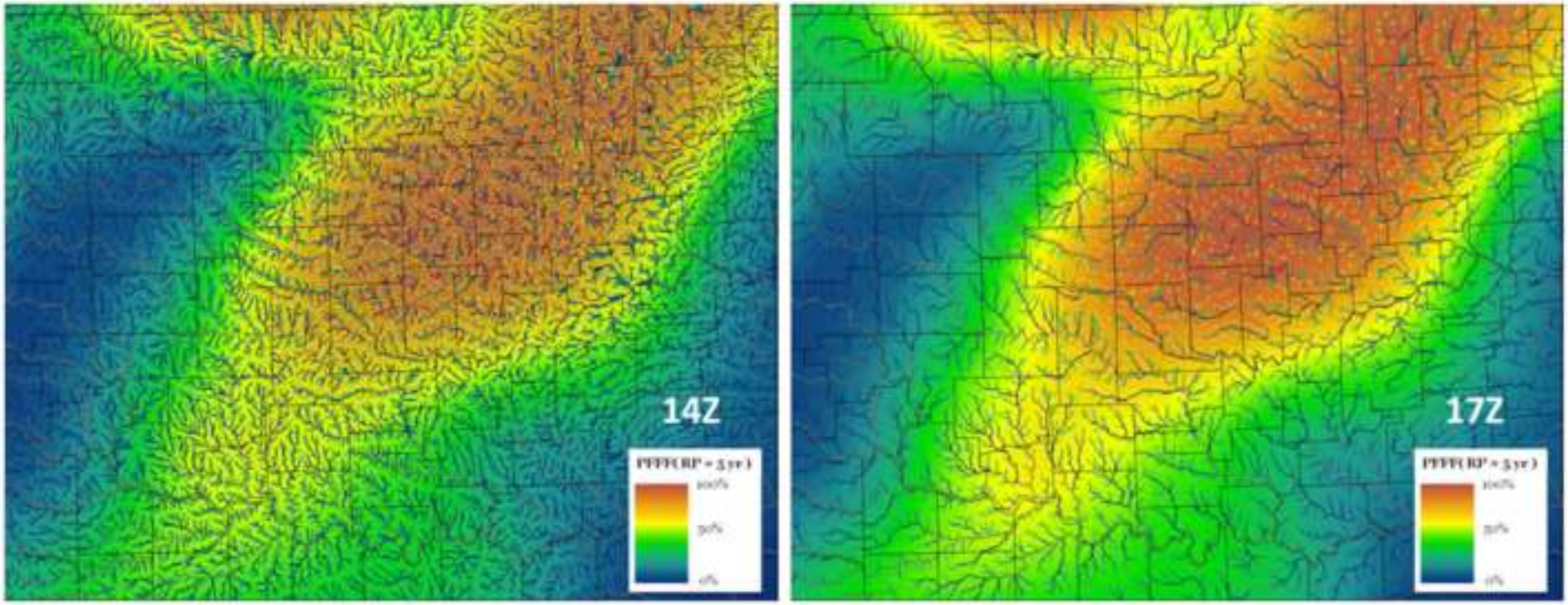


Figure 11

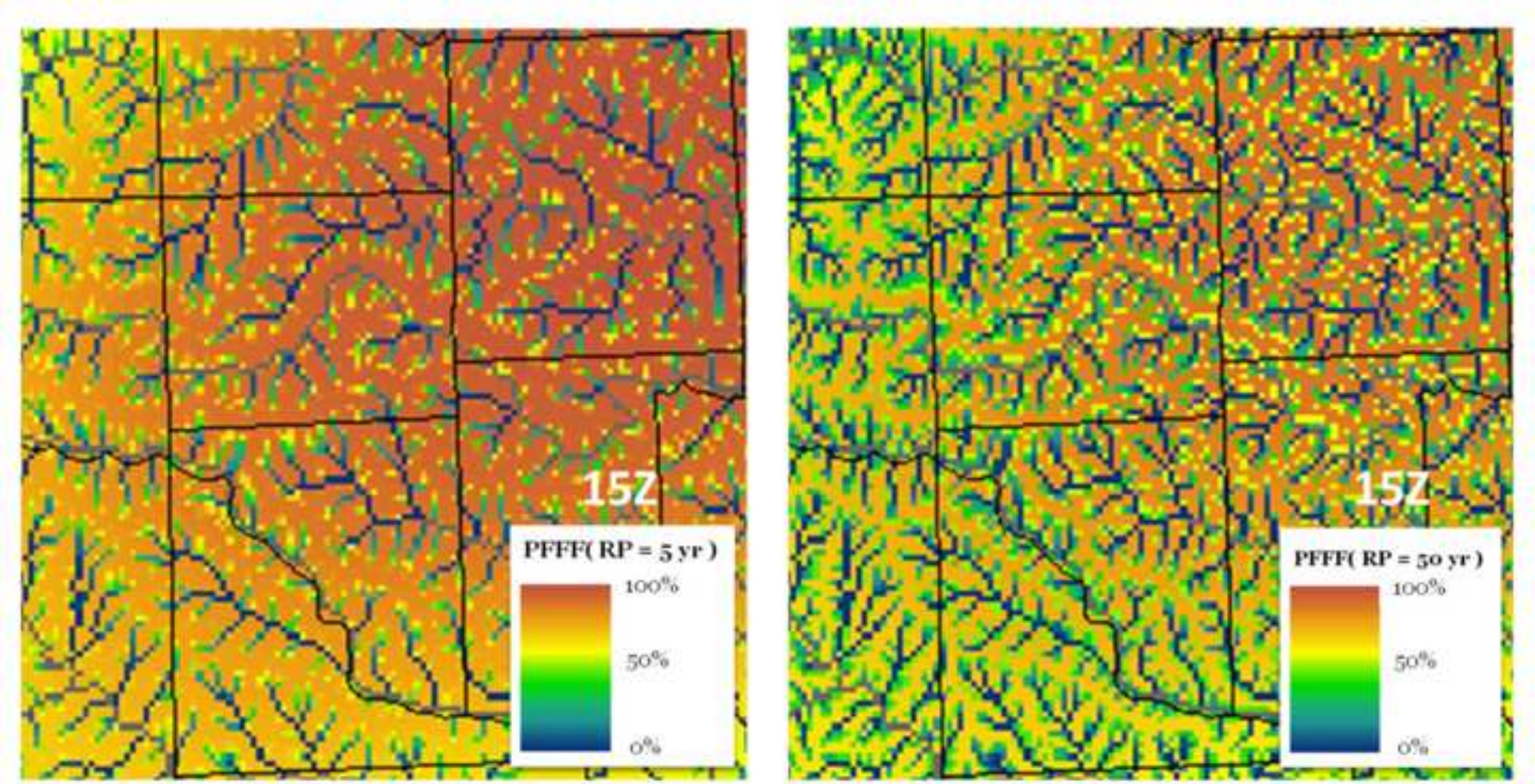
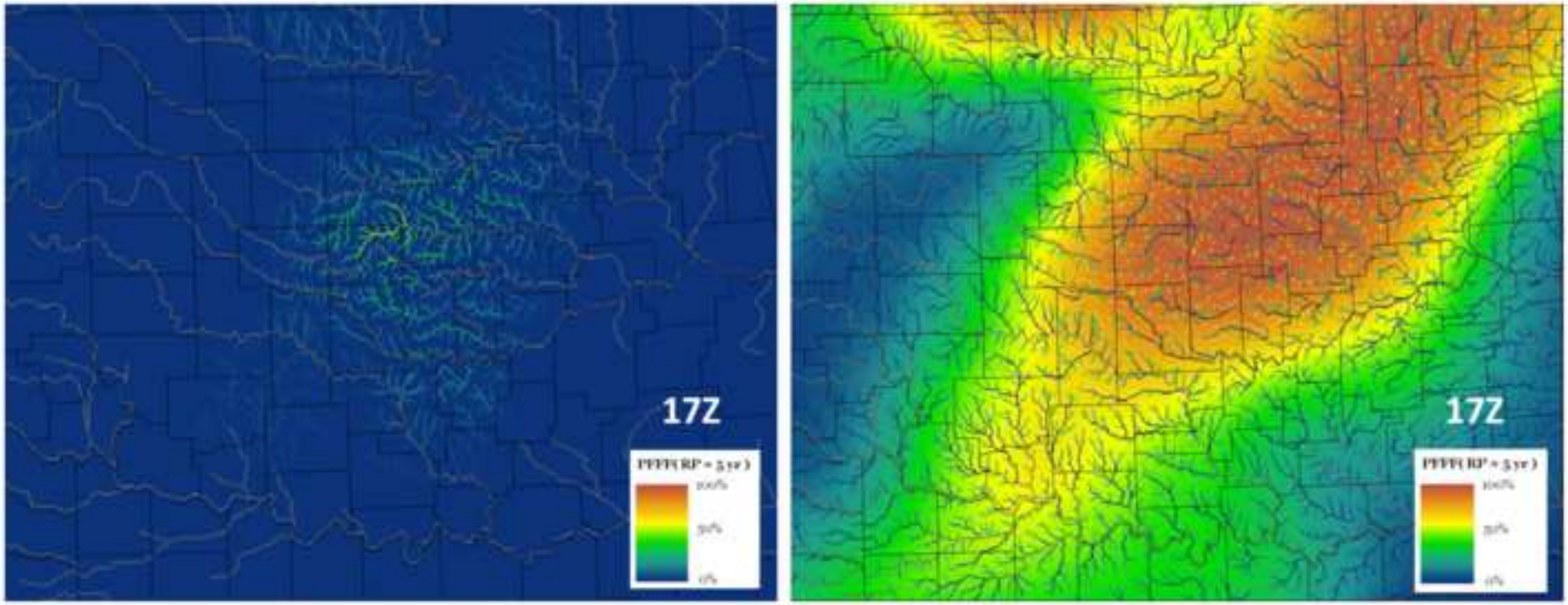


Figure 12


17Z
PFFF(RP = 5 yr)
100%
50%
0%
17Z
PFFF(RP = 5 yr)
100%
50%
0%