THESIS

MODEL POST-PROCESSING FOR THE EXTREMES: IMPROVING FORECASTS OF LOCALLY EXTREME RAINFALL

Submitted By

Gregory Reid Herman

Department of Atmospheric Science

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, CO

Spring 2016

Master's Committee:

Advisor: Russ Schumacher

Elizabeth Barnes
Daniel Cooley

ABSTRACT

MODEL POST-PROCESSING FOR THE EXTREMES: IMPROVING FORECASTS OF LOCALLY EXTREME RAINFALL

This study investigates the science of forecasting locally extreme precipitation events over the contiguous United States from a fixed-frequency perspective, as opposed to the traditionally applied fixed-quantity forecasting perspective. Frequencies are expressed in return periods, or recurrence intervals; return periods between 1-year and 100-years are analyzed for this study. Many different precipitation accumulation intervals may be considered in this perspective; this research chooses to focus on 6- and 24-hour precipitation accumulations. The research presented herein discusses the beginnings of a comprehensive forecast system to probabilistically predict extreme precipitation events using a vast suite of dynamical numerical weather prediction model guidance.

First, a recent climatology of extreme precipitation events is generated using the aforementioned fixed-frequency framework. The climatology created generally conforms with previous extreme precipitation climatologies over the US, with predominantly warm season events east of the continental divide, especially to the north away from major bodies of water, and primarily cool-season events along the Pacific coast. The performance of several operational and quasi-operational models of varying dynamical cores and model resolutions are assessed with respect to their extreme precipitation

characteristics; different biases are observed in different modeling systems, with one model dramatically overestimating extreme precipitation occurrences across the entire US, while another coarser model fails to produce the vast majority of the rarest (50-100+ year) events, especially to the east of the Rockies where most extreme precipitation events are found to be convective in nature. Some models with a longer available record of model data are employed to develop model-specific quantitative precipitation climatologies by parametrically fitting right-skewed distributions to model precipitation data, and applying these fitted climatologies for extreme precipitation forecasting. Lastly, guidance from numerous models is examined and used to generate probabilistic forecasts for locally extreme rainfall events. Numerous methods, from the simple to the complex, are explored for generating forecast probabilities; it is found that more sophisticated methods of generating forecast probabilities from an ensemble of models can significantly improve forecast quality in every metric examined when compared with the most traditional probabilistic forecasting approach. The research concludes with the application of the forecast system to a recent extreme rainfall outbreak which impacted several regions of the United States.

ACKNOWLEDGEMENTS

This research was supported by National Oceanic and Atmospheric Administration grant NA14OAR4320125, Amendment 26, National Science Foundation grant ACI-1450089, and National Aeronautics and Space Administration grant NNX15AD11G.

I would like to thank the Hydrometeorological Design Studies center and National Centers for Environmental Prediction for their previous work on precipitation analysis and freely providing access to the Atlas 14 and Stage IV Precipitation Analysis products, respectively. Numerous people and organizations have been very helpful in generously providing access to both real-time and archive model simulations: Adam Clark and the National Severe Storms Laboratory for providing access to their long-running high-resolution model; Kelly Mahoney, Eric James, and Bob Lipschutz for supplying HRRR data; Tom Hamill and all others who worked to generate the NOAA Second Generation Global Ensemble Reforecast dataset; and NCEP again for their availability of their full suite of model simulations run several times per day.

I would like to thank my advisor, Dr. Russ Schumacher, for his continual guidance, support, and insights throughout this research process, which has proved tremendously valuable on numerous occasions and helped substantially improve the quality of the research presented herein. I would also like to thank my graduate committee, Drs. Elizabeth Barnes and Daniel Cooley, for their dedication and helpful comments and suggestions. I would like to thank the entire Schumacher research group- Erik Nielsen, Stacey Hitchcock, Bob Tournay, Dr. John Peters, Peter Goble, Sam Childs, and Nathan Kelly- for their support throughout the research process and several helpful and productive research discussions throughout. Finally, I wish to acknowledge my family and friends for their perpetual encouragement and support of all of my pursuits.

TABLE OF CONTENTS

List of Figures

List of Tables

List of Acronyms

ADVA: Advanced Algorithm (see chapter 5.1 for description)
AEP: Annual Exceedance Probability (see chapter 2.4 for description)
AI: Accumulation Interval
AMS: Annual Maximum Series (see chapter 2.4 for description)
ARI: Average Recurrence Interval (see chapter 2.4 for description)
BC: Boundary Condition

BMA: Bayesian Model Averaging
BS: Brier Score (see chapter 2.7 for description)
BSS: Brier Skill Score (see chapter 2.7 for description)
CAPE: Convective Available Potential Energy
CDF: Cumulative Density Function (see chapter 2.3 for description)
CLR: Cost-Loss Ratio (see chapter 2.7/5.2 for description)
CMF: Cumulative Mass Function (see chapter 2.3 for description)
CONUS: Contiguous United States
CoV: Coefficient of Variation (see chapter 4.1 for description)
CSGD: Censored, Shifted Gamma Distribution
CT: Contingency Table (see chapter 2.7 for description)
DC: Dynamical Core
DF: Direct Fit
DNUR: Deterministic Neighborhood Uniform Ranks (see chapter 5.1 for description)
DS: Direct Solve (see chapter 2.4 for description)
DV: Democratic Voting
EAS: Earth Atmosphere System
ECMWF: European Center for Medium Range Weather Forecasts
EFI: Extreme Forecast Index
ENUR: Ensemble Neighborhood Uniform Ranks (see chapter 5.1 for description)
EP: Evolutionary Program
EPS: Ensemble Prediction System (see chapter 2.3 for description)
EVDG: Economic Value Diagram (see chapter 2.7 for description)
EXP: Exponential Distribution (see chapter 2.4 for description)
EXPO: Offsetted Exponential Distribution (see chapter 4.1 for description)
EVT: Extreme Value Theory (see chapter 2.4 for description)
FDS: Full Dry Series (see chapter 2.4 for description)
FP: Forecast Probability
FRPSS: Fractions Rank Probability Skill Score (see chapter 2.7 for description)
FS: Forecast System
FSS: Fractions Skill Score (see chapter 2.7 for description)
FTG: Fisher-Tippett-Gnedenko Theorem (see chapter 2.4 for description)
FWS: Full Wet Series (see chapter 2.4 for description)
GAM: Gamma Distribution (see chapter 2.4 for description)
GEFS: Global Ensemble Forecast System (see chapter 2.2 for description)
GEFS/R: Global Ensemble Forecast System Reforecast Version 2 (see chapter 2.2 for description)
GEV: Generalized Right-Skewed Distribution (see chapter 2.4 for description)
GEVO: Offsetted Generalized Right-Skewed Distribution (see chapter 4.1 for description)
GFS: Global Forecast System (see chapter 2.2 for description)
GLO: Generalized Logistic Distribution (see chapter 2.4 for description)
GLOO: Offsetted Generalized Logistic Distribution (see chapter 4.1 for description)
GNO: Generalized Normal Distribution (see chapter 2.4 for description)
GNOO: Offsetted Generalized Normal Distribution (see chapter 4.1 for description)
GPA: Generalized Pareto Distribution (see chapter 2.4 for description)
GPAO: Offsetted Generalized Pareto Distribution (see chapter 4.1 for description)
GUM: Gumbel Distribution (see chapter 2.4 for description)

HDSC: Hydrometeorological Design Studies Center
HP: High Precipitation
HRMD: High Rate, Moderate Duration
HRRR: High Resolution Rapid Refresh (see chapter 2.2 for description)
IC: Initial Condition
INTA: Intermediate Algorithm (see chapter 5.1 for description)
KAP: 4-parameter Kappa Distribution (see chapter 2.4 for description)
KNN: K-Nearest Neighbors (see chapter 2.6 for description)
LAM: Limited Area Model
MCS: Mesoscale Convective System
MLE: Maximum Likelihood Estimation (see chapter 2.4 for description)
MoLM: Method of L-Moments (see chapter 2.4 for description)
MoM: Method of Moments (see chapter 2.4 for description)
MOS: Model Output Statistics (see chapter 2.1 for description)
MP: Model Physics
MRHD: Moderate Rate, High Duration
MSE: Mean Squared Error
NAM-NEST: North American Mesoscale Model High Resolution Nest (see chapter 2.2 for description)
NCAR: National Center for Atmospheric Research
NCEP: National Centers for Environmental Prediction
NDV: Neighborhood Democratic Voting (see chapter 5.1 for description)
NIVA: Naïve Algorithm (see chapter 5.1 for description)
NOAA: National Oceanic and Atmospheric Administration
NPUR: Neighborhood Point Uniform Ranks (see chapter 5.1 for description)
NSSL: National Severs Storms Laboratory
NUR: Neighborhood Uniform Ranks (see chapter 5.1 for description)
NWP: Numerical Weather Prediction
OF: Output Forecast (see chapter 2.3 for description)
ORF: Observed Relative Frequency (see chapter 2.3 for description)
PDF: Probability Density Function (see chapter 2.3 for description)
PDS: Partial Duration Series (see chapter 2.4 for description)
PDV: Point Democratic Voting (see chapter 5.1 for description)
PE3: Pearson Type III Distribution (see chapter 2.4 for description)
PE3O: Offsetted Pearson Type III Distribution (see chapter 4.1 for description)
PFS: Probabilistic Forecast System (see chapter 2.3 for description)
PMF: Probability Mass Function (see chapter 2.3 for description)
PoP: Probability of (measurable) Precipitation
POT: Peaks over Threshold (see chapter 2.4 for description)
PQPF: Probabilistic Quantitative Precipitation Forecast
PUR: Point Uniform Ranks (see chapter 5.1 for description)
QF: Quantile Function (see chapter 2.4 for description)
QPF: Quantitative Precipitation Forecast
RBS: Reliability Brier Score (see chapter 5.2 for description)
RCS: Reliability Confidence Score (see chapter 5.2 for description)
RD: Reliability Diagram (see chapter 2.7 for description)
RL: Reliability Line (see chapter 2.7 for description)

RP: Return Period (see chapter 2.4 for description)
RP(S)S: Rank Probability (Skill) Score (see chapter 2.7 for description)
RPT: Return Period Threshold
RSD: Right-Skewed Distribution (see chapter 2.4 for description)
RSS: Reliability Skill Score (see chapter 5.2 for description)
SVC: Support Vector Classification (see chapter 2.6 for description)
SVM: Support Vector Machine (see chapter 2.6 for description)
SVS: Summary Value Score (see chapter 5.2 for description)
TC: Tropical Cyclone
TF: True Forecast (see chapter 2.3 for description)
UR: Uniform Ranks
VPDF: Verifying PDF (see chapter 2.3 for description)
VS: Value Score (see chapter 2.7 for description)
WEI: Weibull Distribution (see chapter 2.4 for description)
WEIO: Offsetted Weibull Distribution (see chapter 4.1 for description)
WRF: Weather Research and Forecasting Model

1 Introduction, Motivation, and Overview

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 deaths in the United States over the last 30 years than any other single weather hazard, including tornadoes, hurricanes, lightning, and other windstorms. Floods can heavily damage or destroy buildings, roads, crops, and other property; in 2014, flash floods were responsible for more economic damage than any other weather hazard, with nearly $2.5B in reported flash flood damages occurring that year. Though some damages from extreme rainfall and flooding are inevitable, appropriate preparedness can greatly alleviate damages and almost completely eliminate flood fatalities. As such, accurate forecasts of extreme precipitation and flooding are of immense value to society.

Ultimately, flood forecasting is what most directly addresses societal impacts associated with heavy precipitation. Flood forecasting is performed by forcing a hydrologic model with precipitation forecasts from an atmospheric model, or numerical weather prediction (NWP) model. Both observations and modeling has shown that hydrologic response is extremely sensitive to the amount of precipitation, the location where the precipitation falls, and antecedent conditions. Thus, even what is often considered a fairly good precipitation forecast may produce a very inaccurate response in the hydrologic model, resulting in a poor flood forecast. Due to these high sensitivities, using hydrologic models for real-time flood and especially flash flood forecasting is presently exceedingly difficult, and perhaps not yet feasible. There nevertheless exists a strong correlation between precipitation amount and associated impacts. This correspondence is not uniform, however. In some areas, such as the southeast United States, an inch or two of rain over a day-long period is commonplace, and both the native ecosystem/soils and man-made infrastructure are adapted to accommodate this rainfall with

minimal impacts. In areas of the arid west, this amount of precipitation is much rarer, and much larger flood impacts may be experienced. Typically, due to adaptation to the local precipitation climatology, local impacts associated with precipitation are more closely tied to the rarity of receiving a given amount of precipitation over a specified period than impacts being simply associated with a fixed precipitation amount. It therefore follows that, without performing hydrologic modeling, a useful proxy for the impacts of extreme precipitation is the quantification of the rarity of forecasted precipitation accumulation at a given location over a particular length of time. Often in fixed frequency applications when concerned only with rare events, event frequency is expressed by means of return periods (RPs) or, equivalently, average recurrence intervals (ARIs). In this context, an N-year RP refers to a long term average occurrence of once per every N years for the specified location and precipitation accumulation interval (AI), though there will of course be N-year periods experiencing several events and other periods experiencing no events at all. For a given location and AI, the precipitation accumulation required yielding an ARI of exactly N-years is termed the N-year RP threshold. Because of the impacts associated with extreme precipitation, the utility of accurate locally extreme precipitation forecasts, the immense challenges associated with real-time hydrologic modeling, and the utility of precipitation accumulation rarity on hydrologic impacts, the research conducted and presented here seeks to improve real-time forecasts of locally extreme precipitation from the return period framework.

There exist many plausible routes to seek in attempting to achieve the goal of improving forecasts of locally extreme precipitation. The question is which avenue or avenues will best achieve this goal given the NWP models and forecast products in place today. More than a decade ago, Fritsch and Carbone (hereafter FC04) laid out some of the leading challenges in quantitative precipitation forecasts (QPFs) at the time, and many remain true today. They argued that QPF, and in particular warm-season QPF, is the worst forecast predictand of interest in all forecast systems of the time, and the performance gap with other predictands was increasing since warm-season QPFs were not

improving as quickly as other areas. They argue that the warm-season QPF challenge will continue for the foreseeable future, and given the extent of the societal impact of precipitation and especially heavy precipitation, great effort must be invested towards alleviating the QPF deficiencies that existed then and remain today. The article presented a targeted research and development plan for moving forward as a community. Their first key goal was the generation and dissemination of forecast guidance in probabilistic form. FC04 argues that this is critical for several reasons: 1) the importance of probabilistic forecast information for end user decision scenarios and risk management; 2) the limited skill and lead times over which deterministic guidance exhibits skill argue for a probabilistic framework on top of a deterministic foundation; 3) the incomplete representation of moist convection, especially in models which by necessity apply a cumulus presentation fails to adequately capture the statistical properties of moist convection; and 4) statistical post-processing of model forecasts can alleviate inherent model bias and quantify forecast uncertainty, even absent an improved understanding of the physical precipitation processes. FC04 also proposed several additional specific areas that are critical to target and improve going forward: 1) acquiring an improved understanding of the economic and social aspects of QPFs so that the full value of the available meteorological information may be realized; 2) develop new models and refine existing models to better represent physical processes such as cloud microphysics and moist turbulence; 3) improving understanding of microphysics and convective systems, particularly the mechanisms for propagation, dissipation, and regeneration; 4) improving atmospheric observations, with particular emphasis on widening precipitation coverage both at the surface and aloft, and properly observing aerosol extent and composition; 5) improving data assimilation; 6) improving probabilistic forecast guidance through a combination of observationally-derived, model-derived, and blend-derived guidance, depending on the lead time and application; 7) improving QPF verification methods, especially in relation to end user goals; and 8) develop products of use to end users, with a particular emphasis on hydrologic modeling and forecasting. Lastly, FC04 outlined a roadmap for how to best accomplish these

goals. They separated these action items into 'early' and 'continuing' activities; many items, even in the 'early' stage category, are still very much ongoing, and some still in preliminary phases. Important action items that have received considerable improvement since the publication of this manuscript include: 1) "compil[ing] a high-quality, high-resolution database of precipitation properties"; 2) "develop[ing] improved metrics for verifying mesoscale precipitation forecasts in both time and space, especially for guidance provided in gridded probabilistic form"; and 3) "evaluat[ing] the benefit from very-high horizontal and vertical resolution observations over the continent". These advancements have made the ability to advance in many of the other target areas much more feasible than it was a decade ago. While it is not realistic for a single research project to address all of these goals or all of the recommended action items, the research proposed and explored herein attempts to address many of these points which, to date, have been neglected and/or underexplored. In particular, with respect to the proposed action items, this study aims to: 1) "combine ensemble techniques and traditional statistical postprocessing techniques to provide calibrated probabilities, ensemble fields, and unbiased ensemble statistics"; 2) further "develop improved metrics for verifying mesoscale precipitation forecasts … for guidance provided in gridded probabilistic form"; 3) "determine appropriate methodologies to evaluate case-dependent uncertainty for precipitation events"; and to a lesser extent 4) "construct nowcasting techniques that utilize high-resolution observations and numerical model output to generate categorical probabilistic QPFs"; 5) "assess the feasibility of periodically producing a retrospective archive of the high-resolution … model output and ensemble runs of that model"; and 6) "develop techniques to integrate ensemble precipitation forecasts from different forecast systems ranging from nowcasts to regional to medium range to climate into a seamless and consistent set of ensemble forcing". In addressing so many of these action goals, it is hoped that this research will significantly advance the field towards improved probabilistic QPF prediction and in so doing, contribute to the forecasting community at large.

As FC04 alluded, human forecasting in the modern age, despite substantial advances in both numerical models and our physical understanding of atmospheric processes, presents new challenges. There is an overwhelming amount of model guidance produced each day. Many nations and/or regions have their own operational center with their own global model and global ensemble, in addition to possible regional modeling efforts as well. These often run two to four times daily, and operational global ensembles often range in size from 10-50 members- each a distinct run to consider. In the United States, the National Centers for Environmental Prediction (NCEP) runs a full-scale regional 21-member ensemble four times daily as well. In addition to operational products, centers often also have experimental products running on a regular cycle as well: upgrades to existing models running in parallel, high-resolution implementations of existing models, or completely new modeling frameworks. The amount of modeling produced solely from the operational centers around the globe is already daunting for a forecaster to exhaustively inspect and ingest, but it is still only a portion of the total modeling data available. Specialized forecast centers such as the National Severe Storms Laboratory (NSSL) and research institutions such as the National Center for Atmospheric Research (NCAR) have their own modeling efforts, both with deterministic runs and high-resolution ensembles. Many universities also run real-time model simulations with varying levels of dedication, from daily runs of a coarse model to full-scale regional ensembles or multiple runs per day of very high resolution NWP models. All told, there are often hundreds of model runs a forecaster has at his or her disposal to inform their forecast. The forecaster's challenge is to intelligently use all of this guidance to produce the best possible forecast. But with the amount of data available and new guidance constantly emerging, it is nearly impossible for a human to thoroughly inspect and sort through the full suite of information at their disposal. Challenges exist for the modern forecaster on the actual forecasting front as well. Forecasting is not limited to simply making a single deterministic forecast and hoping to minimize the error of that forecast. A robust forecast must also include uncertainty quantification, and a probabilistic assessment

of event likelihood, particularly for rare, high-impact events for which most end users are highly sensitive to the verifying outcome. A computer can ingest this large magnitude of forecast data much more quickly than a human; the question is whether an automated algorithm can use the forecast information as effectively as a skilled human forecaster. The forecast system (FS) developed herein seeks to examine this question in the limited capacity of probabilistic locally extreme precipitation forecasting. The FS will not generate traditional QPF predictions. Instead, it seeks to ingest a large quantity of NWP model guidance from numerous sources and utilize it to generate probabilistic forecasts of locally extreme precipitation of varying degrees of rarity, or extremeness.

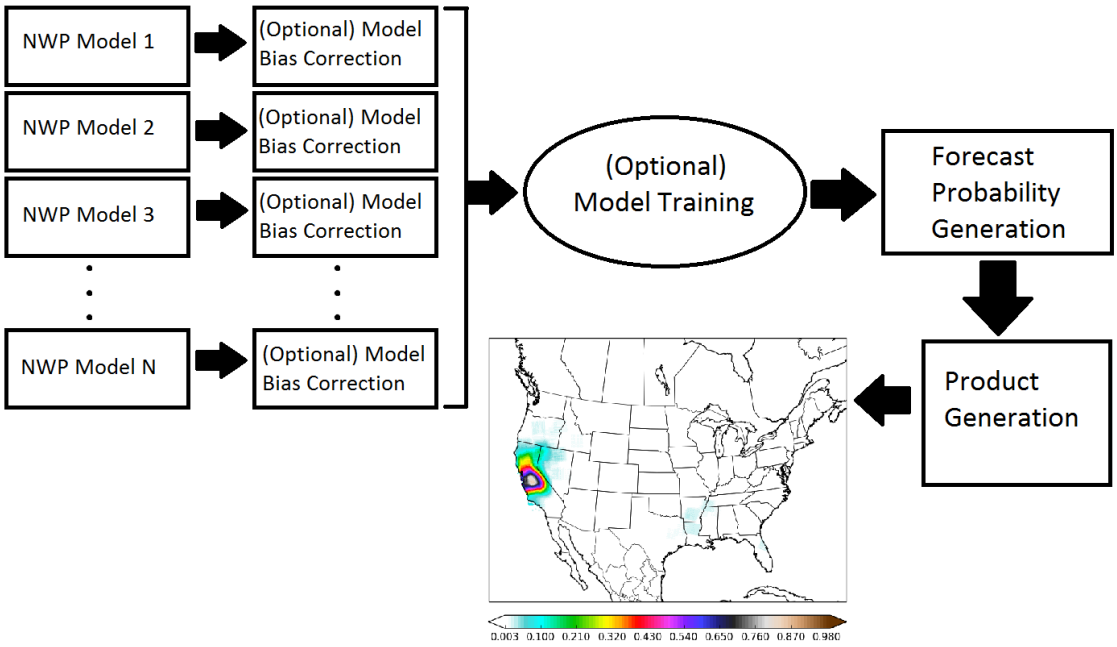

Figure 1.1: Schematic of forecast system pipeline.

Figure 1.1 provides a rough overview of the steps involved with this FS. The process begins with ingesting NWP guidance from different sources and of different scales and resolutions. After providing necessary background in Chapter Two, an examination of the performance and characteristics of these

individual modeling systems with respect to locally extreme precipitation prediction is presented in Chapter Three. As will be discussed in more detail later, individual modeling systems each exhibit biases associated with precipitation; with extreme events, it is often the case that these biases are exacerbated or new biases emerge (e.g. Marzban 1998). These biases vary from model to model. An examination and evaluation of approaches to correct for these individual model biases is made in Chapter Four. For reasons to be explained in more detail later, performance and bias explored in Chapters Three and Four only examine bulk behavior- in the case of bias, forecasts that are always too high or always too low. But for numerous reasons, models can also have different biases under different meteorological regimes, and also perform more or less skillfully on average depending on the meteorological context. Model training uses past forecasts from all members of the ad-hoc ensemble to attempt to identify these dynamic biases and patterns and make appropriate corrections based on historical performance. These methods can also correct for persistent displacement errors more readily than the bulk methods of Chapter Four. Now, having ingested all model guidance and optionally performed any bias correction to the members or the ensemble as a whole, the forecast information is applied towards generating probabilistic forecasts of locally extreme rainfall framed in the context of return periods. Various techniques for probability generation and the ensemble training methods are explored in Chapter Five. An application of the forecast system, including examining the effects of the individual pipeline components, on a recent extreme rainfall 'outbreak' is used to synthesize and solidify the discussion of the previous chapters and is presented in Chapter Six. The overall results of the research will be summarized in Chapter Seven.

2 Background

The research conducted in this study is somewhat technical, borrowing from knowledge of many different fields including meteorology, statistics, mathematics, and computer science. Due to the breadth of the information and knowledge used, this chapter aims to present a concise but sufficiently thorough overview from all of the background fields in order for any scientifically and mathematically inclined reader to adequately comprehend and appreciate the research presented in subsequent chapters. This chapter is laid out in seven sections, each covering a different background area. Section 2.1 presents a brief overview of underlying principles and history of NWP, with an emphasis on statistical forecasting techniques and those methods which pertain to extreme precipitation post-processing. Section 2.2 concerns specific NWP models that will be used as input data for the research presented in future chapters. Section 2.3 provides an introduction to probabilistic forecasting and ensemble prediction systems (EPSs). Section 2.4 gives an introduction to extreme value theory (EVT) and associated applications, while Section 2.5 describes observational precipitation datasets used by this research and a brief overview of the United States extreme precipitation climatology. Section 2.6 provides a brief description of machine learning algorithms applied in subsequent research, and Section 2.7 concludes with a targeted presentation of the forecast validation methods used for this study.

2.1 Numerical Weather Prediction and Statistical Forecasting

2.1.1 Dynamical and Statistical Modeling

There exists a dichotomy of sorts in atmospheric modeling. Traditionally, atmospheric modeling and forecasting is separated into two camps: dynamical modeling and statistical modeling. Both methods have a long history of research and numerous approaches to application. However, all approaches share some commonalities. In dynamical modeling, one begins with a numerical

representation of the current state of the atmosphere. Often, this is represented by having current numerical values for a suite of atmospheric variables on a three-dimensional grid. Equations governing the evolution of the atmosphere are then used in conjunction with the initial state to ascertain a representation of the atmosphere at future times. The process of taking observational data and previous model forecasts to generate a new representation of the current atmospheric state is known as data assimilation; using the equations to predict future states from there is termed model integration. Model resolution refers to the smallest scale physical processes that the dynamical model is able to adequately resolve. Model resolution is largely a function of model grid spacing; an approximate rule-of-thumb being that dynamical models can begin to resolve phenomena occurring on scales at least four to five times the model grid spacing in the dimension of interest. Obtaining an accurate representation of the atmosphere may require these smaller-scale phenomena being represented; this is done by means of parameterization. Common examples include cumulus, microphysics, and land surface parameterizations; these and others will be discussed in more detail in section 2.2.

Statistical modeling does not directly simulate the atmosphere. Instead, it uses historical observations to derive statistical relationships between fields of interest, predictands, and other observables, or predictors. There are a plethora of approaches to implementation, including regression, clustering, and more advanced machine learning approaches. Some of these will be discussed in more detail in section 2.6. Statistical-Dynamical modeling, often considered a subset of statistical modeling, is in essence the application of statistical forecasting approaches to dynamical model variables. This is primarily the approach that will be used in this research. Targeted historical and current developments in statistical-dynamical modeling will be discussed in the subsequent subsections of this section.

2.1.2 Model Output Statistics and Linear Regression

Model Output Statistics, or MOS, is the first major operational implementation of a statistical forecasting system for general-purpose forecasting. Initially developed beginning in 1965, and implemented operationally from 1976 onwards, MOS has been the operational state-of-the-art for four decades. It is based on the simple, yet effective, technique known as *multivariate linear regression* (Glahn and Lowry 1972).

In multivariate linear regression, one begins with a set of predictands of interest. For MOS, this includes temperature, dew point, wind speed, wind direction, precipitation, ceiling, visibility, and cloud cover at a variety of lead times separated by three to twelve hour intervals, depending on the specific MOS product. Associated with each predictand is a set of candidate predictors. For MOS, there is a vast suite of *N* candidate predictors; some are 'static', such as station elevation, latitude, and longitude; sinusoid functions of the time of day and time of year; and climatological weather at the station, while others are 'dynamic', changing from day to day and year to year. 'Dynamic' variables may include both current observations in addition to predictions from an operational dynamical model. It should also be noted that some predictors are 'derived'; they're functions of a base predictor, such as the square of the dynamical model's temperature forecast (Glahn and Lowry 1972). Fitting a regression model involves using training data to fit an equation of the form: $\boldsymbol{y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}$, where **y** is a vector of length *m,* with each element corresponding to a unique observation for the predictand of interest, **X** an *n* by *m* matrix populated by the *n* predictor values corresponding to each of the *m* predictand observations, $\boldsymbol{\beta}$ a weights vector of length *n* relating each predictor to the predictand, and $\boldsymbol{\varepsilon}$ an error vector of length m, effectively the residual between the predictand and the corresponding result of the product of $\boldsymbol{\beta X}$ for each observation (Wilks 2011). The goal in fitting a linear regression model is to minimize the sum of the squared residuals: $\mathrm{SSR}(\beta) = \sum_{q=1}^{M}(y_q - \boldsymbol{x_q\beta})^2 = \sum_{q=1}^{M} {\varepsilon_q}^2$. The weights, or equation coefficients,

$\boldsymbol{\beta}$, are fixed to minimize this function: $\widehat{\boldsymbol{\beta}} = argmin\big(SSR(\boldsymbol{\beta})\big)$. The minimum can be obtained by differentiating the SSR function and equating it to zero:

$$0 = \frac{d}{d\boldsymbol{\beta}} \sum_{q=1}^{M} (y_q - \boldsymbol{x_q}\boldsymbol{\beta})^2 = \frac{d}{d\boldsymbol{\beta}} \sum_{q=1}^{M} {y_q}^2 - 2y_q\boldsymbol{x_q}\boldsymbol{\beta} + \boldsymbol{x_q}^2\boldsymbol{\beta}^2$$

$$= \sum_{q=1}^{M} \frac{d}{d\boldsymbol{\beta}} ({y_q}^2 - 2y_q\boldsymbol{x_q}\boldsymbol{\beta} + \boldsymbol{x_q}^2\boldsymbol{\beta}^2) = 2\big(\boldsymbol{x_q}^2\widehat{\boldsymbol{\beta}} - y_q\boldsymbol{x_q}\big)$$

$\boldsymbol{x^2}\widehat{\boldsymbol{\beta}} = \boldsymbol{yx}$ (note: x is *m*x*n*, y is 1x*m*)

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{x^T x})^{-1}\boldsymbol{x^T y}$$

Once coefficients for the $\boldsymbol{\beta}$ vector have been computed, using the statistical model to generate predictions is trivial; the new predictor values are gathered, and the inner product of the predictor vector $\boldsymbol{x}$ with $\boldsymbol{\beta}$ yields the model prediction for the predictand of interest, $\hat{y}$. However, one important step of the model development process has been overlooked: the selection of predictors to use in the statistical model. In statistics and machine learning, the *bias-variance tradeoff* refers to the problem of minimizing two sources of error in fitting a statistical model. Error from *bias* occurs due to erroneous- often too simple- assumptions about the behavior of the predictand relative to the predictand. *Variance*, in contrast, refers to the statistical model's sensitivity to the training data; a high variance model will change drastically when trained on two different samples extracted from the same population, while a low variance model will not. High variance models are often said to be *overfit*- the model is fitting the noise in the training data rather than just the underlying predictor-predictand relationship- while high bias models are often said to be *underfit*. Ultimately, one aspires to have a model that is both low bias and low variance; however, the bias-variance tradeoff dictates that decreasing model bias results in increasing variance, and decreasing variances increases model bias. In the context of linear regression, the model assumption is that the model predictand may be described

by a linear combination of the model predictors. As fewer predictors that are used, stronger assumptions are made on the predictive capability of the retained predictors, resulting in an increasingly biased model. In contrast, increasing the number of predictors increases the propensity to fit noise in the training data, resulting in higher variance solutions. For this reason, it is not desirable to simply use all available candidate predictors; doing so will often lead to a high variance, overfit model solution. The challenge is to select a subset of the candidate predictors that have a strong predictive relationship with the predictand, but is sufficiently small to avoid overfitting (Murphy 2012). This procedure is known as *feature selection*. MOS implements feature selection through a fairly simple scheme known as the forward stepwise implementation of *screening regression*. Screening regression is a greedy algorithm that operates by, beginning from the pool of candidate predictors, retaining the one with the highest correlation with the predictand. Next, the predictor from the remaining set of candidate predictors that, combined with the set of already-retained predictors, explains the largest proportion of predictand variance, is selected and retained. This procedure is repeated until a termination criterion is satisfied, typically either a fixed number of predictors have been selected, or until further predictor selection fails to explain a specified threshold of additional variance (Glahn and Lowry 1972).

MOS is run one, two or four times daily, for numerous atmospheric variables enumerated in part above, at lead times from 6 hours after initialization to approximately one week, with detailed guidance being available out to 84 hours past initialization. This procedure is used for approximately 1700 stations nationwide, with independent equations being used at each station. Model equations are retrained periodically, and different equations are often applied for different seasons. MOS output trained from the operational Global Forecast System (GFS) and a separate MOS trained from the North American Mesoscale (NAM) model are run and publicly disseminated by the National Weather Service (NWS). Operational MOS products do make quantitative precipitation forecast (QPF) predictions. However, rather than predict QPF directly, the MOS QPF predictand is discretized into categories, or

bins. Category 0 is defined as zero measurable precipitation, Category 1 one hundredth to less than one tenth of an inch, Category 2 from one tenth to less than one quarter of an inch, Category 3 from one quarter to less than one half of an inch, Category 4 from one half to less than one inch, Category 5 from one to two inches, and Category 6 refers to at least two inches of precipitation over the accumulation interval (Dallavalle and Cosgrove 2005; Gilbert et al. 2008). While this has use in ascertaining the general 'wetness' of the day, for a variety of reasons, it has limited use in many direct, quantitative applications of QPF. First, it has no regional awareness of the precipitation climatology in its formulation; 2" of rain may be exceedingly rare in some parts of the country, while relatively common in other area. Second, for extreme precipitation issues, having an unbounded threshold at 2" may not be high enough; it may be the case for some applications that 2" of rain will not yield any problems, but 6" of rain can cause a catastrophe. The MOS QPF formulation has no resolution under these circumstances. Third, even the bounded categories can present substantial problems in these sorts of situations. Suppose a user is interested in 24-hour accumulations, and MOS QPF variables are presented in 12-hour accumulations (as they are, in the short-range message). In the continuous QPF prediction context, this is not a problem: the user simply sums $Q12_1$ and $Q12_2$ to determine Q24. However, in the categorical reality, $Q12_1$ and $Q12_2$ may be "Category 5" and "Category 5", which could correspond to a Q24 anywhere between 2" and 4". Again, this difference may be very significant, with a 2" accumulation not requiring any preparative action, while a 4" accumulation does.

Despite its strengths, MOS does have several limitations which this research attempts to overcome. As noted above, MOS does not attempt to make numerical QPF predictions, limiting its utility in many forecasting applications. Is this simply a flaw in design? No, probably not; linear regression, while powerful, does have limitations. First, it is likely that the relationship between model predictors and continuous QPF predictand are complex, and the relationships may be non-linear. Second, linear regression is not especially well equipped for handling rare or extreme cases; it can be over-influenced

by outliers, leading to poor predictions throughout the QPF spectrum. These reasons, among others, likely motivated the decision to have QPF be a categorical, rather than continuous, predictand. Nevertheless, there is a need for continuous QPF prediction, and also a need for forecasts that can robustly handle extreme cases. Statistical systems research geared towards these forecast applications are explored in the following sections. MOS also does not directly provide any uncertainty information about its QPF values, instead only giving uncertainty by means of the pre-determined categorical ranges above. As will be discussed in sections to follow, probabilistic and uncertainty information can be of immense value to decision-making end users (Wilks 2011; Fritsch and Carbone 2004).

2.1.3 Other Precipitation Post-Processing

Many different methods and techniques exist for QPF post-processing. As alluded in the discussion of MOS above, even in the context of a single algorithm, there are often many different options, including but not limited to how define the QPF predictand (e.g. categorical vs. discrete), the selection of predictors, and the optimization and/or verifying function. Despite the vast array of yet unexplored algorithms and applications, many approaches have already been attempted and pursued. This subsection will discuss a few of the most significant developments, at least in the context of the research conducted herein. It should be noted that this should not be taken as a completely comprehensive literature review, as the full development of precipitation post-processing is too involved to discuss completely here.

A lot of early work on precipitation forecasting had little focus on calibration and post-processing. In the early years of operational global ensembles, more research effort focused on ensemble prediction system (EPS) design and comparing raw EPS derived forecasts, either via the ensemble mean or probabilities based on the proportion of ensemble members exceeding a threshold, to traditional deterministic products. One example of this is Buizza et al. (1999), which verified the

ECMWF ensemble for different seasons, thresholds, and resolutions. Later work (e.g. Clark et al. 2011; Mullen and Buizza 2011) looked at the effect of ensemble size and model resolution on raw ensemble PQPF skill. Other very early work in this area focused on the relationship between probability of precipitation (PoP) and QPF. Wilks (1990) considered categorical QPFs and conditional precipitation distributions, either conditioned on whether measurable precipitation occurred, or conditioned on the approximate subjective PoP issued for a given forecast. A discernable PoP-QPF relationship was identified over all sites examined, namely higher PoPs corresponded to a distribution of precipitation accumulations shifted towards higher amounts. This relationship was exploited to improve probabilistic QPF skill. The concept of using exploiting the PoP-QPF relationship for the purpose of predicting QPF has evolved and been refined over the years, such as in Bremnes (2004). In Schaffer et al. (2011), the relationship was applied in the opposite direction; QPFs were used to generate calibrated PoP forecasts. The authors again demonstrated the utility of this relationship for forecasting purposes, and further demonstrated the utility of using *neighborhoods*- forecasts values surrounding a given forecast point- to inform the local PoP forecast. The neighborhood concept was perhaps first demonstrated to be a computationally-efficient way to generate probabilistic information from a deterministic system in Theis et al. (2005). Here, neighborhoods were applied both spatially and temporally in generating probabilistic QPFs. This technique was also applied to a high-resolution ensemble in Schwartz et al. (2010), and robust result were seen here as well. Neighborhood-based approaches are seen as a method to increase forecast information without the added computational expense of an additional dynamical model run; the idea shows great promise, and will be further explored in detail in Chapter 5.

Hamill et al. (2004) was perhaps the first paper to discuss and advocate using model reforecasts- historical re-runs of a new model- to improve (statistical) forecasts for a variety of fields, including precipitation. Hamill and Whitaker (2006) demonstrated the value of reforecasts for QPF using an extremely coarse T62 version of the Global Forecast System (GFS) model. The authors used forecast

analogs, identifying historical cases that were deemed similar to the current forecast, to generate an ensemble of similar historical cases. The observations from these cases could then be used to predict and/or modify the current QPF, and the authors were able to do so in a way that significantly enhanced forecast skill at a variety of accumulation thresholds. Wilks and Hamill (2007) used the same dataset to compare three different methods- Logistic Regression, Nonhomogeneous Gaussian Regression, and Gaussian Ensemble Dressing- on their ability to make probabilistic forecasts for medium range precipitation. All methods showed some utility at different locations and lead times, and using a long 15-25 year reforecast dataset for training as opposed to a shorter 1-2 year training period significantly improved forecast skill for all algorithms tested, resulting in approximately a one-day improvement in forecast skill. This work was furthered with an Extended Logistic Regression implementation on ECMWF ensemble reforecast data in Roulin and Vannitsem (2012). Further reforecast work continued in Hamill et al. (2008) considered reforecasts both from the GFS used previously in addition to those from the European Center for Medium Range Forecasting (ECMWF). The authors examined multiple ways for generating forecast probabilities using these reforecasts. They first considered 'raw' ensemble probabilities (termed Point Democratic Voting in Chapter 5) and found these to be very unreliable and of minimal or negative skill for both forecast systems examined. Calibration with reforecasts was able to improve reliability substantially and yield forecasts with positive skill. The difference between skill comparing forecasts calibrated with long and short training samples was greatest at higher accumulation thresholds, suggesting the particular importance of increased training data for rare or extreme events. All this research, among others, demonstrating the value of reforecasts for improving forecast skill via post-processing led to the creation of a much higher resolution, albeit still very coarse, T254 11-member global ensemble based on the 2012 version of the Global Ensemble Forecast System (GEFS; Hamill et al. 2013). As will be discussed in more detail, this dataset will be used extensively throughout this research.

Other forecast calibration approaches have been explored as well. Applequist et al. (2002) was one of the earliest studies to comprehensively evaluate numerous methods for generating forecast probabilities (FPs) for precipitation from a deterministic model. While only examining fairly low accumulation thresholds, the study compared linear regression, logistic regression, neural networks, discriminant analysis, and a classifier system for generating FPs. Logistic Regression was found to be the best performing algorithm in this study, with traditional linear regression performing the poorest for the QPF problem. Raftery et al. (2005) introduced the Bayesian Model Averaging (BMA) technique to the field and applied it to the application of ensemble FP calibration; this approach was extended to PQPF in Sloughter et al. (2007). The same group of authors went on to compare other calibration methods (e.g. Gneiting et al. 2007). Yussouf and Stensrud (2008) demonstrated the utility of a simple 12-day running average binning technique for calibrating probabilistic QPFs for a multi-model ensemble.

Many other techniques and applications have been explored. The next section will discuss a subset of those which have targeted extreme event forecasting.

2.1.4 Rare Event Forecasting and Extreme Value Post-Processing[1]

Early work in this area that has been refined and improved in more recent years is the Extreme Forecast Index (EFI) developed at the European Center for Medium Range Forecasting (ECMWF). The EFI, first described in Lalaurette (2003), does not directly forecast any atmospheric field; instead, it attempts to quantify how the probabilistic forecast from an Ensemble Prediction System (EPS) compares with the model climate distribution for the prescribed atmospheric variable, location, and time. In so doing, the EFI acts as a qualitative forecasting tool in alerting forecasters to the potential for highly rare and anomalous occurrences. The EFI is used operationally at the ECMWF today.

[1] This sub-chapter makes extensive references to material presented in subsequent background chapters, particularly Section 2.4. Refer there for more information.

Friederichs and Hense (2007) developed censored quantile regression methodology to statistically downscale extreme precipitation over Germany. Regression is performed and performance assessed for events as rare as the 99th percentile. This is then applied in Friederichs and Hense (2008) towards generating PQPFs for 12-hour precipitation accumulations based on the operational GFS at that time. These concepts are further refined in Friederichs (2010), and later in Bentzien and Friederichs (2012), where the authors use a parametric mixture model approach to aid PQPFing over Germany using a high-resolution time-lagged ensemble. They apply fitted gamma, lognormal, and inverse Gaussian distributions and apply a generalized Pareto tail to aid in forecasting extreme precipitation amounts. This work affirmed previous findings that large amounts of data are needed for extreme precipitation quantiles, and that the behavior in the extremes being different than more typical deficiencies, with the base distributions performing acceptably until the extreme quantiles are reached, when the GPA tail is found to significantly improve model skill.

Various other techniques have been employed for forecasting extremes. Marsh et al. (2012), extending work from Sobash et al. (2011), uses historical model spatial error characteristics in a convection allowing model (CAM) to fit a kernel density function, which is then applied to calibrate probabilistic QPFs from a deterministic simulation. Roebber (2013) used evolutionary program (EP) techniques in a mildly-to-highly idealized frameworks and compared the EP-derived ensemble characteristics with those of a dynamical ensemble. He found that the EP ensemble performed better than the dynamical model ensembles at the extremes; specifically, he noted that EP forecasts had higher *resolution* (see Sections 2.3, 2.7). Williams et al. (2014) compared many of the techniques employed in papers discussed in Section 2.1.3- logistic regression, nonhomogenous Gaussian regression, BMA, and ensemble dressing- in a highly idealized framework and assessed their ability to appropriately perform bias correction in extreme cases. Most methods (except logistic regression) performed similarly well,

but great value was identified in allowing the bias correction to vary as a function of the predictand mean of an EPS. Many other related studies exist.

Recently published Scheuerer and Hamill (2015) is among the first studies to begin investigating the quantitative diagnosis of model QPF climatologies to enhance QPF forecasting. Scheuerer and Hamill use a complex algorithm on the Global Ensemble Forecast System Reforecast (GEFS/R) model, described in Section 2.2 below, to generate probabilistic forecasts over the contiguous United States (CONUS) for 1, 10, and 25 mm exceedance probabilities over 12 hours. They identify the unique challenges faced with QPF post-processing, which will also be conducted in the research to be presented herein: 1) QPFs have a unique probability distribution, with a positive probability of exactly zero precipitation and a continuous distribution of some flavor for positive amounts; 2) Forecast uncertainty is positively correlated with QPF magnitude; and 3) Infrequent occurrence of high precipitation amounts necessitates a vast amount of training data to appropriately handle these cases. They argue that the demand for training data associated with the third challenge is greatly alleviated using parametric, as opposed to non-parametric methods, provided the necessary assumptions made in the application of a parametric technique are sufficiently accurate. This distinction will be discussed in more detail in Section 2.4. Consequently, to combat challenges (1) and (3), they fit censored, shifted gamma distributions (CSGDs) to observed precipitation accumulations. Gamma distributions will be described in more detail in Section 2.4.3; these distributions are shifted such that the valid interval (see Table 2.2 below) may begin below 0, and are censored such that the probability of zero precipitation is the integral of the probability density function (see 2.3.3) from $-\infty$ to 0. A regression model is then fit to link local observed CSGD parameters to the ensemble statistics of the GEFS/R, and this is used to generate PQPFs. The authors found that this approach significantly outperformed the analog method discussed above. They also assessed the algorithm's sensitivity to training data length by testing 1-year and 3-year periods in addition to the primary 12-year dataset; it was found that shorter datasets where highly

prone to overfitting, resulting in significantly diminished forecast skill, reliability, and sharpness (see Sections 3.3 and 3.7), but the problem could be at least moderately improved by the inclusion of points exhibiting similar characteristics in the model training. The study did not, however, investigate truly extreme precipitation thresholds; the highest precipitation threshold examined was 25 mm over 12 hours, which is below the 1-year return period threshold for the same accumulation interval over most regions of the country. Since the algorithmic design and implementation is thought and argued to be appropriate for forecasting extremes, it is presented in this section of the discussion.

2.2 Model Information and History

Table 2.1: Dynamical models used or being planned for use in this research. Data availability, information about horizontal and vertical resolution, and the various parameterizations used in each model is included. Some ensemble systems are grouped into a single entry; for these, the number of members using a particular parameterization is included in parentheses or, if absent, applies to all members. Horizontal grid spacings with slashes denote a nested grid.

Model	Available From	Available To	Horizontal Grid Spacing	Vertical Levels	DA / (IC/BCs)	Cumulus	PBL	Microphysics	Land Surface	SW Rad	LW Rad
GEFS-RFCST	12/84	PRES	T254	42	GFS	SAS2	PM	Zhao, Carr (ZC)	Noah	RRTM	RRTM
GEFS-RT	3/14	PRES	T254	42	GFS	SAS2	PM	ZC	Noah	RRTM2	RRTM
GFS-OLD	6/09	1/15	T574	64	GFS*	ArSc	PM	Simple	Noah	RRTM	RRTM
GFS-NEW	1/15	PRES	T1534	64	GDAS	SAS2	EDMF	ZC	Noah	RRTM	McICA
NAM	6/09	PRES	12km	60	NDAS	BMJ	MYJ	Ferrier	Noah	CSRT	CSRT
NAM-NEST	5/14	PRES	4km	60	NAM	BMJ ***	MYJ	Ferrier	Noah	CSRT	CSRT
SREF	3/14	PRES	16km	35	NDAS, GFS,	BMJ (12), KF	MYJ (19),	Ferrier (17), GFS (2),	Noah	GFDL	GFDL

					RAP; BCs from GEFS	(5), SAS (4)	GFS (2)	WSM6 (2)			
HIRESW-ARW	5/14	PRES	4.2km	40	RAP/GFS	NA	YSU	WSM6	Noah	Dudhia	RRTM
HIRESW-NMMB	5/14	PRES	3.6km	40	RAP/GFS	NA	MYJ	Ferrier	Noah	RRTM	RRTM
NSSL-WRF	6/09	PRES	4km	35	NAM	NA	MYJ	WSM6	Noah	Dudhia	RRTM
CSU-MEM1	6/09	PRES	12/36km	36	GFS	KF	MYJ	WSM6	Noah	Dudhia	RRTM
CSU-MEM2	6/09	PRES	12/36km	36	GFS	G3	YSU	Thompson	Noah	Dudhia	RRTM
CSU-MEM3	6/09	PRES	12/36km	36	NAM	BMJ	MYJ	Goddard	Noah	Goddard	RRTM
CSU-MEM4	6/09	PRES	12/36km	36	NAM	KF	MYJ	Thompson	Noah	Dudhia	RRTM
HRRR	**	PRES	3km	50	RAP/HRRR	NA	MYNN	Thompson	RUC-Smirnova	Goddard	RRTM

*A major upgrade to GFS DA occurred in May 2012.
** Warm seasons 2012, 2013, 2014, and sporadically from 05/14-Present.
*** Operates in a very limited context.

A suite of different NWP models are used throughout this study in different capacities; the details of their use will be further described in subsequent sections. A summary of each model is depicted in Table 2.1. The models used are available through a combination of archived operational runs and reforecasts. Reforecasts have the advantage of model staticity which is often not upheld in operational settings, where models are often updated and revised to manually correct for particular perceived model errors and biases. Substantial model revisions change the bias characteristics and behaviors of

the model in different atmospheric scenarios, and it can therefore be difficult to use for accurate, calibrated model post-processing.

Many of the models appearing in Table 2.1 come from the NWS NWP suite. Their global model, the GFS, has evolved over decades of research and computing advancements, from the Global Spectral Model (GSM) beginning in 1980 to the Nested Grid Model (NGM) from 1987-2000, and later the Aviation (AVN)/Medium Range Forecast (MRF) model which morphed into the modern global system used today. The National Centers for Environmental Prediction (NCEP) also run a global ensemble based on the GFS, known as the Global Ensemble Forecast System (GEFS). The GEFS is a 21-member ensemble which, along with the GFS, is run four times daily out to 384 hours past initialization; ensemble members are perturbed only in their initial conditions (ICs), and not in their model physics. Both the GFS and GEFS undergo periodic changes to correct for observed biases and to improve bulk error characteristics. Major changes occur sporadically, roughly every two years. The most recent major upgrade to the GFS occurred in January 2015; many updates were made, but the most significant was a substantial increase in model resolution to T1534 from T574. Previously, a major upgrade to the model's data assimilation (DA) system occurred in May 2012. Both of these updates significantly impacted the performance and bias characteristics of the operational GFS; for this reason, the GFS has been separated into GFS-OLD and GFS-NEW in Table 2.1 to reflect the model specifications before and after the most significant 1/15 upgrade. The GEFS will undergo a similar upgrade in 2015 or 2016, but has not yet done so at the time of writing this manuscript. Additionally, a recent project to create a long, consistent record of model data used the February 2012 version of the GEFS to generate daily reforecasts for an eleven member ensemble beginning from December 1984 to present. This GEFS-derived dataset will be termed GEFS-RFCST, in order to distinguish it from the operational version, GEFS-RT. The GFS and its derivatives are *spectral* models, which have a different formulation than the *grid-point* formulation used in many models; the GFS and derivatives are the only spectral models used in this research. The GFS and

especially the GEFS are rather course; none of them are convection-allowing, meaning that they all require a cumulus parameterization. Other implementation details can be surmised from Table 2.1.

NCEP also runs numerous regional models, also frequently called limited area models (LAMs), which are used operationally by the NWS. North American regional modeling began with the Limited-area Fine Mesh (LFM) model in 1971, which was used until the implementation of the NGM in 1987. In 1993, the ETA model was implemented for regional modeling, and this improved over numerous upgrades and eventually became the North American Mesoscale (NAM) model used operationally since 2006. The NAM is also run four times daily, out to 84 hours past initialization, and has a horizontal grid spacing of 12 km, compared with the modern GFS's approximately 13km equivalent horizontal grid spacing (Rogers et al. 2009). This is still too coarse to begin to resolve convection, and requires a cumulus parameterization. However, a 4-km grid spacing one-way nested grid (NAM-NEST), which requires very little by means of a cumulus parameterization, has somewhat recently been embedded within the original North American NAM domain to cover the contiguous United States; data from this nest has been stored for this research since May 2014. NCEP also runs two high resolution runs twice daily out to 48 hours using two different dynamical cores, the Weather Research and Forecasting: Nonhydrostatic Mesoscale Model B (WRF-NMMB) core used in the operational NAM, and the WRF: Advanced Research WRF (WRF-ARW; Skamarock and Klemp 2008) core; these are called HIRESW-NMMB and HIRESW-ARW, respectively. Both of these are convection allowing grid point models; their parameterizations are summarized in Table 2.1.

Lastly, there are two other operational products at NCEP being used for this study. The first is a short-term, high-resolution- 3 km horizontal grid spacing- system, the High Resolution Rapid Refresh (HRRR) that is run hourly out to 15 hours, with experimental runs out to 24 hours. This system developed from the NAM-based Rapid Update Cycle (RUC) model implemented in 2005 (Benjamin et al.

2004), and the Rapid Refresh (RR) that replaced it in 2012. Due to its limited forecast duration, it cannot be used for forecasting 24-hour accumulation events, but still has great utility in forecasting near-term, short-duration events. The HRRR was operationally implemented recently, in September 2014, and as such, the HRRR has received numerous major upgrades and changes during its development over the last few months and years (Waxberg 2015). The last NCEP system used here is the Short Range Ensemble Forecast (SREF) system, run four times daily out to 87 hours. With a 16km horizontal grid spacing, the SREF is notably higher resolution than the current GEFS, but still too coarse to allow for explicit convection. Like the GEFS, the SREF consists of 21 members; however, the SREF is perturbed not only in ICs, but has three different DCs- NMM, NMMB, and ARW- with seven ensemble members for each DC. Each of those seven member groups has different ICs and different MPs. Thus, although summarized as one system in Table 2.1, the SREF is in reality 21 different model runs with different numerics, physics, and initial conditions.

Lastly, some NWP models used in this study are not operational, and are run in-house at universities or research institutions. These models often, though not always, have the advantage of more stacitiy for statistical analysis than the operational models; when the models are left undisturbed, their bias characteristics remain the same from season-to-season and year-to-year, allowing for a longer and more robust model dataset to analyze. For example, the National Severe Storms Laboratory (NSSL) runs a convection-allowing 4km WRF-ARW model once daily out to 36 hours (NSSL-WRF). Since June 2009, this model has been run with very limited alterations, with the only one of note being an upgrade in WRF version from 3.1.1 to 3.4.1 in April 2013. The last set of models, the Colorado State University (CSU) 12km WRF ensemble, has been run once daily with the same configuration for its first four members since February 2012. Those members have been or are being reforecasted back to June 2009 to match the data record length of the NSSL-WRF.

2.3 Probabilistic Forecasting and Ensemble Prediction System Fundamentals

2.3.1 Motivation: Uncertainty and Predictability

There are a vast number of sources that yield uncertainty in a dynamical model simulation and ultimately lead to forecast error. Error in a model's analysis, or starting point, is one major source of uncertainty, and many different factors contribute to inaccuracies in assessment of the current atmospheric state. First, it is important to realize that a perfect, error-free analysis would accurately place the position and movement of every particle in the atmosphere. Recognizing this, a perfect analysis of the atmospheric state is, at present, woefully unrealistic, as we do not have atmospheric observations on the particulate scale; even in the immediate vicinity of an observing station, uncertainty is introduced solely from observation resolution. Further, even somehow attaining an accurate record of every particle's position, velocity, etc. in the Earth-atmosphere system (EAS), current dynamical models do not keep track of every particle within the model either, so the information would undoubtedly be accordingly coarsened during the process of model initialization, and this finite model resolution would yield model analysis uncertainty as well. It must be further noted that observation instruments are not perfect either, and errors in their measurements, even where we do have them, introduce uncertainty as well. Other error in the data assimilation process of translating the true atmospheric state to the model atmosphere presents a further source of analysis uncertainty (Kalnay 2003).

Initial conditions (ICs) are one major source of uncertainty, but boundary conditions (BCs) can yield erred representations of the atmospheric state as well. In limited area models (LAMs), the error associated with lateral boundary conditions can introduce error to the model solution even with perfect ICs. However, in all models, top and bottom boundaries exist that can present issues. On the bottom level, improperly resolved topography, incorrect soil moisture measurements, and incorrect

representations of the surface characteristics are all possible source of uncertainty. Improper, artificial upper boundaries in the model present an additional source, and any interaction of the EAS with space is likely to be improperly handled as well (Kalnay 2003).

Even with a perfect analysis of the initial atmospheric state that manages to be represented without error or simplification, an imperfect model will still quickly introduce small errors to the model's projection of the atmospheric state, and from there, non-linear error growth due to chaos will continue to increase the departure of the model solution from reality (Lorenz 1963). Firstly, even neglecting issues with model physics, there are numerous direct problems associated with model numerics. Floating point operations on modern computers have finite precision, and this can lead to error both in how the numbers are stored and in their finite size leading to truncation error. In fact, explicit truncation error was what first led to the discovery of non-linear error growth in modeling the atmosphere. Further, many equations governing the atmosphere involve derivatives, integrals, and other mathematical constructs which may only be approximated by using nearby values in the context of various numerical approximation schemes; these too, introduce error. The finite resolution of numerical models also prevents accurate representation and simulation of many small-scale physical processes. Additionally, problems with model physics compound the problem of model error. Not all atmospheric processes are yet fully understood, and thus not completely accurately represented in numerical models. Additional assumptions are often necessary, which each add a source of uncertainty, and any physics parameterizations inherently are not a 'pure' representation of the simulated process and contribute further error as well (Kalnay 2003).

Despite these sources of uncertainty and non-linear error growth, dynamical systems, the atmosphere included, do not evolve randomly or unboundedly throughout parameter space. Physical laws and relationships between the model variables prevent certain theoretical values, or combination

of values, from being obtained. For example, though valid temperatures, it is exceedingly unlikely- nearly impossible even- that surface temperatures anywhere on Earth will be less than 100K or higher than 400K anytime in the foreseeable future. Similarly, while surface temperatures of 250K and 300K are both observed on Earth with moderate frequency, it is virtually impossible that two adjacent locations only a kilometer or two apart and at similar elevations will simultaneous possess these two values. Instead, absent a major change in external forcing to the system, the atmosphere will evolve in only a small region, or subspace, of the total phase space; this is known as the system's *attractor*. Forecast error E_f can be thought of quantitatively as the distance in phase space between the forecast and true atmospheric state for any given lead time τ. *Error growth* can be similarly characterized at lead time τ as $\frac{dE_f}{dt}(\tau)$. Further, the *intrinsic limit of predictability* may be defined as the critical lead time τ^* satisfying: $\tau^* = \min\left(\tau \in \left\{E\left[E_f(\tau)\right] > E[E_{clim}]\right\}\right)$, the first time that the expected forecast error for robust, unbiased forecasts exceeds the expected error from using climatology as a forecast. Lastly, *error saturation* is defined here to occur at lead time $\tau^\dagger$ when $\frac{dE[E_f]}{dt}\left(\tau^\dagger\right) \approx 0$, and the corresponding saturation value is denoted $E_f^\dagger$ (Eckel 2003).

For these reasons, despite the theoretical capability of a perfect forecast of the state of a deterministic, chaotic system for all time just through a single dynamical simulation, in practice, this, or anything close to it, will not occur in the envisionable future with respect to the atmosphere. Deterministic forecasts will continue to begin with errors that will grow in expectation with lead time until it saturates to twice the mean squared error of the climatological mean. The chaotic nature of the atmosphere sharply limits the utility of directly using deterministic dynamical model output for the purpose of forecasting. Given this, the question is: can we do better, and, if so, how much better can we do? There is no tractable 'cure' for the problems inherent with forecasting a chaotic system. Recognizing the limitations, there are, however, effective mechanisms to cope with the errors and

uncertainty as best as possible. Specifically, in addition to minimize forecast error, it should be noted that there is great utility in accurately quantifying expected forecast error, or forecast uncertainty, and attempt to accomplish both forecast error minimization and uncertainty quantification simultaneously. This can be effectively accomplished by using a combination of different, but realistic, model ICs, possibly in addition to varied, plausible model physics for approximating how the true atmosphere behaves. Repeating this many times yields many different plausible dynamical model forecasts; this collection of model forecasts is termed an *ensemble*, and the process is known as *ensemble forecasting* (Leith 1974). Ensemble forecasting goals and design considerations are discussed in the following sections.

One simple motivation for use of ensemble forecasting comes from the consideration of predictability limits and error saturation. The expected climatological error E_{clim} for a single forecast can be expressed in a mean square error sense over a long record of T observations: $E[E_{clim}] = \overline{E_{clim}} = \frac{1}{T}\sum_{j=1}^{T}(\mu_j - o_j)^2$, where μ denotes the climatological mean and o_j denotes the observation at time j. This can be readily converted to the framework of forecast anomalies *a*- departures from the climatological mean μ- as: $E[E_{clim}] = \frac{1}{T}\sum_{j=1}^{T}((\mu_j - \mu_j) - (o_j - \mu_j))^2 = \frac{1}{T}\sum_{j=1}^{T}(0 - a_j)^2 = \overline{a^2} \equiv E_f^*$. A forecast system in this context is said to be *unbiased* if its forecasts yield climatology-relative forecast anomalies $\hat{a}$ that have a long term expected value of zero. The expected forecast mean square error in the context of anomalies can then be expressed for an unbiased deterministic forecast as: $E[E_{f_{det}}](\tau_{det}{}^{\dagger}) = \frac{1}{T}\sum_{j=1}^{T}(\hat{a_j} - a_j)^2 = \frac{1}{T}\sum_{j=1}^{T}(\hat{a_j} - a_j)(\hat{a_j} - a_j) = \frac{1}{T}\sum_{j=1}^{T}\hat{a_j}{}^2 + a_j{}^2 - 2\hat{a_j}a_j$. Because these forecasts are assumed to be unbiased, $\hat{a_j}$ and a_j are independent and have long term mean values of zero, leading the last covariance term to vanish. Again, the forecasts being unbiased means the variance characteristics are identical over a large number of samples, meaning $E[\hat{a_j}] = E[a_j]$. Thus, we find that the expected mean squared error of an unbiased deterministic forecast at the deterministic

forecast lead time of error saturation ${\tau_{det}}^{\dagger}$ may be expressed by: $E^{\dagger}_{f_{det}} \equiv E[E_{f_{det}}]({\tau_{det}}^{\dagger}) = \frac{1}{T}\sum_{j=1}^{T} \widehat{a_j}^2 + {a_j}^2 = \frac{1}{T}\sum_{j=1}^{T} 2{a_j}^2 = 2\overline{a^2} = 2E_f^*$. For a single deterministic forecast, error saturation is thus found to occur at twice the mean climatological error, or twice the limit of predictability. However, consider instead an ensemble of *n* unbiased deterministic forecasts. The ensemble mean climatology-relative forecast anomaly can be denoted $\hat{\bar{a}}$, and the mean squared error at saturation $E[E_{f_{ens}}]({\tau_{ens}}^{\dagger}) = \frac{1}{T}\sum_{j=1}^{T}(\hat{\bar{a}}_j - a_j)^2 = \frac{1}{T}\sum_{j=1}^{T} \hat{\bar{a}}_j^2 + {a_j}^2 - 2\hat{\bar{a}}_j a_j = \frac{1}{T}\sum_{j=1}^{T} \hat{\bar{a}}_j^2 + {a_j}^2$. Since the last term is again zero by the same arguments above. $\hat{\bar{a}}_j^2$ behaves as:

$\hat{\bar{a}}_j^2 = \frac{1}{n}\sum_{k=1}^{n} a_k \frac{1}{n}\sum_{k=1}^{n} a_k = \frac{1}{n^2}\sum_{k=1}^{n} a_k \sum_{k=1}^{n} a_k = \frac{n}{n^2}\overline{a^2} = \frac{1}{n}\overline{a^2}$, we can write:

$E^{\dagger}_{f_{ens}} \equiv E[E_{f_{ens}}]({\tau_{ens}}^{\dagger}) = \frac{1}{T}\sum_{j=1}^{T} \hat{\bar{a}}_j^2 + {a_j}^2 = \left(1 + \frac{1}{n}\right)\overline{a^2} = \left(1 + \frac{1}{n}\right)E_f^*$. From this, it is readily seen by inspection that error saturation for an ensemble of forecasts (n > 1) occurs with less error than a deterministic forecast. Noting this, and further noting that both deterministic and ensemble based forecasts begin at lead time zero with no baseline error (aside from analysis error), given that expected forecast error increases smoothly and monotonically, it follows that there exists some window of lead times τ where the expected ensemble forecast error is less than both an expected deterministic forecast error and climatology: $E[E_{f_{ens}}](\tau) < E[E_{f_{det}}](\tau)\ and\ E[E_{f_{ens}}](\tau) < E_f^*$. The set of lead times satisfying this condition is referred to as the *ensemble window of utility*, and this is where ensemble forecasting is of particular value. It is posited that, for the extreme precipitation forecasting problem examined in this work, the set of lead times examined herein fall within the ensemble window of utility (Eckel 2003).

2.3.2 Goals of Ensemble Prediction

Qualitatively, there are numerous desired outcomes from the use of ensemble forecasting. A deterministic forecast gives one sense of how the atmosphere may evolve from present; an ensemble aims to give an accurate assessment of the range of possible future evolutions of the atmosphere. Some scenarios have lower sensitivity and are thus more predictable than other situations; the use of

ensembles gives a sense of the flow-dependent error growth, or predictability-of-the-day. Use of only deterministic forecasting, in contrast, yields only information on long-term average predictability, with no information specific to the uncertainty associated with the current forecast. One of the objectives, then, of ensemble forecasting is forecast-specific *uncertainty quantification*. Additionally, as illustrated in section 2.3.1 above, the ensemble mean or consensus forecast can be shown to, on average, have lower forecast error than the use of a single deterministic forecast. While not a major objective of ensemble forecasting, this result is a beneficial side-effect. Ultimately, the chief objective of an ensemble prediction system (EPS), or more generally, a probabilistic forecast system (PFS), is to create the *sharpest* possible output forecast (OF) PDF, while still maintaining forecast *reliability* and *statistical consistency*. As a further requirement, the PFS PDF output must be accessible to end users in a useful and understandable format. Without this, despite having very high theoretical utility, the practical utility of the EPS will be relatively low (Wilks 2011; Eckel 2003).

Statistical consistency requires that the OF PDF corresponds to the true forecast (TF) PDF. The TF PDF is *not* the PDF of the atmosphere at a known future time given the current atmospheric state; assuming that the atmospheric system is deterministic (which we are), then this function would always be a delta function with infinite probability density at the verifying state and zero probability density elsewhere. Let this PDF be denoted the verifying PDF (VPDF). Rather, the TF PDF is a function of both the analysis of the current atmospheric state and the model itself. If the analysis and model are both, in fact, perfect, then the TF PDF is the VPDF. However, both analysis error and model error and the associated uncertainty that each source of error introduces act to both shift and broaden the TF PDF. An ideal ensemble would have infinitely many ensemble members, with appropriate perturbations to accurately capture *all* sources of uncertainty, and would employ either a perfect dynamical model, or if the dynamical model has error, would employ perfect post-processing to appropriate re-map ensemble member atmospheric states to true atmospheric states. In reality, none of this is possible; finite

computing resources limits the number of ensemble members, not all sources of uncertainty are fully understood or accurately quantified, and for reasons explained in 2.3.1, dynamical models are far from perfect. More formally, statistical consistency requires that the mean squared error (MSE) of an ensemble mean equal the mean ensemble member variance:

$$MSE_{\bar{x}} = \overline{\sigma_x}$$

$$\frac{N}{N+1}\frac{1}{D}\sum_{d=1}^{D}\left(\frac{1}{N}\sum_{n=1}^{N}x_{nd} - o_d\right)^2 = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{N-1}\sum_{i=1}^{N}\left(x_{di} - \frac{1}{N}\sum_{n=1}^{N}x_{nd}\right)^2$$

Forecast *reliability* is a related metric that is also a necessary but insufficient condition for forecast skill. Reliability refers to the correspondence between forecast probability (FP) and observed relative frequency (ORF). Ideally, probabilistic forecasts should be reliable: when an 80% probability of event occurrence is forecast, it is desirable that the event actually occur 80% of the time. However, this is not enough. Having a forecast PDF be the climatological PDF for every forecast means forecasting the climatological frequency of event occurrence $\bar{o}$ for any possible event; this is by definition reliable (the ORF over all forecasts is by definition the climatological frequency of occurrence), but has no utility to any end users, as it presents them with no new information. Forecasts must also exhibit some *sharpness*- ability to forecast towards extremes, away from $\bar{o}$. The combination of sharp and reliable forecasts leads to high *resolution* forecasts- those which distinguish events from non-events by forecasting relatively higher FPs when events occur (Wilks 2011).

2.3.3 Probability Density Functions, Cumulative Density Functions, and Quantile Functions

In probability theory, there are three primary ways of expressing the distribution of values that a continuous random variable may take: a probability density function (PDF), cumulative density function (CDF) and quantile function (QF). For a discrete random variable, the corresponding distributions are termed mass functions: probability mass functions (PMFs) and cumulative mass functions (CMFs). The

axioms of probability specify that the total probability of a random variable possessing some value be unity, and further, the probability of an event outcome must be non-negative. By definition, a continuous variable is one of an infinite number of possible values; thus the probability of a continuous random variable having any given value is zero (if the probabilities were non-zero the total probability being unity could not be satisfied). Thus, when discussing the probability of a continuous random variable's value, one must frame the discussion in the context of the variable taking on one of infinitely many values within a range. This motivates the use of *PDFs*, which will typically be denoted *f* throughout unless otherwise specified. A PDF satisfies the following properties: 1) $\forall x\ f(x) \geq 0$; 2) $\int_{-\infty}^{\infty} f(x)dx = 1$; 3) $P(a \leq x \leq b) = \int_{a}^{b} f(x)dx$. Higher *probability density* at a value x indicates higher probability for the variable possessing a value *near* x; but $P(x = a) \neq f(a)$. The PDF framework is often helpful in quantifying rarity of double-bounded events, e.g. between 1 and 2 inches of precipitation in the context of the QPF problem. However, often problems are framed in the context of exceedance thresholds, or single bounded events; in this framework, it is often more desirable to examine the probability distribution in a cumulative framework by means of a CDF. A CDF *F* is defined as a function of the corresponding PDF: $F(x) = \int_{-\infty}^{x} f(u)du$. F(x) then corresponds to $P(U \leq x)$, the probability that the observed value will be less than the input argument. Lastly, in some applications, including many relevant to this study, it is more value to look at the inverse CDF, or QF, x(F). In the context of precipitation accumulation, a CDF answers the question "how rare is it to experience a precipitation amount _?", while the QF answers the question "how much precipitation accumulation is required to attain an event of rarity _?" Since it is often more useful to have the rarity fixed than the threshold, the QF presents advantages over use of the CDF in many instances (Wilks 2011).

2.3.4 Ensemble Configuration

There are many factors to consider in the configuration of an EPS. It is essential that, before making any ensemble configuration decisions, the EPS objectives and resources are first specified. What are the forecast applications? Who are the end users, and what information do they most care about? How powerful of computing resources are available? The optimal EPS configuration may be very different depending on the answer to these sorts of questions. Perhaps the most obvious ensemble configuration parameter is the ensemble size *n*. Required computing power scales linearly with the ensemble size. If available computing resources are effectively infinite, a very large number of ensemble members is, of course, desirable. However, in a limited-resource environment, the desire for a large ensemble must be balanced with the quality of the individual members, quality of any initial condition perturbations, time for ensemble post-processing and calibration, and other computational tasks. A small number of members, perhaps 3-12, is often enough to yield a reasonable ensemble mean for fairly common or routine events, and also give a qualitative sense of relative forecast uncertainty. A larger ensemble size, say 20-30 members, is often necessary to use the ensemble to generate sharp and reliable forecast PDFs. For rare events, many more members are needed to adequately sample the true forecast PDF's tail, with perhaps 50-100 members desired for skilled probabilistic forecasts, perhaps even more for extremely rare events (Eckel 2003). At the same time, individual member forecast skill may be substantially degraded if, to compensate for the increase in ensemble size, ensemble members are coarsened. In a very approximate sense, model run time can be thought to scale inversely proportional to the cube of the model grid spacing *g*. Going from a deterministic run to a 100-member ensemble, for example, would require coarsening the grid spacing by a factor of approximately 4, perhaps slightly more, to conserve use of the computing resource C. Processes that are barely resolved in the deterministic run then, will not be resolved at all in any of the ensemble members, instead likely depending on parameterization of the phenomenon of interest, likely resulting in both worse individual

forecasts and substantially worse sampling of candidate true atmospheric solutions. This tradeoff must always be carefully considered in ensemble design. Another major consideration is how to best generate member perturbations so that all member forecasts are still realistic candidate forecast solutions while still generating sufficient ensemble spread to avoid an overconfident EPS with associated overconfident forecasts. Spread can be achieved through IC perturbations, model physics (MP) perturbations, dynamical core (DC) changes, and other less common alterations such as model terrain and model resolution. It is often desirable for the purpose of ensemble statistics to have each ensemble member be equally likely to verify as truth; changes to MP and DC have the disadvantage of frequently not satisfying this property, giving credence to the use of IC-perturbed ensembles. However, research has found that IC-perturbed ensembles tend to result in an ensemble that is unrealistically underspread due to inability to capture all of the true sources of uncertainty, requiring either unrealistically large IC-perturbations or an attempt to broaden the forecast PDF artificially via post-processing. A combination of these perturbations tends to result in a more realistic forecast PDF, at the cost of complicating the generation and interpretation of the ensemble output (Kalnay 2003, Wilks 2011). Again, these considerations must be handled carefully, and final choices should be optimally tailored to the end users and their relevant forecast applications which will make use of the ensemble information. The research in this study is concerned with extreme precipitation occurring on a variety of scales from the meso-γ to synoptic. For this reason, it is considered desirable to use as large of an ensemble as possible, with less emphasis on the associated complication to the ensemble statistics.

2.4 Extreme Value Theory

Extreme Value Theory (EVT) can be a bit misleading in its name; this body of statistical theory does not strictly concern the modeling of very rare events, but describes the behavior of extremes from groups- the distribution of block maxima or minima. Directly modeling the distribution of right-skewed phenomena, such as daily precipitation, is also of great importance for applications in meteorology,

hydrology, and other fields. Distributions explored in this thesis will all be characterized as Right-Skewed Distributions (RSDs), but only a subset of such distributions applies directly to EVT.

In EVT, the ultimate goal for this application is to obtain accurate QF estimates for large return periods (RPs); that is, for a given RP, obtain an accurate estimate for the precipitation amount corresponding to that frequency of occurrence. In many cases, including in the research discussed herein, the RPs of interest extend well beyond the length of the data record. There exist many approaches to estimate event probabilities and QFs from a data record; the approaches can be classified into either *parametric* or *non-parametric* techniques. Parametric techniques make more assumptions than their non-parametric counterparts; principally, at least when model fitting on raw data, they assume that the input data comes from a known underlying probability distribution, and seek to use the data to optimize estimates for the parameters of underlying probability distribution. As will be seen below, EVT exercises the advantage of not needing to know the underlying probability distribution; raw data is manipulated in such a way that, provided that certain conditions are satisfied, the probability distribution of the manipulated series is known, regardless of the underlying initial distribution. Non-parametric methods, in contrast, do not assume that the training data comes from a known underlying probability distribution, and the number and values of all model parameters are determined dynamically based on the training data. Both classes of approaches have advantages and disadvantages. In general, due to making weaker assumptions, non-parametric techniques are often considered more robust, and several components of the greater forecast model applied in this study are non-parametric algorithms. However, non-parametric approaches tend to extrapolate poorly to data much rarer than what is seen; estimates for events rarer than the data record length tend to be quite poor (Murphy 2012; Scheuerer and Hamill 2015). For this reason, generating estimates based on an assumed underlying probability distribution is considered necessary for this component of the forecast pipeline.

2.4.1 Fisher-Tippett-Gnedenko Theorem

Much of the foundation of EVT makes use of the Fisher-Tippett-Gnedenko (FTG) Theorem. The theorem will be derived briefly here, and summarized below. Let there be a set of *n* independent and identically distributed random variables (IIDRV) $\{X_1,\ldots,X_n\}$; that is, data values which are *independent* of each other, and sampled from the same probability distribution. Let each IIDRV have CDF F: $F(x) = P(X_k < x)$. Further, let $M_n = \max\{X_1, X_2, \ldots, X_n\}$. $P(M_n < x) = P(X_1 < x \;\&\; X_2 < x \;\&\; \ldots \;\&\; X_n < x)$. Thus, by assuming independence, it can be shown that $M_n \sim F^n(x)$, where the ~ notation is used here to mean "is distributed as".

Let x* be the set of values for x satisfying F(x) < 1: $x^* = \sup\{x: F(x) < 1\}$

It can be readily shown: $\lim_{n\to\infty} P(M_n \le x) = \begin{cases} 0 & x < x^* \\ 1 & x > x^* \end{cases}$. This, taken in the limit of large n, is termed an *asymptotic distribution*; and because it is single-valued, it is said to be *degenerate*.

To avoid a degenerate asymptotic distribution, random variable M_n can be normalized using coefficients a_n and b_n, subject to $a_n > 0$, as: $Y_n = \frac{M_n - b_n}{a_n}$. Given how M_n is distributed, shown above, it can be readily shown that $Y_n \sim F^n(a_n x + b_n)$. Suppose there exist a series of coefficients for a_n and b_n such that the asymptotic distribution of Y_n is *non-degenerate* distribution function G(x). Then:

$$\lim_{n\to\infty} F^n(a_n x + b_n) = G$$

$\lim_{n\to\infty} n \log F(a_n x + b_n) = \log G$, taking the logarithm of both sides

$\lim_{n\to\infty} \frac{1}{n(1-F(a_n x+b_n))} = \frac{-1}{\log G}$, noting $\log x \approx x - 1 \; for \; x \approx 1$

The full FTG proof is too involved to show here, but it can be shown that the statement above is equivalent to saying that:

$\lim_{n\to\infty} \frac{U^{\leftarrow}(nx)-b_n}{a_n} = G^{\leftarrow}\left(e^{-1/x}\right)$, where $U \equiv \frac{1}{1-F}$, and $\leftarrow$ denotes the *left-continuous inverse* of a function, defined as: $f^{\leftarrow}(x) \equiv \inf\{y: f(y) > x\}$, where inf denotes the *infimum* of a set.

We can continuize this and define a function D:

$$D(x) \equiv \lim_{t\to\infty} \frac{U^{\leftarrow}(tx) - b_{\lfloor t \rfloor}}{a_{\lfloor t \rfloor}}$$

Assuming without loss of generality that D is continuous at 1, we can further define a function E:

$$E(x) \equiv \lim_{t\to\infty} \frac{U^{\leftarrow}(tx) - U^{\leftarrow}(t)}{a_{\lfloor t \rfloor}} = D(x) - D(1)$$

It can be shown that, because a non-degenerate solution is mandated:

$$E(xy) = E(x)a_{\lfloor y \rfloor} + E(y)$$

Define two new functions:

$$H(x) \equiv E(e^x); Q(x) \equiv \frac{H(x)}{H'(0)}$$

$$E(xy) = H(\log x + \log y) = H(\log x)a_{\lfloor e^y \rfloor} + E(\log y)$$

Using the definition of E, $E(1) = 0 = E(e^0) = H(0)$

Subtracting $H(\log y)$ from both sides and dividing by log(x) yields:

$$\frac{H(\log x + \log y) - H(\log y)}{\log x} = \frac{H(\log x) - H(0)}{\log x} a_{\lfloor y \rfloor}$$

Differentiating: $H'(\log\ y) = H'(0)a_{\lfloor y \rfloor}$

From this, by inspection: $Q(0) = 0; Q'(0) = 1$

Further manipulation gives rise to the following differential equation for Q:

$Q'(z) - 1 = Q(z)Q''(0)$, subject to $Q(0) = 0; Q'(0) = 1$

Denote $\xi \equiv Q''(0)$. Solving yields $Q'(z) = e^{\xi z}$

Re-writing in terms of D: $D(z) = D(1) + H'(0)\frac{z^{\xi}-1}{\xi}$

Taking the left-continuous inverse: $D^{\leftarrow}(z) = \left(1 + \xi\frac{z-D(1)}{H\prime(0)}\right)^{1/\xi}$

By the definition of D and the definition of left-continuous inverse:

$$G(z) = e^{-1/D^{\leftarrow}(z)} = e^{-1\Big/\left(1+\xi\frac{z-D(1)}{H\prime(0)}\right)^{1/\xi}}$$

This can be expressed as:

$$G(H'(0)x + D(1)) = e^{-1\Big/\left(1+\xi\frac{H'(0)x+D(1)-D(1)}{H\prime(0)}\right)^{1/\xi}} = e^{-1/(1+\xi x)^{1/\xi}}$$

Or, more generally:

$$G(ax + b) = e^{-(1+\xi x)^{-1/\xi}}$$

This is the FTG Theorem. More plainly, FTG states that block maxima of renormalized IIDRVs, regardless of the underlying distribution of the individual block elements, converge in distribution to one of the Gumbel, Fréchet, or Weibull families of probability distributions, depending on the value of ξ in the above derivation. A positive ξ indicates belonging to the Fréchet family, negative values indicate membership in the Weibull family, and ξ=0 implies a Gumbel distribution. More succinctly, FTG states that block maxima are distributed as the Generalized Extreme Value distribution, $M_n \sim GEV(\mu, \sigma, \xi)$ (De Haan and Ferreira 2007).

2.4.2 Approaches to Application

2.4.2.1 Annual Maximum Series

This FTG theorem is integral to EVT, including the application to extreme precipitation, streamflow, and flooding. In meteorology and hydrology, the most frequent application is to construct an *Annual Maximum Series* (AMS), with each element being the block maximum of daily precipitation (or streamflow, etc.) over each year in the data record. By FTG, this series will follow a GEV distribution (i.e. a GEV distribution can be accurately fit to these maxima), and the GEV can be employed to derive a relationship between *Annual Exceedance Probability* (AEP) and precipitation threshold by means of the QF. AEPs are readily converted to quantiles and *Average Recurrence Intervals* (ARIs), via the relation: $Quant = 1 - AEP = 1 - \frac{1}{ARI}$. So the 2-year ARI corresponds, for example, to a 50% AEP and the median of the distribution. It should be noted that the ARI is distinct from the RP in that in this framework, the ARI interval is discretized into years- the period does not get shortened for having multiple exceedances in the same year. This is a shortcoming of the AMS framework, at least when the year-independent RP framework is desired. However, a relation does exist relating AMS-based exceedance probabilities and those derived from the year-independent framework described below: $RP = \frac{1}{EP_{PDS}} = \frac{1}{1-e^{-AEP_{AMS}}}$. This relation can be used to relate AMS-derived estimates into the desired framework (e.g. De Haan and Ferreira 2007; Bonnin et al. 2004).

2.4.2.2 Partial Duration Series

An alternative approach to AMS used in EVT is the *Partial Duration Series* (PDS) or *Peaks-over-Threshold* (POT) approach. Instead of simply constructing a time series of annual maxima and fitting a distribution to those values, in PDS, all independent values, or peaks, exceeding a particular specified threshold are extracted from the data record to form a time series, and an RSD is fit to this time series. Recall from the FTG proof that EVT requires that the original random variables be independent. This is a

problem in time series applications, as one day's precipitation, for example, is highly correlated with the surrounding days. This is why, when extracting values to form a derived time series, it is important to take values that are sufficiently temporally separated so as to be considered independent. The other important variable in PDS analysis is the choice of threshold. One popular choice that will be explored in this study is the minimum value of the AMS derived from the same data record.

As FTG establishes, $P(M_n < x) \sim GEV(x; \mu, \sigma, \xi)$. But in this case, we are interested in individual base random variables $X_k \sim F(x)$ exceeding some threshold Θ. The exceedances can be expressed as $y = x - \theta$, and a CDF E of exceedances may be constructed:

$$E_\theta(y) \equiv P(X < \theta + y | X > \theta) = \frac{F(\theta + y) - F(\theta)}{1 - F(\theta)}$$

Though the proof will not be shown here, it can be readily shown that, for sufficiently large Θ, the asymptotic distribution of E_Θ is:

$E_\theta(y) \sim GPA(y;\ \mu, \sigma, \xi)$, where GPA denotes the Generalized Pareto Distribution.

The general procedure in PDS after selecting a threshold Θ and appropriately extracting an independent series S from the complete daily series, a GPA can be fit to the exceedances to yield a conditional distribution: $P(X > \theta + y | X > \theta) = 1 - GPA(y; \mu, \sigma, \xi)$.

The law of total probability may then be applied to back out the unconditional distribution:

$$\begin{aligned} P(X > \theta + y) &= P(X > \theta + y | X > \theta) P(X > \theta) + P(X > \theta + y | X < \theta) P(X < \theta) \\ &\approx \left(1 - GPA(y; \mu, \sigma, \xi)\right) \frac{len(PDS)}{len(Data)} + 0 * \frac{1 - len(PDS)}{len(Data)} \\ &= \left(1 - GPA(y; \mu, \sigma, \xi)\right) \frac{len(PDS)}{len(Data)} \end{aligned}$$

RP thresholds can then be derived using the QF of the unconditional distribution (De Haan and Ferreira 2007).

2.4.2.3 Direct Fits (DF)

The third approach, not directly an application of EVT, is, instead of using threshold exceedances or block maxima, to simply attempt to guess the underlying distribution of the base random variables X_k, which in this study is model QPF values over accumulation interval T. Here, we can't rely too heavily on theory, since, abstractly, if nothing is known about the data record, it can't be expected *a priori* to follow any particular distribution. However, the probability distribution of accumulated precipitation is and has been of great interest to the scientific community for a long time. Approaches have been employed parametrically fitting various probability distributions to precipitation observation records, both Full Wet Series (FWS), which includes only days with measurable observed precipitation, and Full Dry Series (FDS), which includes the entire record, including days without measurable precipitation. This approach is quite simple to apply; simply take the FDS, filter it as necessary in the case of an FWS, and then use the FDS or FWS to estimate appropriate parameter values for an RSD of choice. Unlike the previous approaches, this gives daily exceedance probabilities, so the RPs are given by: $RP = \frac{1}{365.25 * Quant}$, so the quantile for a 100-year precipitation event, for example, is 0.999973. The challenge here is choosing the correct underlying distribution, and coming up with accurate parameter estimation. The best approach with the former is to create a list of candidate distributions and test fits on all of them. Parameter estimation will be discussed more in section 2.4.4. It should be noted that very little, if any, work has been done assessing the underlying probability distribution of model QPF (as opposed to precipitation observations); though it may not be strictly necessary, it is the hope of the author that each model's attractor is close enough to the true atmosphere's attractor such that the

probability distribution of model QPF is in the same family as observed precipitation, even if the parameters vary substantially at local scales.

2.4.3 Right-Skewed Distributions

The tables below summarize the mathematical properties of the RSDs employed in this research (Hosking 1997; Hosking and Wallis 1987; Hosking and Wallis 1993; De Haan and Ferreira 2007). Mathematical intuition on the differences in the distributions may not be readily apparent by inspection of the defining equations; more graphical comparisons will be provided in Chapter 4.

Table 2.2: RSDs used in this research. Full name, abbreviated name, valid interval, and equations for each distribution's PDF and CDF are included. L-moment estimators for each model are included as applicable and possible. All equations expressed such that μ denotes the location parameter, σ denotes the scale parameter, and ξ denotes the shape parameter.

Distribution Name	Exponential	Gamma	Generalized Extreme Value
Abbreviation	EXP	GAM	GEV
Valid Interval	$[0,\infty)$	$[0,\infty)$	$\begin{cases}\left[\mu-\frac{\sigma}{\xi},\infty\right) & \xi>0\\ (-\infty,\infty) & \xi=0\\ \left(-\infty,\mu-\frac{\sigma}{\xi}\right] & \xi<0\end{cases}$
PDF	$f(x;\ \mu,\sigma)=\frac{1}{\sigma}e^{-\frac{x-\mu}{\sigma}}$	$f(x;\ \sigma,\xi)=\frac{1}{\Gamma(\xi)\sigma^{\xi}}x^{\xi-1}e^{-\frac{x}{\sigma}}$ With $\Gamma(\xi)=\int_0^{\infty}z^{\xi-1}e^{-z}dz$	$f(x;\ \mu,\sigma,\xi)=\frac{1}{\sigma}t(x)^{\xi+1}e^{-t(x)}$ With $t(x)=\begin{cases}\left(1+\left(\frac{x-\mu}{\sigma}\right)\xi\right)^{\frac{-1}{\xi}} & \xi\neq\\ e^{-\frac{x-\mu}{\sigma}} & \xi=\end{cases}$
CDF	$F(x;\ \mu,\sigma)=1-e^{-\frac{x-\mu}{\sigma}}$	$F(x;\ \sigma,\xi)=\frac{1}{\Gamma(\xi)}\gamma\left(\xi,\frac{x}{\sigma}\right)$ With $\Gamma(\xi)$ as above; $\gamma(a,b)=\int_0^{b}z^{a-1}e^{-z}dz$	$F(x;\ \mu,\sigma,\xi)=e^{-t(x)}$ t(x) as above

L-Moments Estimators	$\hat{\mu} = l_1 - 2l_2$ $\hat{\sigma} = 2l_2$	$\hat{\mu} = \begin{cases} \left(1-\frac{l_2}{l_1}\right)\frac{0.7213-0.5947\left(1-\frac{l_2}{l_1}\right)}{\left(1+\left(1-\frac{l_2}{l_1}\right)\left(1.2113\left(1-\frac{l_2}{l_1}\right)-2.1817\right)\right)} & \frac{l_2}{l_1} \geq 0{,}5 \\ \frac{1-0.308\pi\left(\frac{l_2}{l_1}\right)^2}{\pi\left(\frac{l_2}{l_1}\right)^2\left(1+\pi\left(\frac{l_2}{l_1}\right)^2\left(0.01765\pi\left(\frac{l_2}{l_1}\right)^2-0.05812\right)\right)} & \frac{l_2}{l_1} < 0.5 \end{cases}$ $\hat{\sigma} = \frac{l_1}{\hat{\mu}}$	*

Table 2.3: Continuation of Table 2.2

Distribution Name	Generalized Logistic	Generalized Normal	Generalized Pareto
Abbreviation	GLO	GNO	GPA
Valid Interval	$(-\infty, \infty)$	$(-\infty, \infty)$	$\begin{cases} [\mu, \infty) & \xi \geq 0 \\ \left[\mu, \mu - \frac{\sigma}{\xi}\right] & \xi < 0 \end{cases}$
PDF	$f(x;\ \mu,\sigma,\xi) = \begin{cases} \frac{\left(1+\frac{x-\mu}{\sigma}\xi\right)^{-1-\frac{1}{\xi}}}{\sigma\left(1+\left(1+\frac{x-\mu}{\sigma}\xi\right)^{-1}\right.} & \\ \frac{e^{-\left(\frac{x-\mu}{\sigma}\right)}}{\sigma\left(1+e^{-\left(\frac{x-\mu}{\sigma}\right)}\right)^2} & \end{cases}$	$f(x;\ \mu,\sigma,\xi) = \frac{\xi}{2\sigma\Gamma\left(\frac{1}{\xi}\right)} e^{-\left(\frac{\lvert x-\mu\rvert}{\sigma}\right)^{\xi}}$	$f(x;\ \mu,\sigma,\xi) = \frac{1}{\sigma}\left(1+\frac{x-\mu}{\sigma}\xi\right)^{-\frac{1}{\xi}-1}$
CDF	$F(x;\ \mu,\sigma,\xi) = \begin{cases} \frac{1}{1+\left(1+\frac{x-\mu}{\sigma}\xi\right)^{-1/\xi}} & \\ \frac{1}{1+e^{-\left(\frac{x-\mu}{\sigma}\right)}} & \end{cases}$	$F(x;\ \mu,\sigma,\xi) = \frac{1}{2} + \frac{x-\mu}{\lvert x-\mu\rvert}\frac{\gamma\left(\frac{1}{\xi}, \left(\frac{\lvert x-\mu\rvert}{\sigma}\right)^{\xi}\right)}{2\Gamma\left(\frac{1}{\xi}\right)}$	$F(x;\ \mu,\sigma,\xi) = 1 - \left(1+\frac{x-\mu}{\sigma}\xi\right)^{-\frac{1}{\xi}}$
L-Moments Estimators	$\hat{\mu} = l_1 - \frac{l_2\sin(-\pi l_3)}{\pi l_3}\frac{1+\frac{\pi l_3}{\sin(-\pi l}}{l_3}$ $\hat{\sigma} = \frac{-l_2\sin(-\pi l_3)}{\pi l_3}$ $\hat{\xi} = -l_3$	$\hat{\xi} = -l_3^2\frac{\left(2.0467+l_3^2\left(-3.6544+l_3^2(1.8397-0.2036l_3^2)\right)\right)}{1+l_3^2\left(-2.0182+l_3^2(1.242-0.2174l_3^2)\right)}$ $E = e^{0.5\hat{\xi}^2}$ $\hat{\sigma} = \frac{l_2\hat{\xi}}{E * \mathrm{erf}\left(\frac{\hat{\xi}}{2}\right)}$ $\hat{\mu} = l_1 + \frac{\hat{\sigma}(E-1)}{\hat{\xi}}$	$\hat{\mu} = l_1 - l_2\left(2+\frac{1-3l_3}{1+l_3}\right)$ $\hat{\sigma} = l_2\left(1+\frac{1-3l_3}{1+l_3}\right)\left(2+\frac{1-3l_3}{1+l_3}\right)$ $\hat{\xi} = \frac{1-3l_3}{1+l_3}$

Table 2.4: Continuation of Table 2.2

Distribution Name	Gumbel	Kappa	Weibull
Abbreviation	GUM	KAP	WEI
Valid Interval	$(-\infty, \infty)$	$\begin{cases} \left[\mu + \frac{\sigma(1-\beta)^{-\alpha}}{\alpha},\ \mu + \frac{\sigma}{\alpha}\right] & \beta > 0, \alpha > 0 \\ [\mu + \sigma \ln \beta\,, \infty) & \beta > 0, \alpha = 0 \\ \left[\mu + \frac{\sigma(1-\beta)^{-\alpha}}{\alpha}, \infty\right) & \beta > 0, \alpha < 0 \\ \left(-\infty,\ \mu + \frac{\sigma}{\alpha}\right) & \beta \leq 0, \alpha > 0 \\ (-\infty,\ \infty) & \beta \leq 0, \alpha = 0 \\ \left[\mu + \frac{\sigma}{\alpha}, \infty\right) & \beta \leq 0, \alpha < 0 \end{cases}$	$[0, \infty)$
PDF	$f(x;\ \mu, \sigma) = \frac{1}{\sigma} e^{-\left(\frac{x-\mu}{\sigma} + e^{-\frac{x-\mu}{\sigma}}\right)}$	$f(x;\ \mu, \sigma, \alpha, \beta) = \left(\left(1 - \beta\left(1 - \alpha\frac{x-\mu}{\sigma}\right)^{\frac{1}{\alpha}}\right)^{\frac{1}{\beta}}\right)^{1-\beta}$	$f(x;\ \sigma, \xi) = \frac{\xi}{\sigma}\left(\frac{x}{\sigma}\right)^{\xi-1} e^{-\left(\frac{x}{\sigma}\right)^{\xi}}$
CDF	$F(x;\ \mu, \sigma) = e^{-e^{-\frac{x-\mu}{\sigma}}}$	$F(x;\ \mu, \sigma, \alpha, \beta) = \left(1 - \beta\left(1 - \alpha\frac{x-\mu}{\sigma}\right)^{\frac{1}{\alpha}}\right)^{\frac{1}{\beta}}$	$F(x;\ \sigma, \xi) = 1 - e^{-\left(\frac{x}{\sigma}\right)^{\xi}}$
L-Moments Estimators	$\hat{\mu} = l_1 - \gamma\frac{l_2}{\ln 2}$ $\hat{\sigma} = \frac{l_2}{\ln 2}$	*	*

*: These distributions do not apply a closed form solution for parameter estimation, instead using an iterative scheme

2.4.4 Parameter Estimation

One of the principal challenges in the application of parametric techniques concerns the question of how to best estimate distribution parameters. Considerable scientific inquiry has been devoted to this research question; several of the most prominent methods are presented here, including the method of moments (MoM), method of L-moments (MoLM), maximum likelihood estimation (MLE), and direct solve (DS).

2.4.4.1 Method of Moments (MoM)

The MoM was the first commonly used method for parameter estimation, and it is now considered rather antiquated for most applications. A brief description is, however, presented here for

historical background and as a baseline for other methods. Again, let X denote a random variable. Then the n^{th} moment of X, with PDF f_X is defined as:

$$M_n(X) = E[X^n] = \int_{-\infty}^{\infty} x^n f_X(x) dx$$

For a finite sample of size *s*, a moment can be estimated as:

$\widehat{M_n}(X) = \frac{1}{s}\sum_{i=1}^{s} x_i{}^n \approx g_n(\theta_1, \ldots, \theta_P)$, where Θ's denote the distribution parameters and g's denote explicit functions of the population parameters. For example, for a normal distribution with parameters μ and σ, $g_1(\mu, \sigma) = \mu$, and $g_2(\mu, \sigma) = \mu^2 + \sigma^2$. The MoM algorithm starts with the first moment, and uses the approximate equality between the sample moment and the explicit formula for the true moment as a function of the distribution parameters to form an equation. This is repeated P times down to the P^{th} moment, yielding a system of P equations and P unknowns. This system of equations can then be solved analytically to yield parameter estimates $\widehat{\theta_1}, \ldots \widehat{\theta_P}$. The primary appeal of this method is that it is very tractable without requiring any additional, external information beyond the initial input data. The estimators are *consistent*, that is, as the sample size gets large, the parameter estimate converges to the true population parameter. However, MoM estimators are often *biased*, meaning that for any finite sample, the difference between the expected value of the true population parameter and parameter estimate is often non-zero (Hansen 1982).

2.4.4.2 Method of L-Moments (MoLM)

The use of sample L-Moments, rather than traditional moments, has been found in many applications to improve the quality of estimates of distribution parameters. From a sorted sample of size s, with x_1 being the smallest element of the sample, the sample L-moment of order *n* may be computed as:

$$\widehat{\lambda_n}(X) = \frac{1}{n\binom{s}{n}} \sum_{i=1}^{s} \left(\sum_{b=0}^{n} -1^b \binom{n}{b} \binom{i-1}{n-b} \binom{s-1}{b} \right) x_i$$

The true L-moment of order *n*, may be expressed using an ordered, ascending, independent sample of size *s* (X_1 smallest, X_s largest) as:

$$\lambda_n(X) = \frac{1}{n} \sum_{i=0}^{n-1} -1^i \binom{n-1}{i} E[X_{n-i}]$$

Since the expected value for an ordered statistic can be readily computed from a known distribution, this equation can be used to derive L-moments as a function of population parameters. Using sample L-moments as population L-moment estimates, much like the MoM, increasing orders of L-moments can be employed until P equations relating the sample L-moments and their parameter-dependent equations are generated. The P unknown parameters can then be solved analytically to yield parameter estimates. The L-moment method's primary advantage over the traditional MoM is it's increased robustness. This appears in two key ways. First, the constraints on the existence of high order L-moments is much looser than those for traditional moments; specifically, the only requirement on the existence of high-order L-moments is that the distribution have a finite mean, while traditional moments require stricter conditions be upheld. More importantly, while still not *resistant statistics*, L-moments are much more robust to outliers or extremes in the sample data when compared with the use of traditional moments. This makes the MoLM especially attractive for extreme value applications (Hosking 1992; Hosking 2006; Hosking and Wallis 1993; Hosking and Wallis 2005; Pilon and Adamowski 1992; Guttman et al. 1993).

2.4.4.3 Maximum Likelihood Estimation (MLE)

The MLE approach to parameter estimation considers the problem from a Bayesian framework. Specifically, given a sample X of *s* IID observations, $X=\{x_1,\ldots,x_s\}$, and a vector of distribution parameters

Θ, the joint density function can be readily computed: $f(X|\theta) = \prod_{i=1}^{s} f(x_i|\theta)$ (using the IID assumption). This is related to $P(X|\theta)$, often expressed as the *likelihood function* ι through Baye's Rule: $P(\theta|X) = \frac{f(X|\theta)P(\theta)}{P(X)}$. Exploiting the monotonicity of the logarithm and applying logarithm identities, this can also be re-written as: $\ln \iota(\theta|X) = \sum_{i=1}^{s} f(x_i|\theta)$. The MLE estimator $\widehat{\theta_{MLE}}$, is then:

$$\widehat{\theta_{MLE}} = \text{argmax}_{\theta}(\ln \iota)$$

MLE, like the methods above, is a *consistent* estimation method- as the sample size gets large, the parameter estimates converge to the true population parameters. With finite sample sizes, evidence suggests that MLE often produced better estimates for a fixed sample size than moment-based methods. However, MLE is more expensive, often lacking an analytical solution and requiring numerical iteration to converge to an estimate. Further, in some instances, no MLE solution exists; this occurs when ι continues to increase without attaining a maximum, or supremum, value. Still, MLE is a very powerful and general method that can be effectively applied to the parameter estimation challenge in many different contexts (Murphy 2012).

2.4.4.4 Direct Solving (DS)

The DS method is a very straight-forward and viable approach to parameter estimation, but appears very infrequently in the literature due to practical constraints on its use. The idea is quite simple: given a distribution D of P parameters, since the equation for D's CDF is known as a function of its parameters, specifying P precipitation threshold, quantile (or CDF-value) pairs yields a system of P equations with P unknown parameters. The system of equations, often with extensive algebra, can then be analytically solved to yield estimates for each of the P parameters. The quality of the estimates is directly proportional to the accuracy of the given quantile, threshold pairs; if each of those is perfect, the parameter estimates will necessarily also be perfect. The challenge, then, is obtaining accurate (quantile, threshold) pairs; it is often infeasible to obtain sufficiently accurate pairings to yield

reasonable parameter estimates, which limits the method's utility in many settings. However, at more common thresholds, where the event ARI is small relative to the data record length, obtaining accurate threshold rarity estimates may be possible, and this would make the use of this method quite attractive.

2.5 Extreme Precipitation and Precipitation Datasets

2.5.1 Precipitation Datasets

2.5.1.1 Stage IV Precipitation

NCEP Stage IV Precipitation Analysis products (Lin and Mitchell 2005) have been created daily in an official capacity since December 2001. Stage IV provides precipitation analyses over the contiguous United States (CONUS) by with hourly and 6-hourly accumulation analyses, and 24-hour accumulation analyses by means of summing four 6-hourly accumulations. Analyses are given on an approximately 4 km grid. Stage IV uses both rain gauge observations and radar-derived rainfall estimates to generate an analysis, and is further quality controlled via NWS River Forecast Centers (RFCs) to assure stray radar artifacts and other spurious anomalies do not appear in the final product. Even with these procedures, Stage IV has numerous deficiencies that will be discussed in further detail in subsequent chapters. However, despite its limitations, the combination of its data record length and analysis quality and resolution are deemed to make it superior to alternative available products for this research, and will be used as the precipitation 'truth' for the purposes of this study.

2.5.1.2 Atlas 14 and Prior Work

The National Oceanic and Atmospheric Administration (NOAA), and specifically the Hydrometeorological Design Studies Center (HDSC), is currently developing an updated assessment of precipitation accumulations to occurrence frequency equivalences for rare events with RPs of 1 to 1000 years over CONUS. In so doing, events may be studied in the context of their climatological rarity rather than a fixed threshold which has different implications over the geographically and

hydrometeorologically diverse CONUS. This product, known as Atlas 14, is an update of work done by Hershfield in 1961, published that year 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. Atlas 2, using AMS methods to convert to PDS statistics, fit a 2-parameter GUM distribution to station gauge data for 6- and 24-hour accumulation intervals to derive 2- to 100-year return period estimates. Topographically-aware formulas were then derived and applied to extend those estimates to all points (Miller et al. 1973). However, only 2- and 100-year return period thresholds have been digitized; the author manually calculated other RP thresholds using the DS method (see 2.4.4.4) for the Gumbel distribution. Atlas 2 frequency estimates remain the most up to date estimates for five northwestern states: Idaho, Montana, Oregon, Washington, and Wyoming. TP-40 methods are nearly identical, also using AMS to PDS conversion and the GUM distribution (Hershfield 1961). TP-40 estimates are the most recent in place for Texas and New England, including New York. In all other states, Atlas 14 updates have superseded the Atlas 2 and TP-40 estimates in place previously. In addition to having several decades of new data with increased station density to improve precipitation frequency estimates, Atlas 14 uses more sophisticated methods for deriving estimates than its predecessors. A suite of different RSDs were fit to precipitation data, using the MoLM for parameter estimations; goodness of fit tests such as Kolmogorov-Smirnov 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 which 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-year up to 1000-years, and estimates are available for accumulation intervals ranging from minutes to months (Bonnin et al. 2004; Bonnin et al. 2006; Perica et al. 2011; Perica et al. 2013).

2.5.2 Modes of Extreme Precipitation

Precipitation accumulation can be described as

$$P = \int_{t_o}^{t_f} R(t)dt = \bar{R}(t_f - t_o) = \bar{R}D$$

where P is the precipitation accumulation and R is the instantaneous precipitation rate (Doswell et al. 1996). It follows that, in order to receive large precipitation accumulations P, the product of average precipitation rate and precipitation duration must be very high; either instantaneous rain rates must be exceptionally high, or moderate rain rates must exist for a long duration, or somewhere in between. The moderate rate, high duration (MRHD) events require a source of ample moisture and an additional source for persistent lift. This class of event, as will be discussed in more detail in the section below, is most often seen in association with *atmospheric river* events along the US west coast, where an atmospheric river acts to transport tropical moisture poleward to the mid-latitudes, and the coastal topography acts as a constant forcing for ascent with tropical moisture from the atmospheric river being brought directly into mountainous topography (Ralph and Dettinger 2011). However, atmospheric instability in these regimes is typically insufficient to develop the very high rain rates required for the high rate, moderate duration (HRMD) event class. Unlike the predominantly stratiform precipitation observed in MRHD events, HRMD events are almost exclusively convective in nature. Extreme precipitation has been observed in a wide variety of convective regimes, from Mesoscale Convective Systems (MCSs) to Squall Lines to High Precipitation (HP) Supercells (Schumacher and Johnson 2006). All of these phenomena occur somewhat routinely without producing extreme precipitation, often because their storm/system motion is too high, resulting in a precipitation duration that's too short for any given location to receive extreme precipitation amounts. Receiving extreme amounts from convective systems often requires that the system's propagation opposes its cell motion (which is strongly related

to the mean flow/wind) to yield slow storm motions (Schumacher and Johnson 2006). Extreme precipitation can also occur in association with Tropical Cyclones (TCs); for reasons that will not be discussed in detail here, TCs generate rainbands with very intense thunderstorms which constitute the TC *eyewall*. TCs, despite being very different in nature to most events in the same class, would usually be most accurately classed as HRMD events, but if a TC stalls over land due to interaction with topography or some other reason, TC rainfall can often be accurately classified as high rate, high duration.

2.5.3 Extreme Precipitation Climatology

One of the most comprehensive studies of extreme precipitation climatology from the RP framework was conducted by Stevenson and Schumacher (2014). The study focused exclusively on states lying entirely east of the continental divide, and due to the timing of the analysis, Atlas 14 data was not yet available, and the TP-40 grids were used instead. The study looked at accumulation intervals of 1, 6, and 24 hours, and RPs of 50 and 100 years. By inspection of the grids (see Figure 3.1 later), it is apparent that use of the RP framework departs substantially from the traditional fixed threshold framework; for a given RP, the ratio between the highest and lowest precipitation thresholds is often two to three or more. The general patterns exhibited are largely what one would intuitively expect; higher precipitation accumulations are required for the same frequency of occurrence near the Gulf Coast, with notably less accumulation required near the Canadian border to the north and approaching the Rocky Mountains to the west. Higher thresholds are, of course, ubiquitously observed for the higher 100-year RP compared with 50-years.

In terms of event analysis over a 10-year period from 2001-2011, 24-hour events at both RP thresholds examined exhibit a broad peak in frequency over the summer months, which precedes a somewhat sharp decline in frequency during the autumn and follows a sharp ascent in the late spring.

Almost no 24-hour events were found to occur during the winter months. There was found, however, to be some regional variation, with the Plains region exhibiting an earlier peak in May and June; the identified Southeast events occurring almost exclusively in August and September in association primarily with tropical cyclone activity; the Northeast region occurring in August, September, and October; and the Ohio-Mississippi Valley region experiencing two peaks in frequency- in May and September- with moderate frequency of occurrence throughout the summer months. Identified events were classified into three categories: 1) Mesoscale Convective System (MCS), 2) Synoptic, and 3) Tropical; it was found that a substantial majority of CONUS heavy precipitation events east of the continental divide occurred in association with MCSs. 6-hour events exhibited a fairly similar pattern to the 24-hour events, except that: 1) OH-MS Valley no longer showed a bi-modal peak pattern, instead with a single broad summertime peak; 2) the Plains peak shifted more towards the summer, instead centered about June; 3) the Northeast October maximum largely disappeared; and 4) while August and especially September remained the Southeast peak, events did occur in that region outside those months. 1-hour events occurred again similarly to 6-hour events, but the Northeast region event frequency shifted further towards the summer months with a July-August maximum and very few September-October events, and the Southeast peak continued to broaden, with many one-hour events occurring in July in that region. The one-hour events were also examined with regard to time of day; though there were some minor regional variations, all regions experienced most 1-hour 50- and 100-year events during the late afternoon, evening, or early nighttime hours, from roughly 16:00 to local midnight. While few or no studies have looked in depth at the climatology of extreme precipitation events west of the continental divide in an RP framework, similar studies in a fixed-threshold framework suggest that extreme events over the US west coast occur primarily in association with synoptic systems and atmospheric river events in the cool season- autumn and winter- and do not occur through the MCS or Tropical modes that dominate much of the rest of the country.

2.6 Machine Learning

All machine learning algorithms used in this study are, at least as applied here, supervised learning models used for the purpose of classification. The prediction problem here is said to be *supervised* because each predictive model is being trained and tuned on *labeled* data, that is, historical data from which the outcome- observed RP exceedances- are known. In the case of categorical RP exceedance forecasting, a finite number of possible observations exist: 100-year exceeded, 50-year exceeded but not 100, 25 exceeded but not 50, 10 but not 25, 5 but not 10, 2 but not 5, 1 but not 2, 1-year RP not exceeded. This can also be reduced to a series of binary problems with regards to a particular RP threshold exceedance (e.g. two classes: 10-year RP exceeded, 10-year RP not exceeded). Because the forecast problem involves predicting a discrete category rather than a quantified, numeric predictand, the machine learning task is deemed to be a *classification* problem rather than a *regression* problem. As a broad overview, each of these supervised predictive models ingests as input numerous labeled *training examples* and uses these to train a final predictive model, which serves as the output of the training phase of this process. Specifically, the outputted, trained model ingests one or more *unlabeled* examples and outputs a prediction- either a single best-guess classification or assigns a verifying probability to each possible classification category for the true label of each input example. Aside from the label, each *training example* also possesses with it a representation of the information available on which to make a prediction. This is typically formatted as a list of predictors, or features, which altogether comprise a *feature vector* (Murphy 2012). For this forecasting application, the features are forecast variables from NWP models used in the PFS's NWP ensemble, with each individual feature corresponding to a specific atmospheric field forecast at a given latitude and longitude for a specific NWP model, depending on what information is being used to train the predictive model.

2.6.1 Logistic Regression

Perhaps the most basic, fundamental method for developing a statistical model is linear regression, as used in MOS and elsewhere. The idea is to express the predictand of interest as a linear combination of the input predictors, or features. Suppose one has *n* records, or training examples, of the form: $< y_i, \vec{F}_i = [1, x_1, \dots, x_m] >$, with each example possessing a vector of *m* features, along with the verifying observation y. The idea of linear regression is to express the predictand of interest as a linear combination of the input predictors, or features: $y_i = \vec{\beta} \cdot \vec{F_i} + \varepsilon_i$, where β is a vector of predictor coefficients, and ε is an error term. This algorithm is powerful for many applications, but has its limitations. Principally, linear regression produces predictand estimates on the spectrum $(-\infty, \infty)$, while probabilities occur on the spectrum [0,1], and observations are members of the set {0,1}. For this reason, linear regression is fundamentally a theoretically flawed approach for the application of probabilistic prediction. This is because it belongs to a class of algorithms for the purpose of *regression*- prediction of a continuous predictand, which is the wrong class of algorithms to apply to the forecast problem examined here. Instead, one seeks robust *classification* algorithms- those that predict which of a discrete set of categories an example belongs.

The name "logistic regression" is quite a misnomer, since in fact it is not a regression algorithm but instead a classification algorithm. Logistic regression (LOG_REG) does, however, share a great deal in common with linear regression. Both techniques are instantiations of the Generalized Linear Model (GLM); the two approaches simply have different underlying assumptions. Three requirements must be satisfied for any instantiation of the GLM. First, a linear predictor η is required; that is, the predictand may be expressed as a function of η, which in turn may be expressed as a linear combination of the input features $\eta = \vec{\beta} \cdot \vec{F}$. There must also be an identified probability distribution, or *distribution function*, of the predictand mean. Lastly, there must be a *link function* connecting the predictor η with mean of the distribution function, υ. In linear regression, the distribution function is the normal

distribution: $f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, and the link function is the identity function: υ = η. In LOG_REG, the Bernoulli distribution serves as the distribution function: $f(x|p) = p\big(\delta(x-1)\big) + (1-p)\big(\delta(x)\big)$, where δ() denotes the Dirac Delta Function, and *p* is the input parameter indicating the probability of occurrence. LOG_REG's link function is the logit, or inverse sigmoid function, which can be re-written as the inverse link function: $\upsilon = \frac{e^{\eta}}{1+e^{\eta}}$. As desired, for any real value for η, υ now possesses a value on the interval [0,1], corresponding to the probability of the feature vector $\vec{F}$ corresponding to η belongs to the positive verification category (Wilks 2011).

2.6.2 Decision Trees and Random Forests

2.6.2.1 Decision Trees

Decision trees are one fairly basic method for approaching classification problems. Decision trees, for the purposes of this study, consist of a network of two types of nodes: *decision nodes* and *leaf nodes*. Decision nodes each have exactly two children, which may be either decision nodes or leaf nodes, with a binary split based on the numeric value of a single feature from the an input example's feature vector. A leaf node has no children and instead, makes a categorical prediction of the verifying class of the input example based on the leaf's relationship to its ancestor nodes. For a given input example, one always begins at a decision tree's *root*, and at each decision node, compares the value of its feature to the critical threshold of the corresponding feature prescribed for that decision node. If the example's feature exceeds the node's critical threshold, the tree is traversed to the node's right child; otherwise, the tree is traversed to the left child. This process is repeated until a leaf node is reached; at this point, the value corresponding to the leaf becomes the predicted verifying category for the input example. In this way, the predictive model acts to make a categorical prediction by means of a conjunction of boolean variables derived from an example's feature vector. Once a decision tree is built, determining a

prediction given a feature vector is rather straight-forward; the challenge comes in the training phase in constructing the tree.

The two primary questions that must be addressed in constructing a decision tree are:

1a) At a given juncture, how is it determined what feature to split on?

1b) After determining a splitting feature, what determines the critical threshold?

2) What determines when to stop node splitting, and thus create a leaf node?

Suppose a decision tree is trained on *n* training examples, each with a feature vector F of length m. At a given node *k*, the candidate splits S consist of a feature f and threshold θ, $S = (f, \theta)$. The set of training examples that traverse the developing tree to reach k is denoted Q. S partitions Q into Q_{left} and Q_{right} by: $Q_{left} = \forall \langle y, F \rangle (F[f] < \theta)$; $Q_{right} = \forall \langle y, F \rangle (F[f] \geq \theta)$. There is said to be *impurity* I at k based on S; that is given by $I(Q, S) = \frac{len(Q_{left})}{len(Q)} H\left(Q_{left}(S)\right) + \frac{len(Q_{right})}{len(Q)} H\left(Q_{right}(S)\right)$, where H is the *impurity function*. Among the candidate splits S at k, the chosen split S* is the split satisfying: $S^* = argmin_S\big(I(Q, S)\big)$. This process of greedy split selection is continued recursively until the *termination criterion* is satisfied.

Traditionally, the *termination criterion* is simply that a node k is a *decision node* unless $len(Q) = 1$, in which case a leaf node with prediction y_0 ($Q = \langle y_o, F_o \rangle$). However, recursing this deep is very susceptible to fitting the noise of the training data, thereby *overfitting* the predictive model and degrading its generalized skill. To alleviate this concern, often more liberal termination criterion are applied, such as creating a leaf node whenever $len(Q) \leq len_{min}$; $len_{min} > 1$, or by imposing a maximum allowable depth D of the tree $depth(k) \leq D$.

2.6.2.2 Random Forests (RAND_FOR)

Decision trees can be a powerful approach for a wide array of applications, but they also have several significant drawbacks. First, they are widely regarded as low *bias*, high *variance* solutions. That is, minimal error is introduced by erroneous or oversimplistic assumptions in the model formulation, but the model formulation is very sensitive to the input data upon which it trains, which results in large error when extrapolating to other test data. More succinctly, decision trees are very prone to *overfitting* the training data: fitting to the noise of the training data rather than just the underlying relationships. This flaw substantially diminishes the utility of decision trees as a general predictive model. Second, the decision tree framework does not robustly extend to a probabilistic framework, since leaf nodes make deterministic predictions based on the mode verifying category of the subset of the training data reaching each respective node; applying a probabilistic prediction to individual leaf nodes greatly compounds the overfitting problem. It has been demonstrated that using many different decision trees to form, in aggregate, a predictive model can significantly decrease the model *variance* with only a slight increase to the model *bias*, provided the trees are sufficiently uncorrelated. This is the idea behind *random forests* (Breiman 2001).

The challenge with random forests is: how does one generate a large set (forest) of reasonably skillful decision trees that are not strongly correlated? The procedure described above for generating a decision tree from training data is *deterministic*, that is, a given set of training data will always produce the same decision tree via that algorithm. A forest of identical decision trees adds no value over using a single decision tree. The extra process for random forest generation is twofold: *tree bagging* and *feature bagging*. To generate a forest of size B from the *n* training examples, *tree bagging* involves the application of a simple bootstrapping procedure. Specifically, one samples, with replacement, *n* training examples from the original set, and uses this derived set to construct a decision tree using the method described above. This process is repeated B times to form a forest. Overfitting due to correlated trees

can still occur under this approach if a small subset of the original feature space are much more robust predictors of the verifying category than the rest. To overcome this problem, only a random subset of the *m* original input features are considered at each decision node; the size of the random subset is denoted here as $Z; 1 \leq Z \leq m$.

2.6.3 K-Nearest Neighbors Classification

Applying straightforward clustering techniques for classification problems can prove highly effective despite its simplicity. Perhaps the best known, the K-nearest neighbors clustering algorithm is explored here. KNN and other clustering algorithms have the unique property, compared with the other machine learning algorithms discussed here, that it is *non-generalizing*; test example predictions are made purely based on the proximity to training examples, rather than applying a fitted model which is extrapolated based on the training data. This is very advantageous when decision boundaries are highly erratic and non-linear, as other methods will tend to produce *biased* solutions in these instances. However, its inability to identify patterns in the training data can also cause it to use training data less efficiently than other algorithms in many instances.

K-Nearest Neighbors classification makes predictions based on a weighted vote of the K training examples judged most similar to the test example. Similarity of two data points is determined by a distance metric D applied to the points' feature vectors F_1 and F_2, smaller distances being more similar. The most commonly used distance metric is also the most intuitive, the Euclidean Distance:

$$D_{Euclid}(F_1, F_2) = \sqrt{\sum_{i=1}^{m} (F_1[i] - F_2[i])^2}$$

Distances between the test example and all training examples are computed, and the training examples associated with the K smallest computed distances comprise the set of voting neighbors. Each neighbor

votes in accordance with its associated verifying observation in the training data to yield a set of votes, or predictions, V. The final prediction of the KNN algorithm is then the product of matrix V and a normalized weights vector W. Traditionally, W's elements are all 1/K, so that each member has an equal vote, but may be chosen to instead vary with weights inversely proportional to distance (Murphy 2012).

2.6.4 Boosting

The basic concept of *boosting* (Friedman 2001) is that a large ensemble of weak learners- very high bias, very low variance models- can form a strong learner. Decision trees are a popular choice of weak learner for boosting, and were selected as the ensemble members for this study. Decision trees may be thought of as partitioning the m-dimensional feature space R^m into different segments, and then assigning a verifying category to each fragment based on the mode verifying category of the training data in that subspace. Each decision tree b can thus be characterized by its basis, or predictive, function h: $h_b(F) = \sum_{j=1}^{J} RP_{bj} X_{bj}\,, where\ X_{bj} = \begin{cases} 1\ F \in R_{bj} \\ 0\ F \notin R_{bj} \end{cases}$, where J is the number of segmented regions of feature space, RP_{bj} is the return period category assigned to the j'th segment of feature space for the b'th tree, and R_{bj} refers to that corresponding region. F here is the feature vector, which specifies a location in feature space. The net model M can then be expressed as a weighted sum of the basis functions:

$M(x) = \sum_{b=1}^{B} h_b(F)\gamma_b$, where γ_b's are coefficients.

At any step b-1, the b^{th} tree is constructed so as to minimize the loss function *L* satisfying:

$$M_b(F) = M_{b-1}(F) + argmin_h \sum_{i=1}^{n} L\big(y_i, M_{b-1}(F_i) - h(F)\big)$$

In gradient boosting, this minimization problem is accomplished by gradient descent:

$$M_b(F) = M_{b-1}(F) + \gamma_b \sum_{i=1}^{n} \nabla_M L\big(y_i, M_{b-1}(F_i)\big)$$

Where y_i corresponds to the verifying category of the i^{th} training example, and with:

$$\gamma_b = argmin_\gamma \sum_{i=1}^{n} L\left(y_i, M_{b-1}(F_i) - \gamma \frac{\partial L\big(y_i, M_{b-1}(F_i)\big)}{\partial M_{b-1}(F_i)}\right)$$

For this study, the chosen loss function for probabilistic QPF recurrence interval classification was *multinomial deviance*.

This process is repeated B times to form an ensemble of size B. Lastly, two extensions of this procedure attempting to reduce the *variance* of the final ensemble are explored in this study. The first is a *learning rate* where the iterative model avoids over-adjusting to new members by applying a dampening coefficient v: $M_b(F) = M_{b-1}(F) + \nu\gamma_b h_b(F)$. The second approach is taken from the idea of random forest creation: use only a subset of the total training data for each new decision tree. Instead of creating a new sample of size n, sampled with replacement from the original dataset, however, αn- where α is the *subsampling coefficient* between 0 and 1- training examples are sampled, without replacement, from the original training data.

2.6.5 Support Vector Machines (SVM)

Despite being rather abstract and difficult to interpret, both in the formulation and the output, Support Vector Classification (SVC; Cortes and Vapnik 1995) is an extremely powerful method which presents numerous advantageous. Due to its versatility, it generally extends to high dimensional feature spaces better than the other algorithms employed in this study, and can still work effectively even when the dimensionality of the feature space is larger than the number of training examples. Aside from the difficulty in physically interpreting the output of the fitted model, the primary drawbacks of this approach are that it cannot directly solve a multi-class classification problem, and also cannot directly

assign probabilities to its predictions. These limitations suggest at first glance that this method may be a poor choice for the problem of probabilistic forecasting of categorical RP exceedances, but due the power of the technique in addition to available workarounds, SVC is still examined here.

Support vector machines (SVMs) aim to define the hyperplane(s) which separates the training examples according to their respective labels and maintains as large of a margin as possible from any training example so as to minimize generalization error. Consider a two class problem, where, without loss of generality, all training examples can be associated with either class A, with a value of 1, or class B, with a value of -1. Each of the *n* training examples has an associated observation y; these may be assembled to a single observation vector Y of length *n*, which is thus comprised of elements $y_i \in \{1, -1\}$. Any hyperplane in the *m*-dimensional feature space can be described by: $\vec{n} \cdot \vec{F} - b = 0$, where b is a scalar and $\vec{n}$ is a vector normal to the hyperplane. In the event that the training data are *linearly separable* in the feature space, then a set of two hyperplanes may be considered: $\vec{n} \cdot \vec{F} - b = 1$ and $\vec{n} \cdot \vec{F} - b = -1$; these planes correspond to the nearest boundaries corresponding to each class. As stated above, SVMs are *maximum-margin classifiers*, that is, they seek to maximize the margin, or distance, between these two bounding planes. It can be readily shown that the margin between these hyperplanes may be expressed as: $\frac{2}{\|\vec{n}\|}$, where $\|\vec{n}\|$ denotes the norm of the vector defining the hyperplane. Thus, to maximize the separation margin, $\|\vec{n}\|$ must be minimized, subject to the constraint that no training example is misclassified. This can be readily expressed as an optimization problem:

Minimize $\|\vec{n}\|$ subject to: $y_i\left(\vec{n} \cdot \overrightarrow{F_i} - b\right) \geq 1, \forall i \in \{1, \ldots, n\}$

This problem can be readily solved (Cortes and Vapnik 1995); the implementation details will not be discussed here.

The approach above works well for training data that are *linearly separable* in feature space, but in general, this is not the case. The SVC approach may be generalized to allow for misclassifications; this is accomplished by creating a slack vector Ξ, whose elements ξ_i, allow misclassification by changing the constraints to the minimization problem to:

Minimize $\|\vec{n}\| + C\sum_{i=1}^{n}\xi_i$ subject to: $y_i\left(\vec{n}\cdot\overrightarrow{F_i} - b\right) \geq 1 - \xi_i$, for $\forall i \in \{1, \ldots, n\}$

In the expression above, C corresponds to the *penalty term,* or *inverse regularization coefficient;* it determines how smooth the decision surface should be, with a low value implying a highly regularized, low variance, high bias solution with smooth classification boundaries, while a high value implies a high variance, low bias solution that attempts to classify all of the training examples as they are actually labeled.

Even the extension above only allows for linear classification: the hyperplane must be defined as a linear combination of the original features. However, in many problems, a non-linear decision boundary better captures the true relationships between the input features and true classifications. This limitation can be solved too by use of *kernels* and the *kernel trick* (Murphy 2012). The mathematics of kernel theory and the kernel trick in particular are interesting, but not fundamental to an elementary understanding of SVC and thus will not be discussed here. Succinctly stated, the kernel trick exploits the fact that for some non-linear transformation ϕ to a feature vector F, $\varphi(F)$, the inner product of two such transformed vectors F_i and F_j may be expressed by a kernel k: $k\left(F_i, F_j\right) = \varphi(F_i)\cdot\varphi\left(F_j\right)$. This can be applied to transform the data into a much higher dimensional space, sometimes even infinite-dimensional, where the optimal decision boundary is linear in the transformed space. Applying this transformation, the problem formulation stays the same, except the kernel function replaces the inner product in the optimization constraints. Many different choices of kernels exist; some popular choices that will be explored in this study are: 1) Linear $k\left(\vec{a}, \vec{b}\right) = \vec{a}\cdot\vec{b}$, 2) Polynomial of degree d $k\left(\vec{a}, \vec{b}\right) =$

$\left(\gamma\left(\vec{a}\cdot\vec{b}\right)+r\right)^{d}$, 3) Radial basis function (RBF) $k\left(\vec{a},\vec{b}\right)=e^{-\gamma\left\|\vec{a}-\vec{b}\right\|^{2}}$, and 4) Sigmoid $k\left(\vec{a},\vec{b}\right)=$ $\tanh\left(\gamma\left(\vec{a}\cdot\vec{b}\right)+r\right)$, where γ and r are scalar constants that may be tuned.

The final limitations, namely (1) inability to extend to multi-class problems, and (2) inability to extend to probabilistic output, present more genuine problems in that they don't have 'pure' solutions. The former limitation has numerous possible workarounds. The approach utilized in this study applies a "one-versus-one" approach where $n_{classes}(n_{classes}-1)/2$ classifiers are fit to the training data, with each classifier corresponds to a unique pair of classification labels. The aggregate of classifiers is then used to make final class predictions. Probability estimates are made using a version of Platt Scaling; the method as applied here is both quite esoteric and *ad hoc*. As such, the method does have some known theoretical issues; principally, the predicted class in the deterministic problem may not have the plurality of the probability assignment in the probabilistic output. The details of the probability assignment phase will not be discussed here; for more information, see the Sci-Kit Learn User's Guide (Pedregosa et al. 2011).

2.7 Forecast Verification

2.7.1 Skill Scores

2.7.1.1 Brier Skill Score

The Brier Score (BS) is defined over N evaluation points as

$BS=\frac{1}{N}\sum_{j=1}^{N}(E_j-FP_j)^2$, with an observed event E defined by

$E_j=\begin{cases}1 & P_j\geq\theta_j\\ 0 & P_j<\theta_j\end{cases}$ and forecast probability $FP_j=P\left(P_j\geq\theta_j\right)$ corresponding to the event of whether the observed precipitation at point j, P_j, exceeds the critical precipitation threshold at that point, Θ_j, and the corresponding forecast probability (Brier 1950; Wilks 2011). Extending this to a latitude longitude grid

of Φ latitudes and L longitudes over D evaluation periods, the aggregated Brier Score becomes:

$BS_{agg} = \sum_{d=1}^{D} \sum_{y=1}^{\Phi} \sum_{x=1}^{L} \left(E_{yxd} - FP_{yxd}\right)^2$.

The Brier Score can be expressed as a skill score (BSS) by comparing with the Brier Score obtained from a reference forecast:

$$BSS = 1 - \frac{BS_{agg}}{BS_{agg_{ref}}}$$

Two common choices of reference forecast are climatology ($FP_{yxd} = FP_{yx_{clim}} \approx \frac{1}{D}\sum_{d=1}^{D} E_{yxd}$), which with a known R-year recurrence interval can simply be expressed as $FP_{yxd} = \frac{1}{365.25*R}$, and the worst possible forecast ($FP_{yxd} = 1 - E_{yxd}$). The resulting aggregated Brier Scores will be referred to as BS_{ref} and BS_{worst}, respectively, with corresponding skill scores of BSS and BSS_{best}. Taking note that $BS_{worst} = \Phi LD$, $BSS_{best} = 1 - \frac{BS_{agg}}{\Phi LD}$ (Wilks 2011).

The BS can also be decomposed into distinct components (Murphy 1973), each with a physical interpretation. Let the total number of N forecasts be subdividable into T distinct subcollections, with each forecast belonging to exactly one subcollection. The climatological frequency of event occurrence can be expressed as $\bar{o} = \frac{1}{N}\sum_{j=1}^{N} E_j$, and $\overline{o_t} = \frac{1}{\sum_{j=1}^{N} \begin{cases} 1 & j \in t \\ 0 & j \notin t \end{cases}} \sum_{j=1}^{N} \begin{cases} E_j & j \in t \\ 0 & j \notin t \end{cases} = \frac{1}{N_t}\sum_{j=1}^{N} E_j^t$.

$$BS = \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} (FP_j - E_j)(FP_j - \bar{o_t}) + E_j(1 - \bar{o_t}) & j \in t \\ 0 & j \notin t \end{cases}$$

$$= \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} (FP_j - E_j)(FP_j - \bar{o_t}) + \bar{o} - \bar{o_t}^2 & j \in t \\ 0 & j \notin t \end{cases}$$

$$= \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} (FP_j - E_j)(FP_j - \bar{o_t}) + \bar{o}(1 - \bar{o}) - (\bar{o_t}^2 - \bar{o}^2) & j \in t \\ 0 & j \notin t \end{cases}$$

$$= \bar{o}(1 - \bar{o}) + \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} (FP_j - E_j)(FP_j - \bar{o_t}) - (\bar{o_t}^2 + \bar{o}^2 - 2\bar{o}^2) & j \in t \\ 0 & j \notin t \end{cases} =$$

$$= \bar{o}(1 - \bar{o}) + \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} FP_j^2 - FP_j E_j - FP_j \bar{o_t} + E_j \bar{o_t} - (\bar{o_t}^2 + \bar{o}^2 - 2\bar{o}\bar{o_t}) & j \in t \\ 0 & j \notin t \end{cases}$$

$$= \bar{o}(1 - \bar{o}) + \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} FP_j^2 - 2FP_j \bar{o_t} + \bar{o_t}^2 - (\bar{o_t} - \bar{o})^2 & j \in t \\ 0 & j \notin t \end{cases}$$

$$= \bar{o}(1 - \bar{o}) + \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} (\bar{o_t} - \bar{o})^2 & j \in t \\ 0 & j \notin t \end{cases} + \frac{1}{N}\sum_{t=1}^{T} N_t \sum_{j=1}^{N} \begin{cases} (\bar{o_t} - FP_j)^2 & j \in t \\ 0 & j \notin t \end{cases}$$

$$= Uncertainty + Resolution + Reliability$$

This decomposition to the BS can be seen in reliability diagrams, as will be shown in section 2.7.3.

2.7.1.2 Fractions Skill Score

Though the Brier Skill Score allows for a simple and intuitive metric for evaluating forecast skill in a probabilistic framework, it does have several limitations. Principally, the BSS only compares the forecast probability assigned in direct collocation with event occurrence. Small displacement errors associated with a feature in a forecast may thus yield a skill score just as poor as a feature that misses the existence of the feature completely, but the former forecast still has much more utility to decision-makers and represents a solution much closer to reality than the latter forecast. The Fractions Skill

Score was developed by Roberts and Lean (2008), motivated in part to address some of the limitations of the BSS. Unlike the BSS, the Fractions Skill Score approach outlines a *neighborhood* within a fixed distance of an *evaluation point*. Within this neighborhood, the fraction of points within the neighborhood observed to have a verifying event is compared with the summed FP of all points within the neighborhood. In this way, a model is not penalized for displacing FPs slightly away from the verifying area, so long as the observed events and probabilities occur within the same neighborhood. This is advantageous in that a forecast is credited for assigning higher FPs in the vicinity of a verifying observation. It does, however, have the corollary disadvantage that a forecast is *not* credited for getting the exact positioning of an event correct; it will receive the same score as the displaced forecast provided that the summed FPs are identical and both occurring entirely within the neighborhood, or evaluation region, of the evaluation point.

For an evaluation radius r, over N evaluation points and D evaluation times, the Aggregated Fractions Skill Score (FSS) is given by:

$FSS = 1 - \frac{\sum_{d=1}^{D}\sum_{j=1}^{N}(O_{jd}-M_{jd})^2}{\sum_{d=1}^{D}\sum_{j=1}^{N}{O_{jd}}^2+{M_{jd}}^2}$, where

$O_{jd} = \frac{1}{(2r+1)^2}\sum_{y=lat_j-r}^{lat_j+r}\sum_{x=lon_j-r}^{lon_j+r}E_{yxd}$ and $M_{jd} = \frac{1}{(2r+1)^2}\sum_{y=lat_j-r}^{lat_j+r}\sum_{x=lon_j-r}^{lon_j+r}FP_{yxd}$, with

$E_{yxd} = \begin{cases} 1 & P_{yxd} \geq \theta_{yx} \\ 0 & P_{yxd} < \theta_{yx} \end{cases}$. Here, P_{yxd} denotes the observed precipitation at latitude y and longitude x accumulated over the period corresponding to observation record d, while θ_{yx} corresponds to the critical precipitation threshold, in this case the Z-year T-hour return period threshold of interest for the location indicated by latitude y and longitude x. FP_{yxd} corresponds to the forecast probability of the observed rainfall exceeding the critical precipitation threshold of interest at location (y,x) $P\left(P_{yxd} \geq \theta_{yx}\right)$.

The Fractions Skill Score exhibits interesting limiting behavior in both limits of evaluation radius. In the limit of small evaluation radius, where the evaluation radius is zero, the FSS reduces to simple point-by-point comparisons with observed 'fractions' as either 1 if the event was observed at that point, and 0 if it was not. Thus the FSS at evaluation radius 0 is simply the BSS with a reference forecast as the worst possible forecast, assigning zero probability whenever an event occurred, and probability one whenever an event did not occur (see BSS_{best} above). As such, the numerator of FSS for a single observation record, $\sum_{j=1}^{N}(O_{jd} - M_{jd})^2$, is referred to as the Fractions Brier Score (FBS), while the denominator $\sum_{j=1}^{N} {O_{jd}}^2 + {M_{jd}}^2$ is called the Worst Possible Fractions Brier Score (FBS_{Worst}). In the limit of large evaluation radius, the FSS reduces to a function of the Frequency Bias $\frac{f_o}{f_m}$, namely: $FSS_{r\to\infty} = \frac{2f_o f_m}{{f_o}^2+{f_m}^2}$, where f_o is the total frequency of observed events and f_m is the total frequency of forecast events.

2.7.1.3 Rank Probability Skill Score

The FSS and BSS metrics are designed for a forecast problem which aims to predict an event with a binary outcome: either it occurs, or it doesn't. For some purposes, this is certainly how you want to verify your forecast. For example, if something is very sensitive to the air temperature dropping below freezing, the user only cares about whether the temperature dropped below freezing; if it didn't drop below freezing, it doesn't matter whether the minimum temperature was 33°F or 40°F, and the verification metric should reflect this ambivalence. In the context of this research, it isn't quite so clear-cut; the impacts will be different and change at different thresholds for different users, but will generally increase monotonically with increasing local rainfall amount. For this reason, there may be some motivation to penalize "near-miss" forecasts less than complete busts. For example, at a given point with a critical rainfall threshold of 100 mm over some fixed period, with two forecasts issuing identical FPs of exceeding this threshold, and in one case 0 mm fall over the forecast period and 95 mm fall within

the other forecast period, there is reason to think that, in this forecast case, the latter forecast should be considered "better" than the former. At the same time, however, the problem of forecasting explicit QPF values adds a new layer of difficulty and complexity- compared with simply forecasting probabilities of exceeding various thresholds- that are beyond the scope of this present research.

The Rank Probability Score (RPS; Epstein 1969) presents an alternative which compromises between the all-or-nothing nature of the BSS and FSS, and the absolute error metrics such as root mean squared error or mean absolute error which require explicit precipitation accumulation forecasts and have many additional errors with the non-linear scaling occurring with extreme precipitation. The RPS metric is designed for a forecast problem where 1) the forecasts are probabilistic rather than deterministic, 2) there are several possible discrete verifying categories, and 3) those categories are *ordinal* rather than *nominal*, that is, they have a natural ordering. The forecast problem specific to this research can be framed in a way that satisfies all of these criteria. In the contexts of using the BSS or FSS, a separate calculation is made for each recurrence interval R of interest, with each FP simply corresponding to the $P\left(P_{yxd} > \theta_{R_{yxd}}\right)$. But in this "all-or-nothing" framework, a forecast for a 25-year event where a 10-year event verifies (but a 25-year event does not) will be treated the same as a case where a 1-year event does not verify. Given the high uncertainty, high impact, small-scale nature of many of these locally heavy rainfall events, it seems that giving partial credit for the 10-year event is a desirable verification property. Instead of viewing the forecast system as a series of binary probabilistic forecast problems, the forecasts can instead be framed in a single multi-category probabilistic forecast problem. For a set of return periods {1-year, 2-year, 5-year, 10-year, 25-year, 50-year, 100-year}, instead of making seven independent forecasts for the probability of exceeding each threshold and verifying each separately, the FS can instead predict a single eight-element probability vector, with the contents being $f = \langle P\left(P_{yxd} < \theta_{1_{yx}}\right), P\left(\theta_{2_{yx}} > P_{yxd} \geq \theta_{1_{yx}}\right), P\left(\theta_{5_{yx}} > P_{yxd} \geq \theta_{2_{yx}}\right), P\left(\theta_{10_{yx}} >\right.$

$P_{yxd} \geq \theta_{5_{yx}}\Big), P\left(\theta_{25_{yx}} > P_{yxd} \geq \theta_{10_{yx}}\right), P\left(\theta_{50_{yx}} > P_{yxd} \geq \theta_{25_{yx}}\right), P\left(\theta_{100_{yx}} > P_{yxd} \geq \theta_{50_{yx}}\right), P\left(P_{yxd} \geq \theta_{100_{yx}}\right)\rangle$ and the vector sum always totaling unity. The corresponding observation vector $o = \langle E_{<1_{yx}}, E_{1,2_{yx}}, E_{2,5_{yx}}, E_{5,10_{yx}}, E_{10,25_{yx}}, E_{25,50_{yx}}, E_{50,100_{yx}}, E_{>100_{yx}}\rangle$, with

$$E_{a,b} = \begin{cases} 1 & \theta_{b_{yx}} > P_{yxd} \geq \theta_{a_{yx}} \\ 0 & P_{yxd} > \theta_{b_{yx}} \; or \; P_{yxd} < \theta_{a_{yx}} \end{cases}$$

For the general case for a forecast problem with K ordered categories, the RPS may be expressed as:

$RPS = \sum_{m=1}^{K}\left(\sum_{j=1}^{m} f_j - \sum_{j=1}^{m} o_j\right)^2$, and this can be further aggregated over all evaluation points and forecast periods and compared with climatology to expressed as a Rank Probability Skill Score (RPSS):

$$RPSS = 1.0 - \frac{RPS}{RPS_{clim}} = 1.0 - \frac{\sum_{d=1}^{D}\sum_{y=1}^{\Phi}\sum_{x=1}^{L}\sum_{m=1}^{K}\left(\sum_{j=1}^{m} f_{jyxd} - \sum_{j=1}^{m} o_{jyxd}\right)^2}{\sum_{d=1}^{D}\sum_{y=1}^{\Phi}\sum_{x=1}^{L}\sum_{m=1}^{K}\left(\sum_{j=1}^{m} f_{clim_{jyxd}} - \sum_{j=1}^{m} o_{jyxd}\right)^2}.$$

The RPS is also an extension of the BS, reducing to it in the case of K=2 categories. This also means it suffers from the same spatial displacement errors as the BS that motivated the use of the FSS. However, the FSS and RPSS can be readily combined to form the Fractions Rank Probability Skill Score (FRPSS) which combines the advantages of both approaches. The FRPSS can be expressed by:

$$FRPSS = 1.0 - \frac{\sum_{d=1}^{D}\sum_{y=1}^{\Phi}\sum_{x=1}^{L}\sum_{m=1}^{K}\left(\sum_{j=1}^{m}\frac{1}{(2r+1)^2}\sum_{b=y-r}^{y+r}\sum_{a=x-r}^{x+r} E_{bad} - \sum_{j=1}^{m}\frac{1}{(2r+1)^2}\sum_{b=y-r}^{y+r}\sum_{a=x-r}^{x+r} FP_{bad}\right)^2}{\sum_{d=1}^{D}\sum_{y=1}^{\Phi}\sum_{x=1}^{L}\sum_{m=1}^{K}\left(\sum_{j=1}^{m}\frac{1}{(2r+1)^2}\sum_{b=y-r}^{y+r}\sum_{a=x-r}^{x+r} E_{bad} - \sum_{j=1}^{m}\frac{1}{(2r+1)^2}\sum_{b=y-r}^{y+r}\sum_{a=x-r}^{x+r} FP_{clim_{bad}}\right)^2}$$

2.7.2 Forecast Value and Value Scores

Forecast skill, though important, is not the end-all for forecast evaluation. Skill quantifies forecast accuracy- how 'good' the forecasts are. It does not, however, quantify how *valuable* the forecasts are, and ultimately, if the forecasts add no value, it does not really matter how *skillful* they are;

they're simply not of use to end users. Assessing forecast *value* is ultimately the most important metric to assess whether the forecasts are actually benefitting the end users, which in many instances is society at large. However, it is also one of the most difficult aspects of a forecast to objectively quantify. Different users have varying sensitivity and risk tolerance. Users have different critical impact thresholds; one farm may be on the 5-year floodplain, a second farm is on the 10-year floodplain but off the 5-year plain, and a third may be in between, with the farm flooding at an 8-year return period. Some user's losses will be closer to the all-or-nothing framework, while others may incur additional losses monotonically with increasing rainfall. A general framework to assess the forecast value with every possible theoretical user is untenable; however, a simple cost/loss contingency table based framework provides a useful, tangible means to assess forecast value and utility (Wilks 2011).

The Cost/Loss Model is predicated on two basic premises: 1) Given a preparation action A and observed precipitation accumulation P, the cost or damages are known, static, and quantifiable; 2) The costs/damages are a scaled Heaviside step function in the P dimension- that is, there are no costs until a certain critical threshold P_{crit} is exceeded, at which point some non-zero loss L is inflicted, and no subsequent precipitation beyond P_{crit} will result in costs different from L. These premises allow for the use of a basic contingency table framework. In one dimension, either P_{crit} was exceeded or it was not. In the other dimension, either the user can take protective action or not. Suppose taking protective action will cost \$C, but then no further costs will be inflicted regardless of whether P_{crit} is exceeded. If protective action is not taken and P_{crit} is exceeded, \$L losses are inflicted, with $L > C$. The contingency table (CT) can then be expressed as:

Table 2.5: Classic cost-loss model contingency table. C denotes the cost of preparing, L the loss burdened when the event occurs given no preparation.

Cost to User	Observed	Not Observed
Prepare	C	C
Don't Prepare	L	0

Under this, the expected cost E[$] with a probability of affirmative verification p is:

$$E[\$]_{prepare} = C, E[\$]_{\neg prepare} = Lp$$

The break-even point between the two action points occurs when $C = Lp \rightarrow p = \frac{C}{L} = \alpha \equiv$ Cost-Loss Ratio. The 'interesting' cases are restricted to users satisfying $0 < \alpha < 1$, otherwise one strategy strictly dominates the other.

In this framework, forecast value may be quantified by means of a value score (VS). For all cases, which may be enumerated as all latitudes Y, longitudes X, and times D, generate a suite of contingency tables for different user sensitivities as expressed through the cost-loss ratio α, with α ranging from 0 to 1. Each constructed table assumes rational users, who will prepare when the FP exceeds their cost-loss ratio, and not take mitigative action otherwise.

$$PREP = \begin{cases} 1 & Prepare \\ 0 & Don't\ Prepare \end{cases} = \begin{cases} 1 & FP_{yxd} \geq \alpha \\ 0 & FP_{yxd} < \alpha \end{cases}$$

The jargon of contingency tables is most often expressed as:

Table 2.6: Traditional contingency table terminology, as will be used in this research.

Contingency Table	Observed	Not Observed
Prepare	Hits (HITs)	False Alarms (FAs)
Don't Prepare	Misses (MISSs)	Correct Rejections (CRs)

The VS aims to quantify, in the CT framework, how the long-term expected economic cost of an end user of interest, with a specific cost-loss ratio α, compares with both acting based on the

climatological frequency of event occurrence $\bar{o} = \frac{HITs+MISSs}{TOT}$; $TOT = HITs + FAs + MISSs + CRs$, and a perfect forecast in which the user prepares for the event when and only when it occurs. The general form can be expressed as: $VS = \frac{E_{clim}-E_{fcst}}{E_{clim}-E_{perf}}$. Like a skill score, the VS is unity when the predictions of a forecast system cannot be improved, and zero when using the forecasts does not save the user of interest anything relative to acting based on climatological frequency of occurrence.

Using climatology as a forecast, a user will either always prepare or never prepare, whichever is cheaper in the long-run. This can be expressed for a single case, using the cost table above, with an expected cost of: $E_{clim} = \min(C, \bar{o}L)$, with the first argument corresponding to the protection cost, and the last as the expected non-protection cost.

It is readily seen that the perfect forecast has an expected cost of $E_{perf} = \bar{o}C$ per forecast.

The cost of using the forecast system being analyzed over the forecast period from which the contingency table is simply the product of the contingency table counts with the cost table: $E_{fcst} = HITs * C + FAs * C + MISSs * L$

Multiplying the E_{clim} and E_{perf} expressions by the total number of cases in the forecast period, *TOT*, yields:

$$VS = \frac{\mathrm{TOT} * \min(C, \bar{o}L) - (HITs * C + FAs * C + MISSs * L)}{\mathrm{TOT} * \min(C, \bar{o}L) - TOT * \bar{o}C}$$

Defining $h = \frac{HITs}{TOT}$; $f = \frac{FAs}{TOT}$; $m = \frac{MISSs}{TOT}$; $r = \frac{CRs}{TOT}$ and dividing out TOT*L yields:

$$VS = \frac{\min(\alpha, \bar{o}) - (h + f)\alpha - m}{\min(\alpha, \bar{o}) - \bar{o}\alpha}$$

After computing VS's for a suite of α's, a plot of VS as a function of α can be created; this type of figure is known as an Economic Value Diagram (EVDG).

In many cases, but especially in the instance of extreme events, it is often unrealistic that a user can fully protect against the event; if it occurs, in many instance one would still expect to endure more losses than if it didn't, even when prepared. The CT framework introduced above can be readily generalized to include a *baseline loss* L_0, in addition to an *extra loss* L_{ext}. The cost table is modified to read:

Table 2.7: Modified contingency table to accommodate a mitigated loss. L_0 denotes the base loss, L_{ext} denotes the additional loss beyond the base loss if prepared, and C denotes the cost of preparing.

Cost to User	Observed	Not Observed
Prepare	$C + L_0$	C
Don't Prepare	$L_0 + L_{ext}$	0

In this framework, the new costs are: $E_{clim} = \text{TOT} * \min(C + \bar{o}L_0, \bar{o}L_0 + \bar{o}L_{ext})$; $E_{perf} = TOT * (\bar{o}C + \bar{o}L_0)$; $E_{fcst} = HITs * C + HITs * L_0 + FAs * C + MISSs * L_0 + MISSs * L_{ext}$

And the associated VS formula:

$$VS$$

$$= \frac{\text{TOT} * \min(C + \bar{o}L_0, \bar{o}L_0 + \bar{o}L_{ext}) - (HITs * C + HITs * L_0 + FAs * C + MISSs * L_0 + MISSs * L_{ext})}{\text{TOT} * \min(C + \bar{o}L_0, \bar{o}L_0 + \bar{o}L_{ext}) - TOT * (\bar{o}C + \bar{o}L_0)}$$

$$= \frac{\min(C + \bar{o}L_0, \bar{o}L_0 + \bar{o}L_{ext}) - (hC + hL_0 + fC + mL_0 + mL_{ext})}{\min(C + \bar{o}L_0, \bar{o}L_0 + \bar{o}L_{ext}) - (\bar{o}C + \bar{o}L_0)}$$

$$= \frac{\min(C, \bar{o}L_{ext}) - (hC + fC + mL_{ext})}{\min(C, \bar{o}L_{ext}) - \bar{o}C} = \frac{\min(\alpha_{ext}, \bar{o}) - ((h+f)\alpha_{ext} + m)}{\min(\alpha_{ext}, \bar{o}) - \bar{o}\alpha_{ext}},$$

With $\alpha_{ext} = \frac{C}{L_{ext}}$

Noting that the break-even point in this framework occurs when:

$C + L_0 p = L_0 p + L_{ext} p \rightarrow \alpha \equiv p = \frac{C}{L_{ext}} = \alpha_{ext}$, it is noted that the addition of a baseline loss does not change the VS metric (Zhu et al. 2002; Mylne 2002).

2.7.3 Reliability Diagrams

As explained in section 2.3.2, forecast reliability, namely the ability for FP = ORF, is a desirable property of a PFS. PFS reliability is most often assessed by means of a reliability diagram. Over a large number of verification records, forecasts are binned into clusters of approximately equal FP for an event of interest. For each bin, the fraction of corresponding event occurrences is computed; this approximates the PFS's ORF for that FP. Then, mean bin FP is plotted against bin ORF, with ORF on the ordinate and FP on the abscissa, and (FP,ORF) points are connected to form a reliability line (RL). The closer the reliability line tracks to the one-to-one FP=ORF line, the more reliable the PFS (Wilks 2011). However, the reliability line may also be used to make more specific diagnoses. An 'S' shape reliability line indicates an overspread ensemble yielding underconfident predictions; conversely, an 'inverted S' line corresponds to an underspread EPS with overconfident probabilities. Further, $\int_{x=0}^{1} RL(x) - x\, dx > 0$ indicates a negatively biased PFS, issuing too low of FPs. Similarly, $\int_{x=0}^{1} RL(x) - x\, dx < 0$ suggests a positively biased PFS, thinking events are more likely to occur than they actually are. Some additional features frequently accompany an RL on an RD. Often, a horizontal line is placed through the RD indicating the climatological event ORF; this line also corresponds to the "zero resolution" line, since if the RL falls along this line, it does not distinguish at all between events and non-events. An extension of this is to draw the line which tracks the mid-point between the zero resolution line and the one-to-one line; this line is referred to the "no skill" line, since points along this line neither add nor subtract to the BSS, which as shown in 3.7.1.1, is readily decomposable in the framework of a reliability diagram. To give the reader better context, a normalized histogram of bin sizes also typically accompanies an RL. A reliability diagram with all of these additional properties includes is often referred to as an "attributes diagram" (Hsu and Murphy 1986).

3 Model Diagnostics and Evaluation

Given the impact of extreme precipitation on life, property, and industry, it is critical to forecast locally extreme precipitation events as well as current resources allow. In order to do this, 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. 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 to forecast locally extreme rainfall events.

3.1 Data & Methods

Analysis of model performance in the return period (RP)/recurrence interval framework first requires establishing the actual numerical thresholds corresponding to the RPs of interest for all locations of interest. This paper seeks to assess model performance in all regions of the contiguous United States (CONUS); thus nationwide threshold grids are required. RPs of 1-, 2-, 5-, 10-, 25-, 50-, and 100-years were evaluated for this verification work. Due to 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 RPs for various accumulation intervals (AIs). Over the past several years, NOAA has made a major effort to update these estimates via the Atlas 14 project. At the time of this writing, Atlas 14 has updated previous RP estimate for the majority of CONUS; however, the northwest: Washington, Oregon, Idaho, Montana, and Wyoming; northeast: New York, Vermont, New Hampshire, Maine, Massachusetts, Connecticut, and Rhode Island; and Texas have

not yet received published updated estimates from Atlas 14[2]. For these areas, older estimates using older data and methods were used to complete the map of CONUS estimates. For more specific details about each of the datasets used, consult section 2.5.1.2 of this document.

After establishing CONUS-wide RP thresholds for the 6- and 24-hour AIs, these were used as the basis for determining the climatology of locally extreme precipitation events as discernable from Stage IV precipitation analysis (see section 2.5.1.1 for description) over the last several years. Further, forecasts of several numerical weather prediction (NWP) models were used to assess individual model characteristics and biases in the forecasting of extreme precipitation and also to quantify model skill. Depending on model availability, either the 09 June 2009 to 30 August 2014 period or the shorter 12 August 2014 to 11 August 2015 period was selected for evaluation and comparison. The Global Ensemble Forecast System Reforecast Version 2 (GEFS/R) and National Severe Storms Laboratory Weather Research and Forecasting (NSSL-WRF) models were evaluated for the 24-hour AI over the longer 2009-2014 period for 12Z-12Z forecasts off of each model's 00Z initialization; the NSSL-WRF was also quantitatively evaluated over the same evaluation period for 6-hour forecasts for four different times of day: 00-06Z, 06-12Z, 12-18Z, and 18-00Z. The beginning of this period coincides with the implementation of a significant update to the NSSL-WRF model. The GEFS/R, by design, has no model changes of any kind during this analysis period; 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. A broader comparison of several Convection Allowing Models (CAMs) was conducted over the shorter 2014-2015 period comparing the NSSL-WRF, the North American Mesoscale 4km Nest (NAM-NEST), and experimental version of the High Resolution Rapid Refresh (HRRR). The beginning of this period coincides with a major update to the NAM-NEST which is thought to have significantly altered the model's bias characteristics in Quantitative Precipitation

[2] After conducting this research and writing this manuscript, it was reported that updated Atlas 14 estimates for the northeastern states would be released on 10/1/2015.

Forecasts (QPFs). More details about all of these models can be found in Section 2.2. All of these forecasts were compared with Stage IV analysis over the respective periods. Beyond struggles with data coverage in complex terrain, Stage IV analysis, despite the internal quality control (QC) prior to public release, exhibits several additional aberrations which require additional QC procedures. Specifically, additional QC procedures are performed to alleviate two significant problems. First, 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 very high associated automated accumulated precipitation amounts. Due to other priorities with the internal QC process by RFCs, this is not always QC'd out in the final Stage IV product. Second, on some days, large regions of exceptionally high precipitation totals are reported and, for whatever reason, are not removed from the final product. A combination of automated and manual means is used to combat these issues. As RPT exceedances, by definition, for any given instance occur with some known specified frequency *p*. Correlations of model QPF time series was 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 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 where removed from the dataset. O

Over the extended period analysis, quantitative assessments of model skill were also conducted by means of the Fractions Skill Score (FSS). Details on this skill metric are provided in Section 2.7.1.2.

3.2 Results

Model analysis and verification is presented first in a broad, qualitative context to give an appreciation for approximate model performance and characteristics for as many models as possible, in addition to providing an overview of the characteristics of the climatology of extreme precipitation in the US in the context of datasets used in this study. Later results will attempt to quantify model bias characteristics and model skill.

The composite threshold maps for a 24-hour AI appear in Figure 3.1. Panels (a), (b), (c), (d), (e), (f), (g) correspond to the 1-, 2-, 5-, 10-, 25-, 50-, and 100-year RPs, respectively. As expected, the thresholds increase monotonically with increasing RP. For the 1-year RP, several parts of the country, in particular areas of the arid and intermountain west, have 24-hour RPTs of less than 25 mm, or one inch. A few locations even have 2-year 24-hour RP thresholds below one inch. In contrast, much wetter regions of the country such as the Pacific coastal mountains and southeast Gulf Coast region experience one-year 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 the traditional FT 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 required precipitation accumulation for a fixed frequency of occurrence. At the 100-year RP threshold there are parts of the west with thresholds below 50 mm- lower than the 1-year RP in other parts of CONUS- and other places where the thresholds are in excess of 500 mm. Lastly, 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 TX/OK and NY/PA borders in Figure 3.1d. 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 MT and WY in Figure 3.1a.

Corresponding grids for the 6-hour AI appear in Figure 3.2. Unsurprisingly, many of the patterns seen here are very similar to those seen in Figure 3.1, just with all-around lower thresholds due to the shortened AI. 1-year 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-year RP thresholds range from near 30 mm to approximately 300 mm accumulations. However, the biggest qualitative differences between this figure and Figure 3.1 relate to regional differences in the nature of extreme precipitation events. In the west, extreme precipitation events are predominantly long-duration stratiform events in which abundant moisture is advected from the ocean to the land and precipitates out from lift by the coastal topography. Because extreme precipitation events in this area tend to be long duration, low-moderate intensity, there is a large difference between the 6- and 24-hour AI thresholds. In contrast, many extreme precipitation events in the east are driven by convective cells or convective systems, and tend to be shorter-duration, higher-intensity events than seen in the west. As such, the difference between the 6- and 24-hour thresholds is not as large, since many of the heavy precipitation events occur predominantly within a 6-hour window anyways. For example, the 1-year RP for the Olympic Mountains of Washington reach nearly 200 mm for the 24-hour AI, but is reduced to roughly 80 mm at the same location for the 6-hour AI. Compare with central IA, where the 1-year 24-hour thresholds are near 65 mm, but the 1-year 6-hour thresholds are down to only near 55 mm. Many of the data source contrasts still exist in Figure 3.2; the NY/PA difference is significantly amplified in the 6-hour thresholds compared with the 24-hour threshold differences.

Figure 3.3 compares forecasted 2-year 6-hour threshold exceedances for the NSSL-WRF, HRRR, and NAM-NEST in panels (a), (b), and (c), respectively, against the observed exceedances as discerned

from Stage IV precipitation analysis appearing in panel (d). Figure 3.3 analysis is confined to the 00-06Z valid period for each day from 12 August 2014 through 11 August 2015. The plots illustrate the established climatology of extreme precipitation events over CONUS. Most events, both forecast and observed, occur during the cool season months of October through 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 US from TX up through ND is seen to experience most events during the early part of the warm season- primarily from April through July, while the eastern states have almost no identified events in April and May, with almost all events seen between June and September. The southeast US has perhaps the most diverse collection of identified events seasonally, with several events being identified in both the warm and cool seasons. However, there is a distinct maximum in identified events during the mid-to-late hurricane season, from August through October, as many extreme precipitation events in this region are associated with tropical cyclone activity. Over the plains and Midwest, it is also of note 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. Lastly, some data anomalies are worth noting. There is a stark decline in the number of observed events crossing from PA into NY and further into New England; this is also reflective of the dramatic increase in 2-year 6-hour RP thresholds noted in Figure 3.2b. Given that the number of events in PA and surrounding states to the south and west is in rough proximity to the number of 2-year events one would anticipate seeing over a one-year period for a given 6-hour interval, this suggests that the 6-hour thresholds derived from TP-40 data are likely too high. There is also only a small number of identified observed events in the southwest US in AZ, UT, NV, and SE CA. Unlike the NE US 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 here Atlas 14 threshold estimates, but instead due to poor precipitation verification estimates due to the complex terrain and poor radar coverage of this region.

Comparing the model forecasts, it is immediately apparent upon inspection of Figure 3.3 that the NAM-NEST forecasts many more 2-year 6-hour events over the 00-06Z time of day compared with either other convection allowing model (CAM) analyzed here, in addition to the Stage IV analysis. This is true essentially throughout CONUS, but is especially amplified over much of the western US. As noted in the methods, a fair amount of HRRR data is missing from the verification period, and as a result, Figure 3.3b underrepresents the number of events forecasted. All the models have similar seasonal characteristics to event forecasts to the truth seen in Stage IV analysis. The NSSL-WRF plot, Figure 3.3a, looks closest to the verification plot, Figure 3.3d, perhaps suggesting that this model performed the most skillfully over CONUS for the prediction of 2-year 00-06Z events over the verification period. However, the analysis performed here is only qualitative, and no such conclusions about model skill may be definitively reached. It does, however, appear to have the best frequency bias characteristics, with NAM-NESTs frequency bias being very large. The seasonality of local extreme precipitation events in the CAMs assessed here appears to be in approximate agreement with observations at each location, though springtime events appear to be especially overforecast in many of the CAMs in the intermountain west, southeast, and Mid-Atlantic regions.

Figure 3.4 shows analysis of forecasted and observed 50-year 6-hour events over the same 1-year analysis period as above, again for the 00-06Z time of day. While some of the overall patterns and comparisons discerned from Figure 3.3 continue here, the differences between the plots are more profound. Observed events from Stage IV analysis are found to occur for this time of day predominantly in a corridor just to the east of the Rocky Mountains, with many events identified in New Mexico, west Texas, Oklahoma, Kansas, North Dakota, and the eastern halves of Colorado, Wyoming, and Montana.

While some events are identified outside this region, the vast majority of identified events occur in this corridor during the warm season, likely due to convective storms developing off the mountains during warm season afternoons, around 20-23Z. This signature is not nearly as apparent in any of the CAMs. The NSSL-WRF, in Figure 3.4a, has a slight concentration of events in the same areas as observations, but tends to predict many more events in the west and, to a lesser extent, areas in the east. The HRRR (Figure 3.4b) does not have a concentration to the immediate east of the Rockies at all, with more events identified in the west, in the Ohio River Valley, and in the upper southeast and southern Mid-Atlantic regions than to the immediate east of the Rockies. Like in Figure 3.3c, the NAM-NEST in Figure 3.4c has so many identified events that it obscures the slight enhanced concentration of events to the east of the Rockies, as many events are also forecasted to the east and west as well. With many fewer identified events, it is also possible to identify the quality of some specific model forecasts for individual events, such as extreme rainfall in Texas in May 2015- the HRRR and NSSL-WRF tended to handle the eastern TX flooding event reasonably well, while the western TX event was more poorly forecasted. The January 2015 Washington extreme rainfall was best predicted by the NSSL-WRF and NAM-NEST, though the NAM-NEST produces so many false alarms that the signal is not as apparent. The June 2015 Montana event is also seen in the NSSL-WRF and HRRR forecasts in addition to appearing in the Stage IV analysis.

Figures 3.5 and 3.6 present the same information as Figures 3.3 and 3.4, respectively, except for the 06-12Z valid period instead of 00-06Z. The seasonality is still much the same as in the 00-06Z period, but the concentration of identified events to the east of the Rockies has shifted further east, with most points now appearing in the Great Plains region from Texas north through South Dakota and east into Arkansas into Indiana. Most major identified 50-year events as seen in Figure 3.6d, such as the event in September 2014 in S AZ, the event in the same month in E TX, the event in E IN in August, and some of the west coast events during various months, were reasonably well predicted with corresponding circles

appearing in most or all of the CAMs in each of these cases. However, the CAMs appear to have several major false alarm forecast busts; the NSSL-WRF for example predicts a large 50+ year 6-hour event in E MT and another in E NC/VA, neither of which verify. All three models predict an event in W OR in November 2014 that doesn't verify, and the NAM-NEST, like in the 00-06Z analysis, predicts many events all over CONUS that do not verify. In general, there also appears to be slightly fewer events seen at both the 2- and 50-year RP compared with the 00-06Z period, perhaps in association with diminished convection associated with the absence of solar heating.

The 12-18Z analysis is depicted in Figures 3.7 and 3.8 for the 2- and 50-year RPs, respectively. Panels 3.7b and 3.8b should be interpreted with great caution, since a substantial amount of 12Z initialized HRRR data was not available to the author at the time of this writing, including entire months of autumn 2014. As such, far fewer forecast events will appear on these plots than were likely actually forecast by the model during the analysis period. Unsurprisingly, the fewest number of events are seen in this morning period, both in the model forecasts and in observations. The locations of event occurrence in Figure 3.8d are also considerably shifted; contrasting, for example, the period 12 hours apart- as seen in Figure 3.4d- one sees that the region immediately east of the Rockies- highlighted in the 00-06Z period- has very few events in the 12-18Z period. The locations of event occurrence are considerably more scattered in the 12-18Z period, with the most events seen in the northern high plains and in the south and southeast. As seen in previously discussed periods, the NAM-NEST continues to considerably overforecast the number of extreme precipitation events for both RPs over almost all of CONUS, with a smaller, but still positive, bias observed in the NSSL-WRF. This appears to be especially true in the west, where both models have multiple significant false alarms- even at the 50-year RP- during west coast cool season events. The NSSL-WRF appears to have reasonable frequency bias characteristics over the eastern two thirds of CONUS, though the bubble correspondence between in particular Figures 3.8a and 8d appears to be poorer than in the 00-06 and 06-12Z periods.

Finally, analysis from the last six hour period- from 18-00Z- appears in Figures 3.9 and 3.10, again for the 2-year and 50-year return period thresholds. This period sees a considerable uptick in events relative to the six hours prior, and appears to be more on-par with the two other six-hour periods analyzed here. Two regions stand out as particularly vulnerable to extreme precipitation during this time of day. The first region exists along and immediately to the east of the Rocky Mountains, even to the west of the region identified in the 00-06Z period. As discussed there, as the Rockies frequently act as a source for convective initiation, and this is the first period in the west with considerable solar heating after sunrise, it makes intuitive sense that heavy precipitation is seen developing in this region in the 18-00Z period, and then propagating progressively to the east in the 00-06 and 06-12Z periods. For 18-00Z, there are almost no observed events in the Mississippi Valley region at the 50-year RP, and a considerable decline in event density is seen in Figure 3.9d for the 2-year RP as well. However, to the east near and along the Atlantic coast, a considerable number of events are seen from Florida up through Maine predominantly during the summer. This signature is well-captured in each of the CAMs analyzed here, each having a local minimum in the Mississippi Valley region; however, all of the models again had too many forecasted events in the west relative to Stage IV. This is likely a combination of model bias and poor observational coverage in the complex terrain of the intermountain west leading to an underrepresentation of actual events in the Stage IV analysis. The very large bubble sizes in Figures 3.9c and 10c indicate that the NAM-NEST is particularly biased during this period relative to the others, especially in the intermountain west and southwest.

Results for 24-hour 12-12Z precipitation events using the longer 09 June 2009-30 August 2014 analysis period is depicted in Figure 3.11. As explained in Section 3.2, the GEFS/R replaces the NAM-NEST and HRRR for the 24-hour accumulation analysis. It is immediately apparent that the coarse GEFS/R model forecasts far fewer events at both the 10-year and particularly the 100-year RP than are actually observed. This is in stark contrast to the CAMs in the six-hour 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. Closer inspection of, for example, Figure 3.11d confirms this: almost all of the GEFS/R forecasted events occur either in 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. 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 (compare Figures 3.11c and e) 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 associated swaths of heaviest precipitation are clearly outlined in Figure 3.11d and less clearly in Figure 3.11f. Other events include the major September 2009 southeast flooding which involved exceptional synoptic-scale moisture transport into the region, the Arizona flooding of January 2010, and a stretch of exceptionally wet systems in Montana during May and June 2013. The intensity of other synoptic-scale systems that produced 100-year 24-hour precipitation events, such as Hurricane Sandy in October 2012 (see red bubbles in Figure 3.11f) were not forecasted by GEFS/R. The NSSL-WRF (Figure 3.11b) still performed better with some of the tropical cyclones, more adequately forecasting the rainfall amounts associated with both Sandy and also with Tropical Storm Debby, which affected northern Florida in June 2012. 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 (Figure 3.11e) is also well captured (Figure 3.11a). As in the six hour 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-hour locally extreme precipitation analysis as discerned by Stage IV precipitation analysis.

The frequency bias characteristics of the CAMs compared in Figures 3.3-3.10 is summarized quantitatively in Figure 3.12. Confirming what was seen in Figures 3.3-3.10, for all four times of day, the most events are seen during the warm season months, with less events per month witnessed during the cool season. In the 00-06 and 06-12Z periods the number of events per month varies on average by approximately an order of magnitude. In contrast, in the 12-18Z 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. This result is especially pronounced at the 50-year RP, as many cool-season months have few or no observed events during the 1-year analysis period. As previously indicated, the 12-18Z period is found to have the fewest number of observed events, with 00-06Z being the largest. Also, despite large month-to-month variability, the average ratio of 2-year events to 50-year events is approximately 25, as should be anticipated.

Comparing models, the NAM-NEST does indeed forecast many more events than are observed. In the 00-06Z period, the NAM-NEST predicts more events than were observed for each month at the 2-year RP; in the warm season months, the NAM-NEST forecasts roughly 3-4 times as many events as are observed during the 00-06Z period, while for the 50-year RP the difference is often a factor of 5-6 and is as large as an order of magnitude difference for some warm-season months. The bias is roughly similar for the 06-12Z and 12-18Z periods, but is further amplified in the 18-00Z period, with most warm season

months having an order of magnitude more forecasted events than are observed, and in some cases the difference is larger than a factor of 20 for the 50-year RP. 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-year RPs. For most warm-season months, it is seen to be slightly positively biased relative to Stage IV at the 2-year RP, but the difference is typically less than a factor of two. Though the month-to-month results are more variable for the 50-year 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-year events were observed during the cool season months to draw any definitive conclusions about the bias characteristics of the CAMs during this period. Though re-scaled in proportion to the fraction of forecasts missing from each months, due to the amount of missing data, HRRR forecast characteristics must be interpreted with care, especially for the 12-18Z and 18-00Z periods where more forecasts are missing. However the bias characteristics appear to be quite good, with the HRRR event count lines most closely tracking the Stage IV lines during the 00-06Z period. During the other periods, the HRRR is seen to be the only analyzed CAM to consistently underpredict events during the winter and spring, with the NSSL-WRF having the best bias characteristics during these months for all times of day, but the HRRR still generally has the frequency bias closest to unity during the late spring and summer of the three models assessed here.

Similar analysis for the 24-hour accumulation interval is illustrated in Figure 3.13, with lines for all seven RPs included. The seasonal cycle of events is similar to that of the 6-hour 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 two from unity each month. However, at higher RPs, some

discernable biases begin to emerge. The NSSL-WRF overforecasts high RP 24-hour locally extreme precipitation events during the cool season months, with observed biases of approximately 3-5 during this period. Figure 3.11 correctly gives the impression that GEFS/R greatly underpredicts high RP events. Figure 3.13 reveals that the overall bias characteristics, while almost always underpredicting, are not terrible for most months. In July however, GEFS/R underforecasts the number of events by an exceptional two orders of magnitude relative to the reasonably large number of events observed during that month over the five-plus year analysis period.

The remaining analysis quantitatively evaluates model skill over the longer 09 June 2009-30 August 2014 analysis period. Fraction skill score results at each RP assessed in this study for 6-hour NSSL-WRF forecasts between forecast hours 12 and 36 on the 00Z initialization are depicted in Figure 3.14. As expected, FSS generally increases with increasing evaluation radius, and the highest scores are seen at lower RPs, which- being on average less extreme- are typically more predictable. There are some exceptions; for example, the results for the 10-year RP are almost uniformly lower than the corresponding 25-year RP FSS results at each forecast period. This anomaly is likely attributable to a combination of random sampling and issues associated with the automated QC of the Stage IV analysis yielding a superior representation of reality at the 25-year RP relative to the 10-year RP observations. At all evaluation radii and RPs, the highest scores by a considerable amount are observed with the 12-18 hour forecast period. This is likely attributable to the combination of 1) the fact that, as previously noted, a lower proportion of events in this six-hour 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; and 2) this is the earliest forecast period analyzed with respect to forecast initialization time and thus has had the least time for non-linear error growth. In a similar vein, the FSSs at all evaluation radii are very low for the 50- and 100-year RPs at the 24 and 30 hour lead times, probably due to these periods being the farthest from 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 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 hour forecast lead times, better even than the 5-year RP verification for the last periods at the high evaluation radii.

A graphical representation of NSSL-WRF model skill for 6-hour locally extreme rainfall forecasts over CONUS appears in Figure 3.15. These plots have been aggregated over all four individual 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 Figure 3.15a. In contrast, at the high RPs, for example in Figure 3.15d, often none or just one event occurred over a given region during the 5-year analysis period, resulting in the plot highlighting areas where the model handled one event well. Panels 15a and b 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. Areas of the mountain and desert west and southwest were not well handled; this was seen qualitatively in many of the bubble plots. The same two events reappear in panels c and d 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 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 on forecasting the extent of locally extreme rainfall. Other events, such as Tropical Storm Debby, were reasonably well forecast during the 25-year RP, but not at the 100-year 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 many of the bubble plots. For the most extreme event threshold, the Floridian peninsula and the central plains were the areas with the worst verifying forecasts.

Figure 3.16 provides regional skill analysis for different times of day as opposed to the different RPs shown in Figure 3.15. Much of the variance in the northeast comes from the largely coincidental timing of tropical cyclone lanfall for the storms that impacted the region during this period. However, some trends can be more definitively discerned. Though 2-year, 6-hour events were observed during all four periods (see Figures 3.3d, 3.5d, 3.7d, and 3.9d) in the Pacific Northwest, skill was observed to be notably higher during 06-18Z than 18-06Z. In regions that experience extreme precipitation events from various types of systems and meteorological conditions, such as the southeast US, one sees substantially elevated skill during the less convective 12-18Z period in comparison to 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 06-12Z period, likely attributable to the loally anomalous time of occurrence of these types of events; typically ongoing convection is well to the east of this region at this time. The highest forecast skill in the upper midwest is seen in the 00-06Z period, while across CONUS in the desert southwest, skill is higher at this time and lowest six hours later in the 06-12Z period. Splotches 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.

Summary agregated FSS analysis for 24-hour accumulations at all RPs is provided for both the GEFS/R and NSSL-WRF models in Figure 3.17. The general findings are again as expected, with FSS generally increasing with increasing evaluation radius, and generally decreasing with increasing RP. Comparing Figures 3.14 and 3.17, skill scores are generally higher for the 24-hour accumulation interval compared with any of the 6-hour 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-hour events occurring in association with longer-lived, larger-scale processes as opposed to some 6-hour events which occur at higher frequency in association with isolated convective cells. At low and very high RPs, the NSSL-WRF appreciably outperforms the GEFS/R at all evaluation radii examined here. This should be anticipated, as, all else equal, the improved model resolution and ability to explicity simulate convection without use of a cumulus parameterization should tend to produce more realistic andskillful representation of clouds and pecipitation. However, at the middle RPs- from 10 to 50 years- the GEFS/R is competitive with and often even outperforms the GEFS/R. This may have to do with the types of systems that are found to exeed 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-year RP or greater, may often require a combination of large scale forcing and meso- and smaller scale forcing which can not be adequately simulated by GEFS/R and other models of similar horizontal resolution.

Regional representation of model performance for 24-hour events appears in Figures 3.18 and 3.19 for the NSSL-WRF and GEFS/R, respectively; and regional model performance is directly compared via Figure 3.20. Inspecting the 1-year RP verification, which provides a broader perspective of general model performance with locally heavy rainfall, it is apparent that slightly elevated skill is seen over much of the west and east coast, while skill over the plains, midwest, and southeast is depressed. 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-year RP in association with forecast quality of individual events; California, Arizona, Montana, the Mid-Atlantic, the IA/MO/IL region, and various parts of the southeast US begin to stand out as areas of locally enhanced forecast skill. These regions are further highlighted moving to the 25-year RP in Figure 3.18c; 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 occur in association primarily with particular extreme precipitation events discussed with Figure 3.11. 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 associated with locally enhanced precipitation forecst skill. The 100-year RP sees a further degredation of forecast skill overall, with particular impact on the California, Arizona, and midwest regions; skill in MT and FL remains rather high.

The trends in the GEFS/R are rather similar. Comparing the two at the 1-year RP (Figure 3.20a), 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 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, midwest, and 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 with the GEFS/R which does not benefit much from this property. Inspecting the 5-year and 25-year comparisons in Figures 3.20b and 3.20c, 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, while GEFS/R had better forecasts for the aforementioned January 2010 and September 2009 events in Arizona and the southeast United States, respetively, 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-year RP (Figure 3.20d).

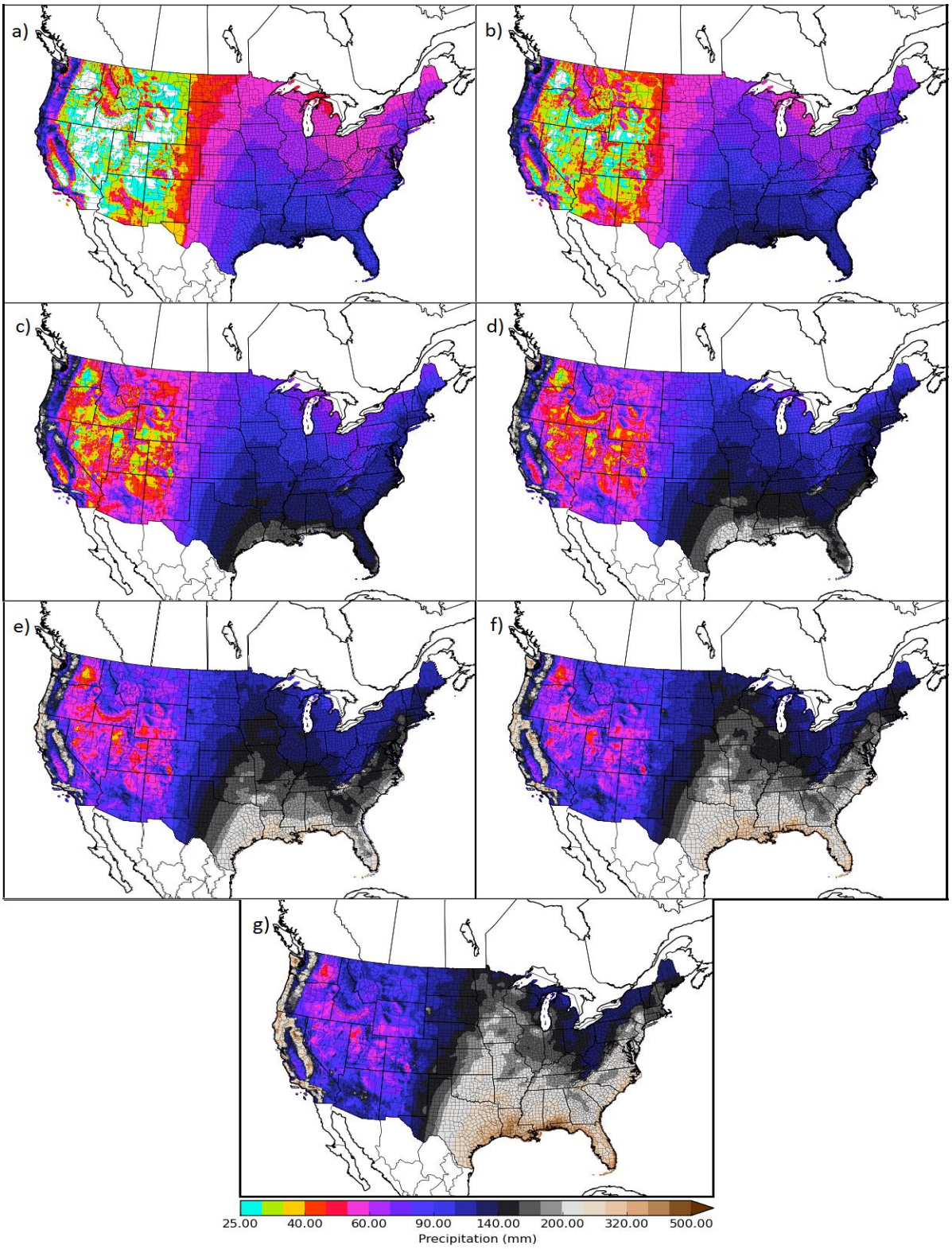

Figure 3.1 Return period thresholds over CONUS for a 24-hour accumulation interval. Panels (a)-(g) correspond to 1, 2, 5, 10, 25, 50, and 100 year return period thresholds, respectively. Threshold sources come from a combination of Atlas 14, TP-40, and Atlas 2 data as described in the paper text.

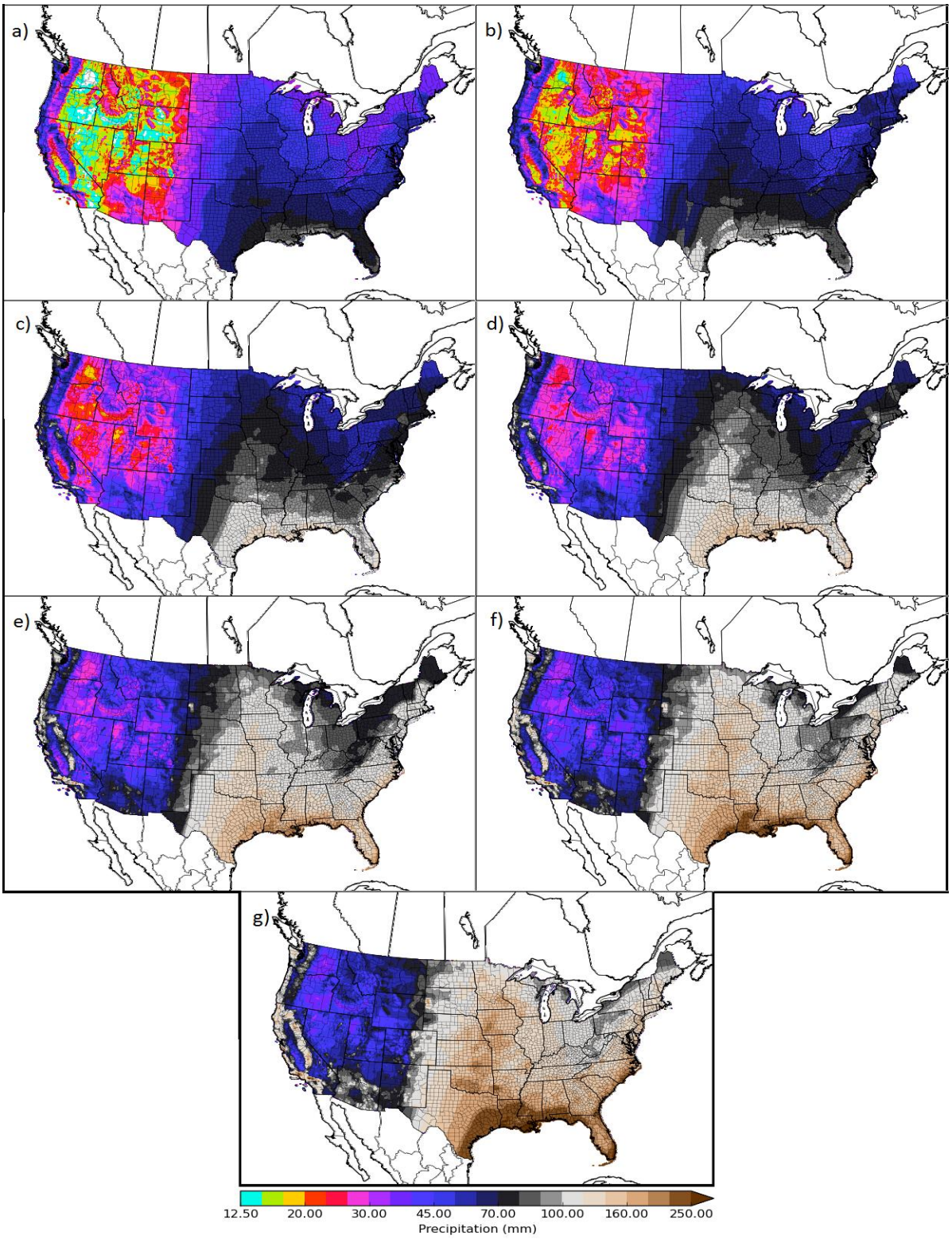

Figure 3.2: Return period thresholds over CONUS for a 6-hour accumulation interval. Panels (a)-(g) correspond to 1, 2, 5, 10, 25, 50, and 100 year return period thresholds, respectively. Threshold sources come from a combination of Atlas 14, TP-40, and Atlas 2 data as described in the paper text.

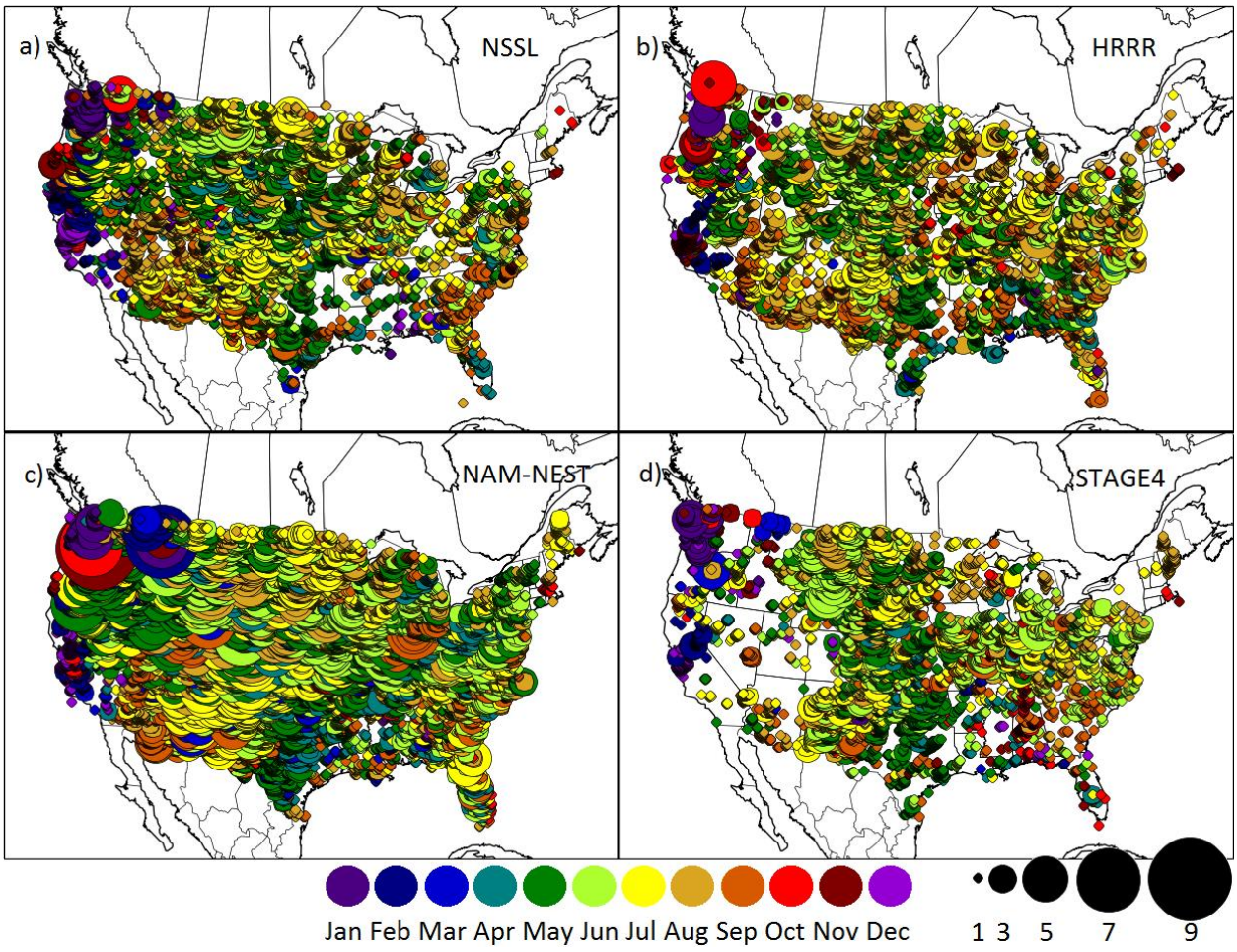

Figure 3.3 Forecasted and observed events of exceedances of the 2-year return period for a 6-hour accumulation interval, as illustrated in Figure 3.2b, over the 00-06Z period from 12 August 2014 through 11 August 2015. Circles indicate an observed or forecasted event at the location of circle center; circle size is proportional to number of events, with a larger circle indicating more events at that location. Black circles in the lower left indicate the circle size corresponding to a given number of events at a particular point. Panel (a) corresponds to forecasted events from the NSSL-WRF 24-30 hour precipitation accumulation from 00Z initialization. Panel (b) corresponds to the 0-6 hour forecast of the HRRR from 00Z initializations. Panel (c) corresponds to forecasted events from the 24-30 hour forecast of the operational 4km NAM-NEST initialized at 00Z. Panel (d) corresponds to observed exceedances of the local 2-year 6-hour threshold 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. Every other grid point in each dimension is assessed in constructing circles; thus, only one quarter of the total number of grid points is analyzed.

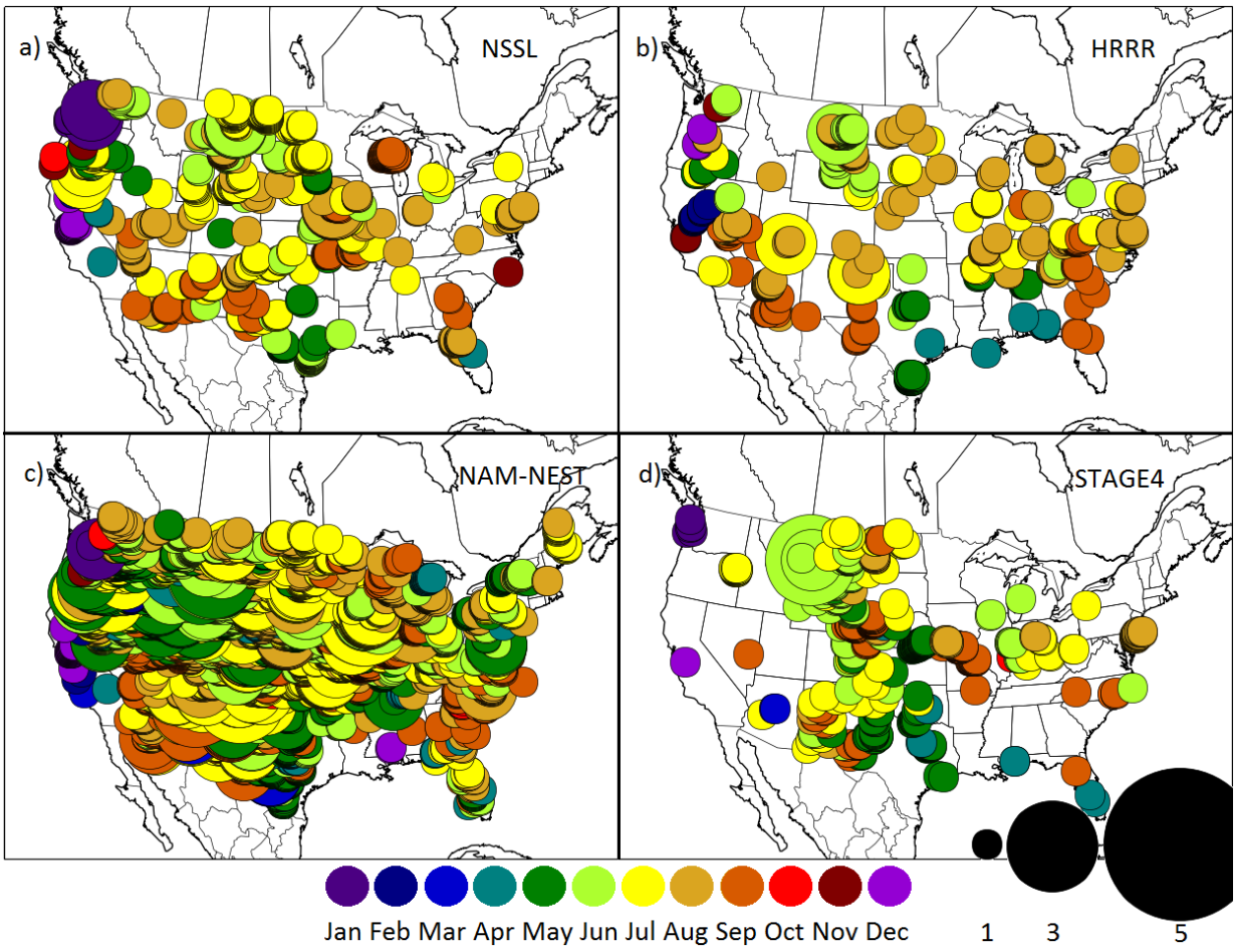

Figure 3.4: As in Figure 3.3, but for the 50-year return period thresholds.

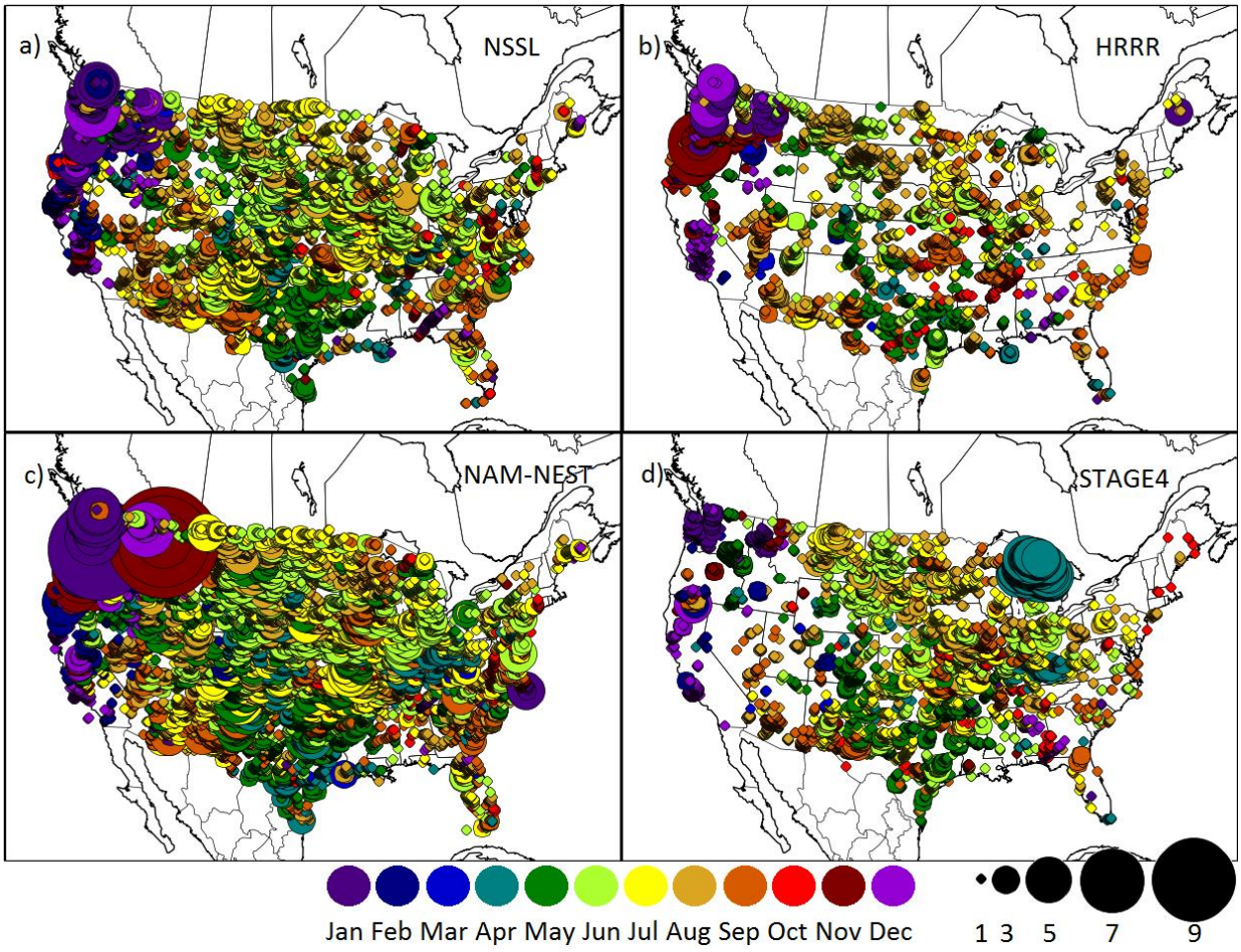

Figure 3.5: As in Figure 3.3, but for the 06-12Z period. NSSL-WRF, NAM-NEST, and HRRR forecasts are taken from the 6-12 hour precipitation forecast from the 00Z initialization.

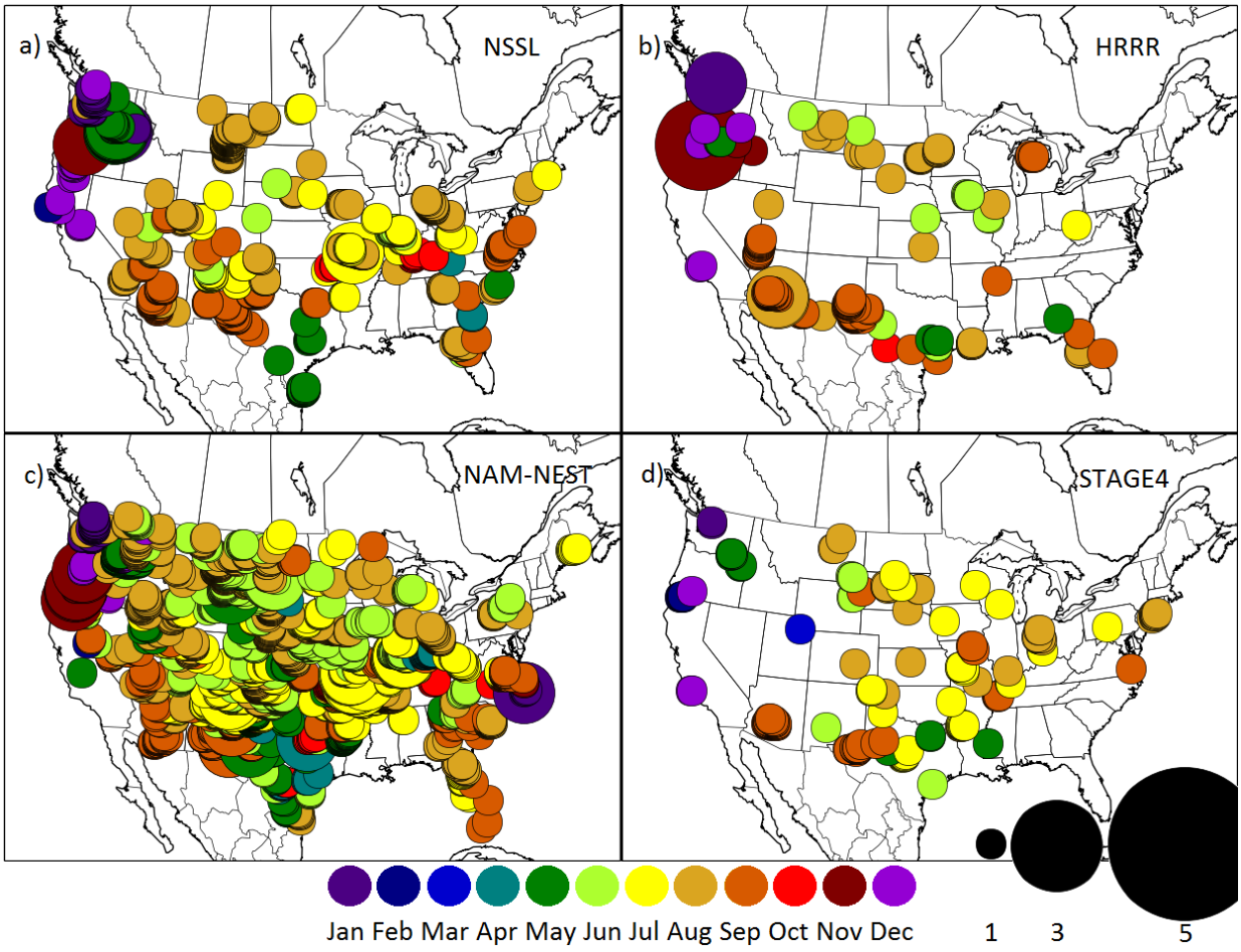

Figure 3.6: As in Figure 3.5, but for the 50-year return period thresholds.

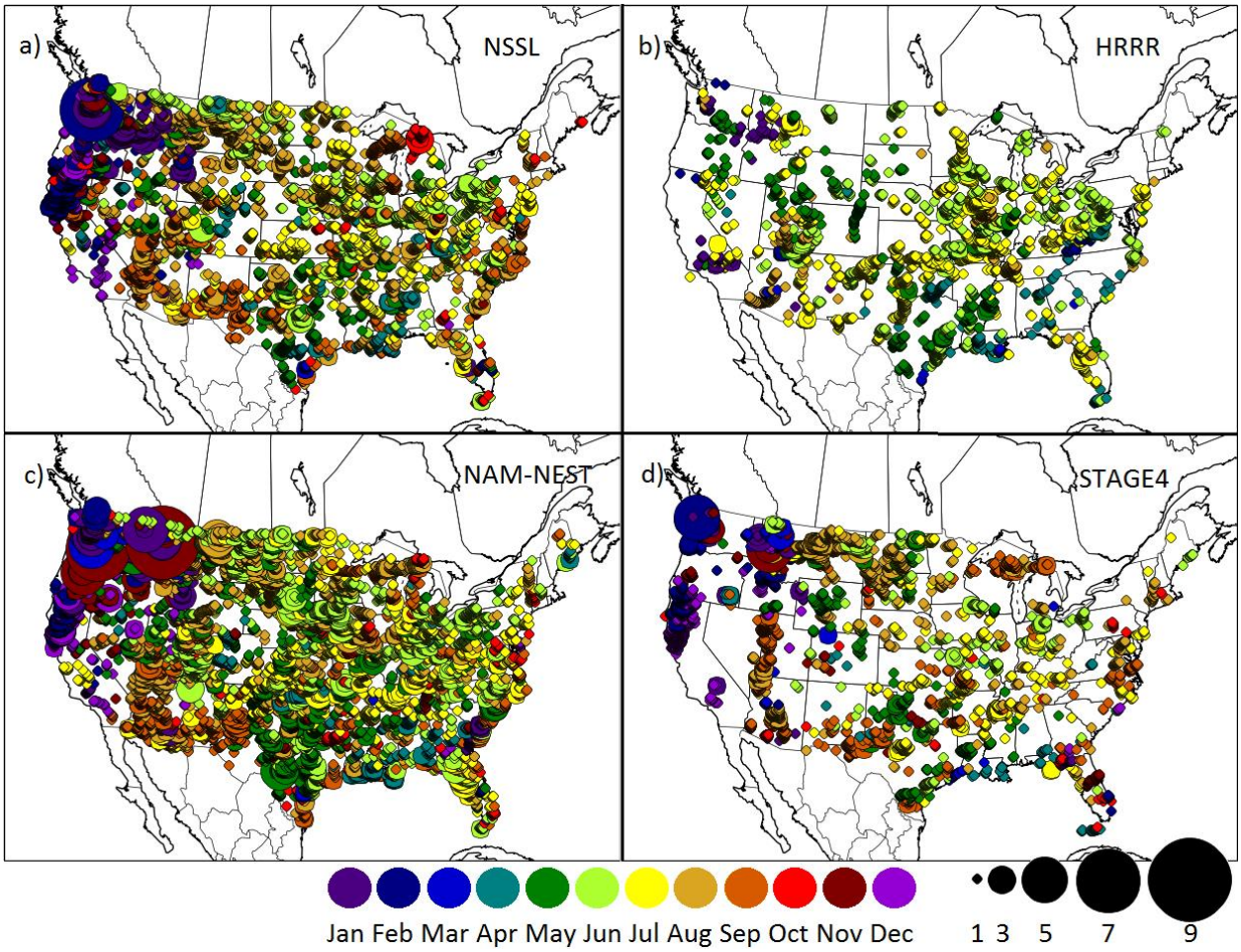

Figure 3.7: Same as Figure 3.3, but for the 12-18Z period. NSSL-WRF and NAM-NEST forecasts are taken from the 12-18 hour precipitation forecast from the 00Z initialization. HRRR forecasts are taken from the 0-6 hour forecast from the 12Z initialization.

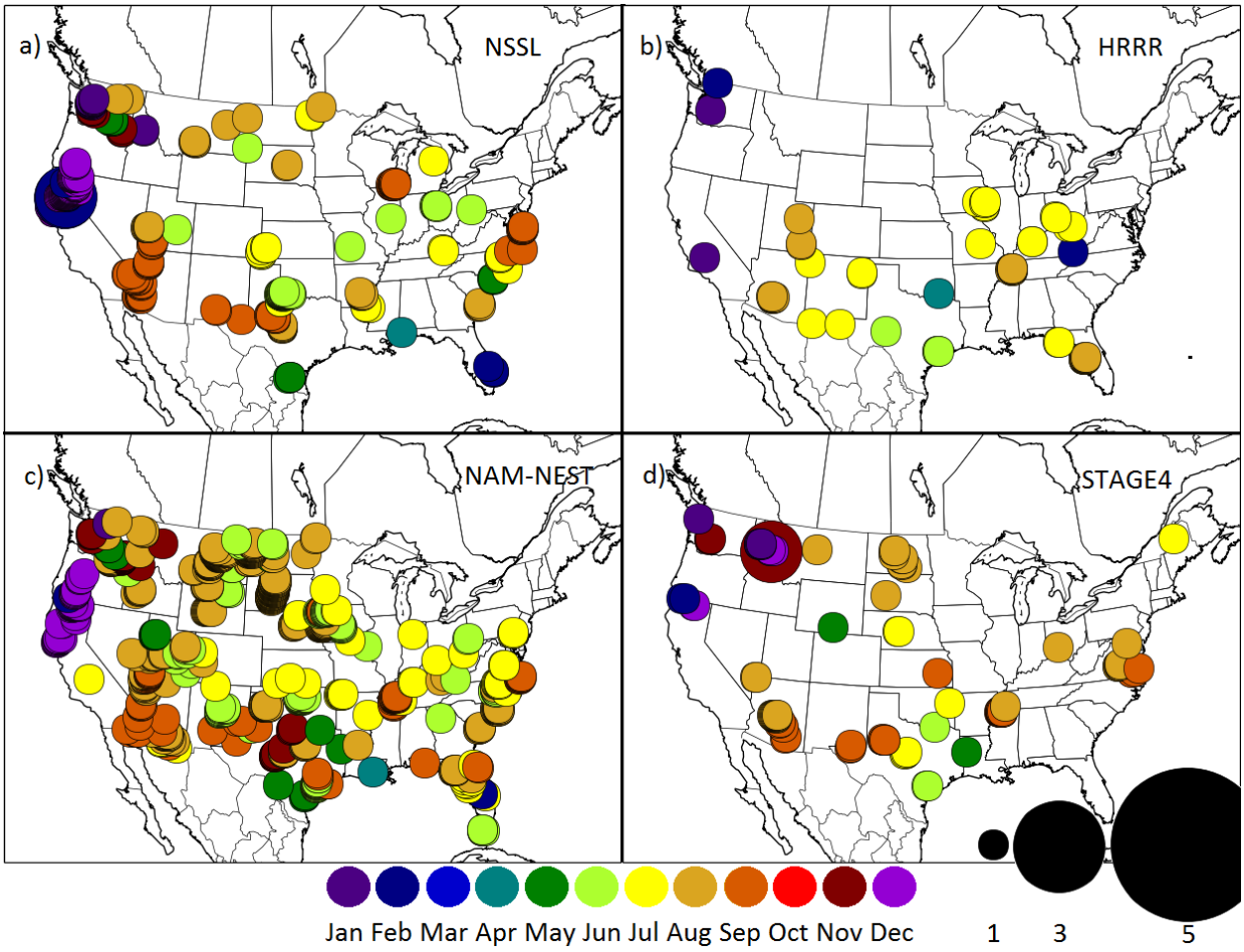

Figure 3.8: Same as Figure 3.7, but for 50-year return period thresholds.

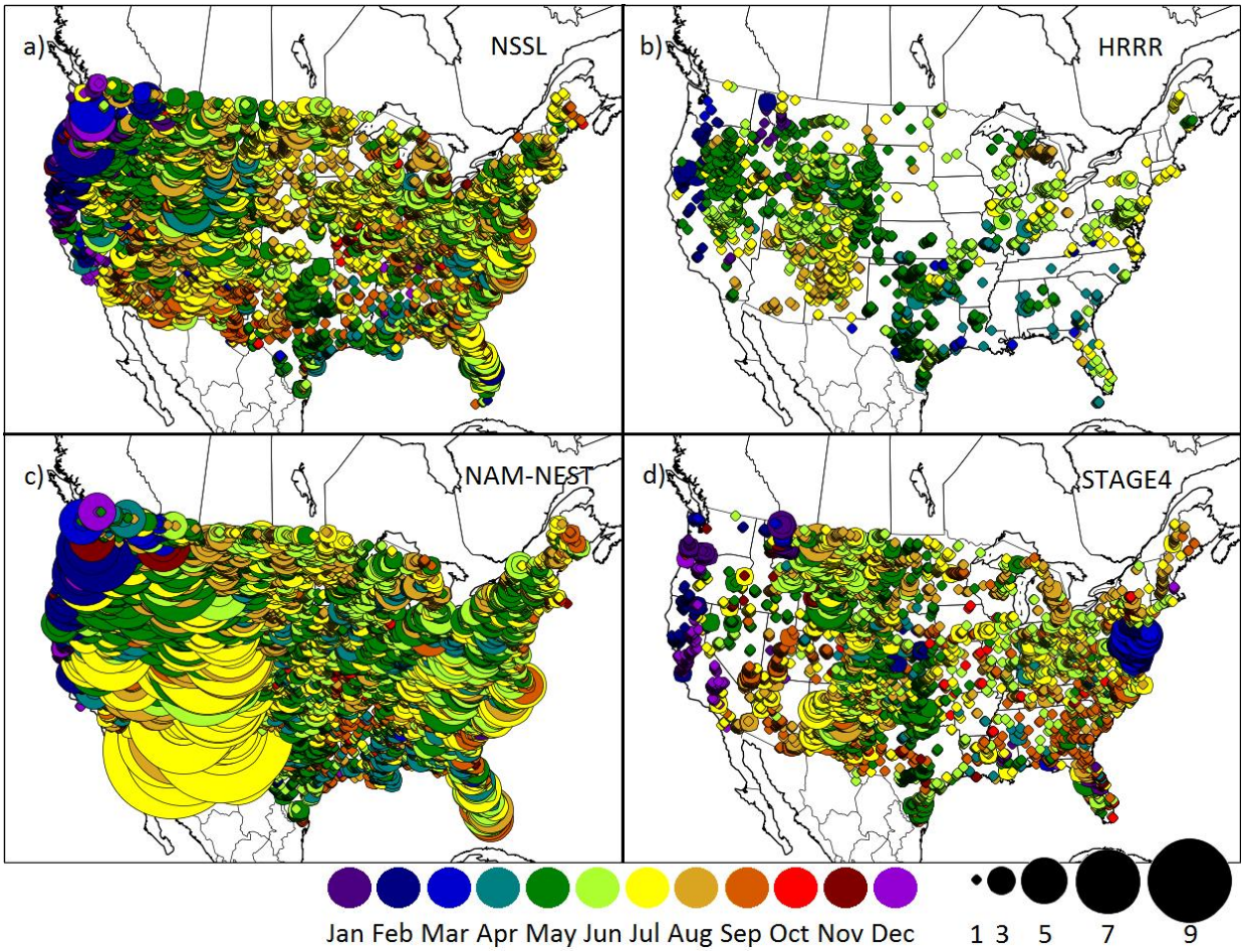

Figure 3.9: Same as Figure 3.3, but for the 18-00Z period. NSSL-WRF and NAM-NEST forecasts are taken from the 18-24 hour precipitation forecast from the 00Z initialization. HRRR forecasts are taken from the 6-12 hour forecast from the 12Z initialization.

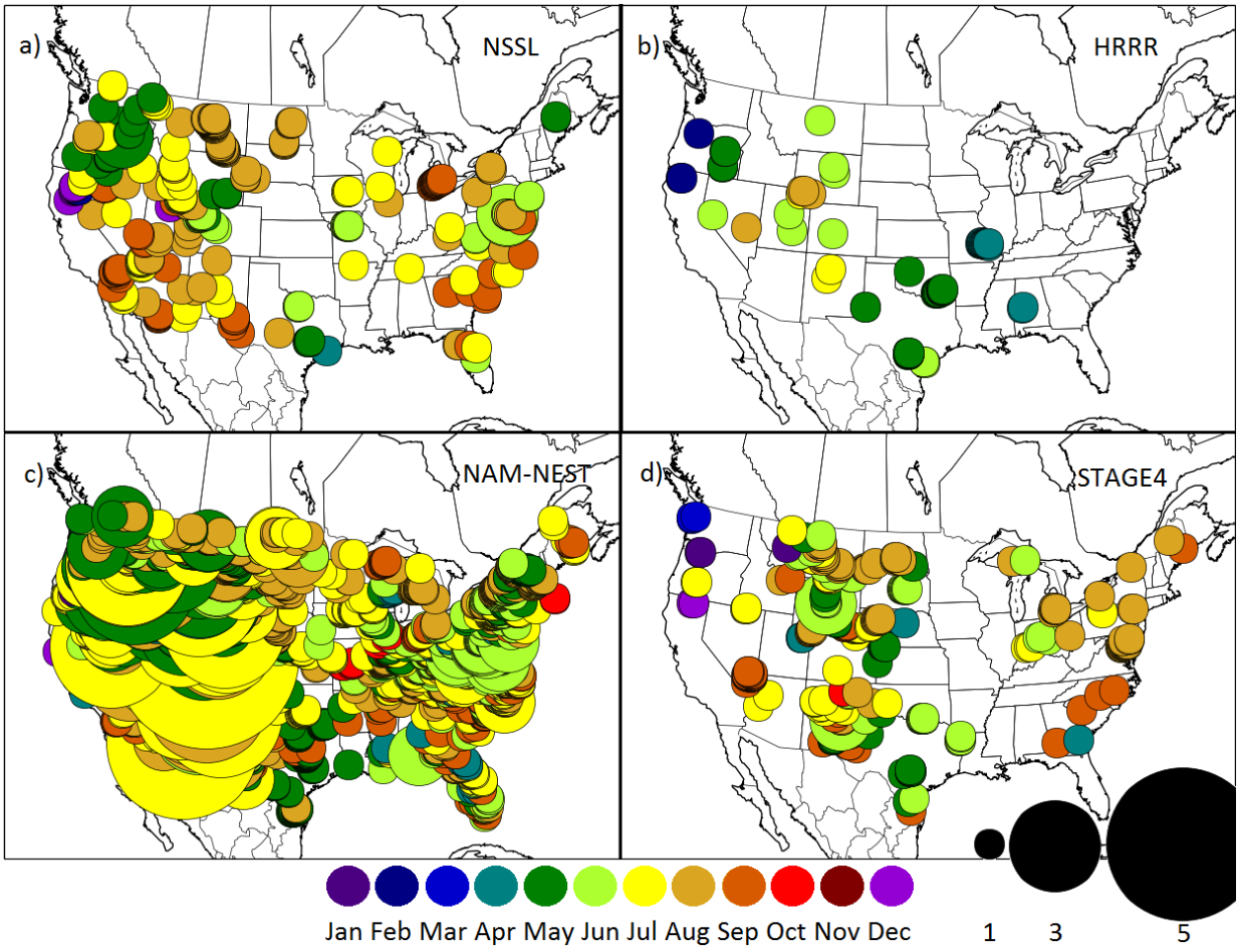

Figure 3.10: Same as Figure 3.9, but for 50-year return period thresholds.

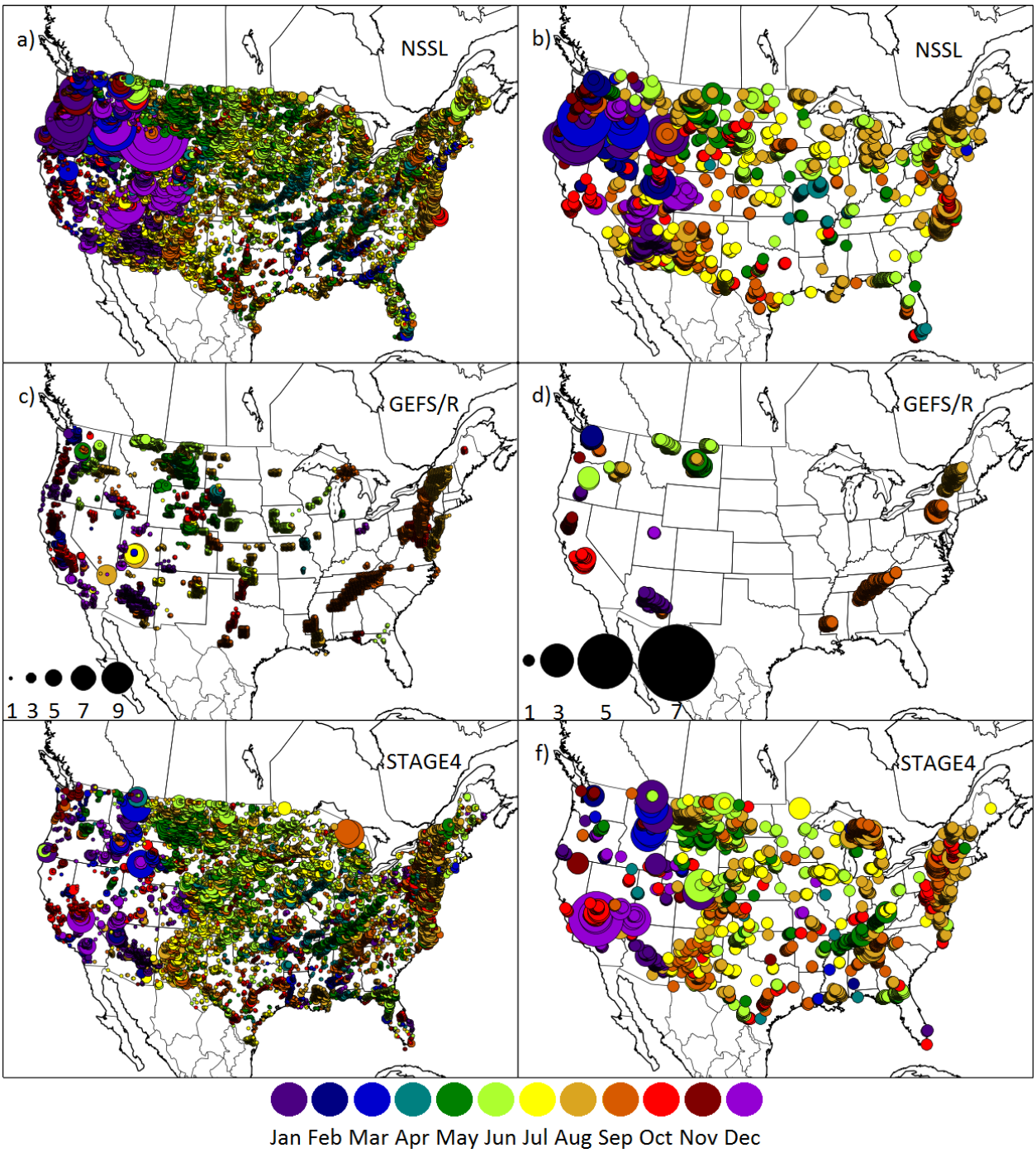

Figure 3.11: Forecasted and observed events of exceedances of two return period thresholds for a 24-hour accumulation interval, as illustrated in Figure 3.1(d) and (g), over the period from 09 June 2009 through 30 August 2014. Circles indicate an observed or forecasted event at the location of circle center; circle size is proportional to number of events, with a larger circle indicating more events at that location. Panels (a) and (b) correspond to forecasted events from the NSSL-WRF 12-36 hour precipitation accumulation from the 00Z initialization at the 10- and 100-year return period thresholds, respectively. Panels (c) and (d) correspond to 12-36 hour forecasts from the 00Z initialization of the GEFS control member, again for 10- and 100-year return periods, respectively. Panels (e) and (f) correspond to observed exceedances of the local 10-year and 100-year thresholds 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.

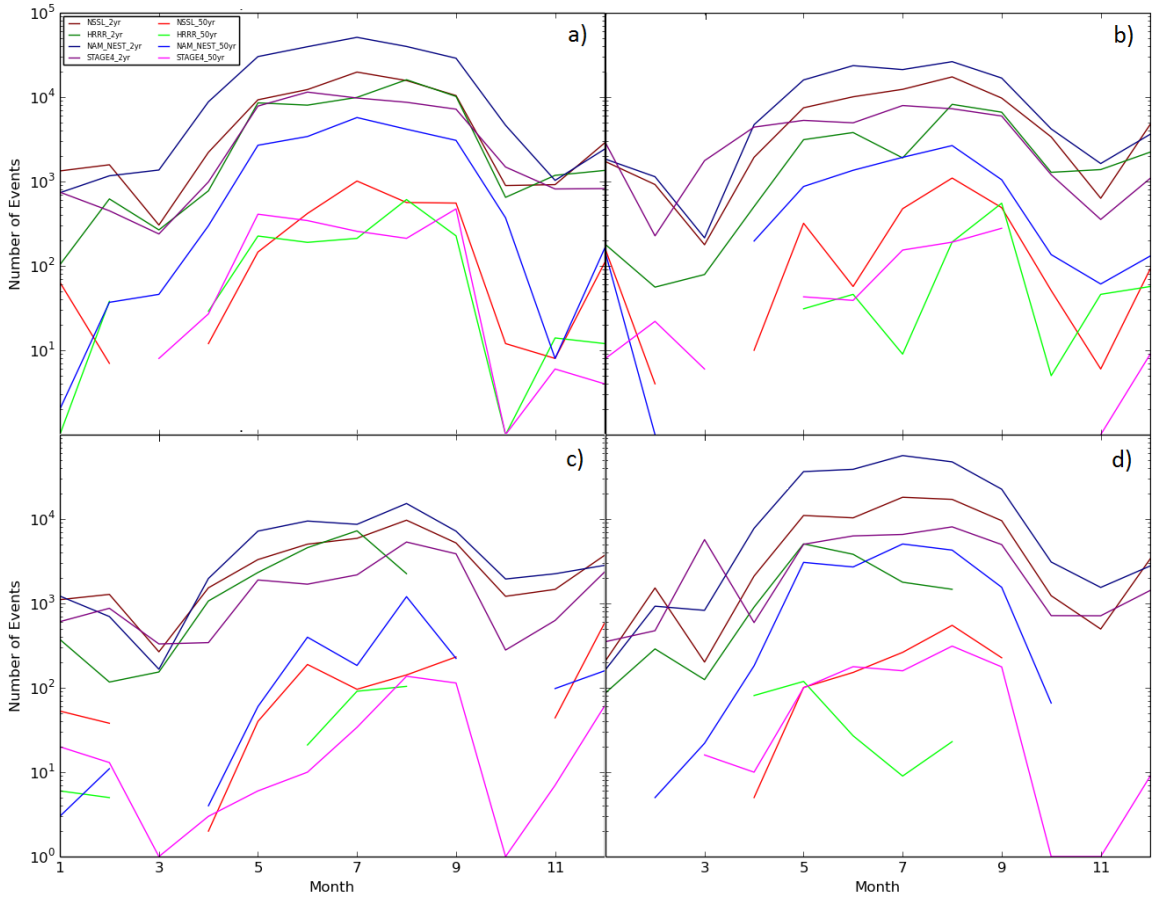

Figure 3.12: Total number of events forecasted or observed over the 12 August 2014-11 August 2015 verification period for 6-hour precipitation accumulations for the NSSL-WRF, HRRR, and NAM-NEST models compared against Stage IV Precipitation Analysis. Counts for the 00-06Z time of day is plotted in panel (a), 06-12Z in panel (b), 12-18Z in panel (c), and 18-00Z in panel (d). Event count is plotted on a logarithmic scale; return periods of 2 and 50 years are shown for each data source for each 6-hour accumulation period. 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. Significant amounts of HRRR data are missing from the verification period; HRRR event counts have been naively rescaled in proportion to the number of missing dates in each respective month.

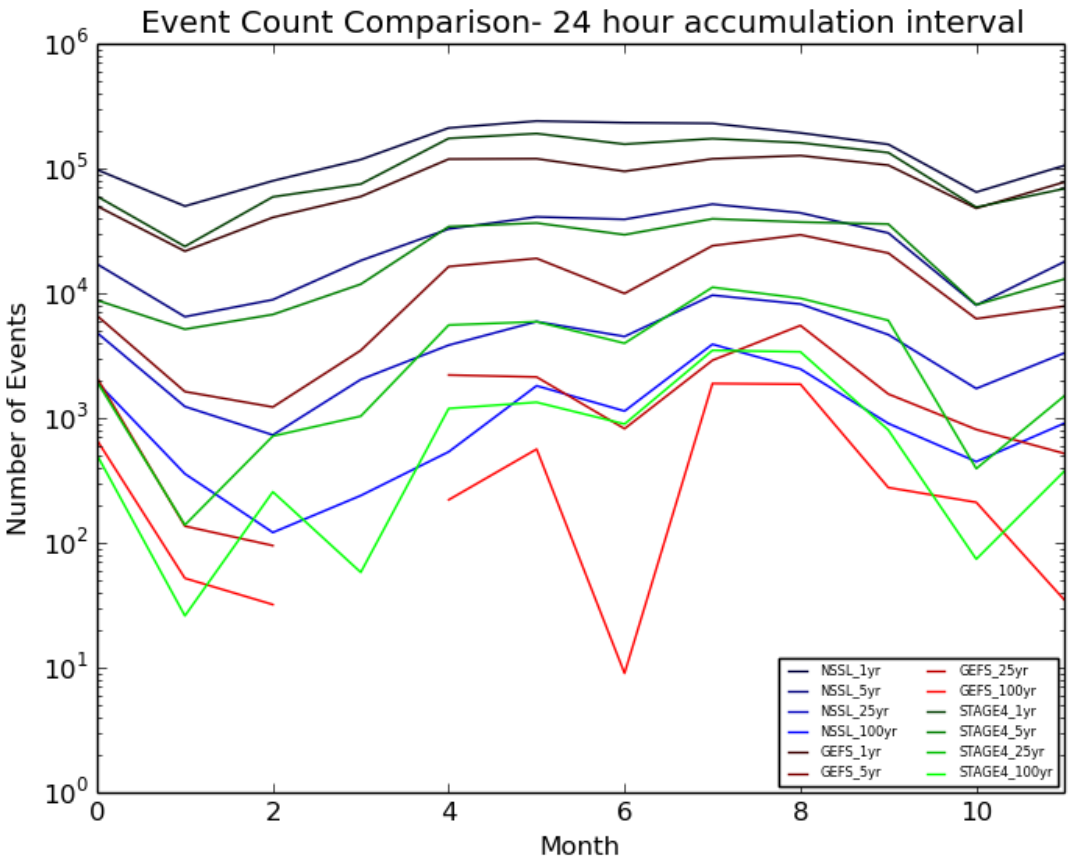

Figure 3.13: Total number of events forecasted or observed over the 06 June 2009-30 August 2014 verification period for 24-hour 12-12Z precipitation accumulations for the NSSL-WRF and GEFS models compared against Stage IV Precipitation Analysis. Event count is plotted on a logarithmic scale; return periods of 1, 5, 25, and 100 years 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.

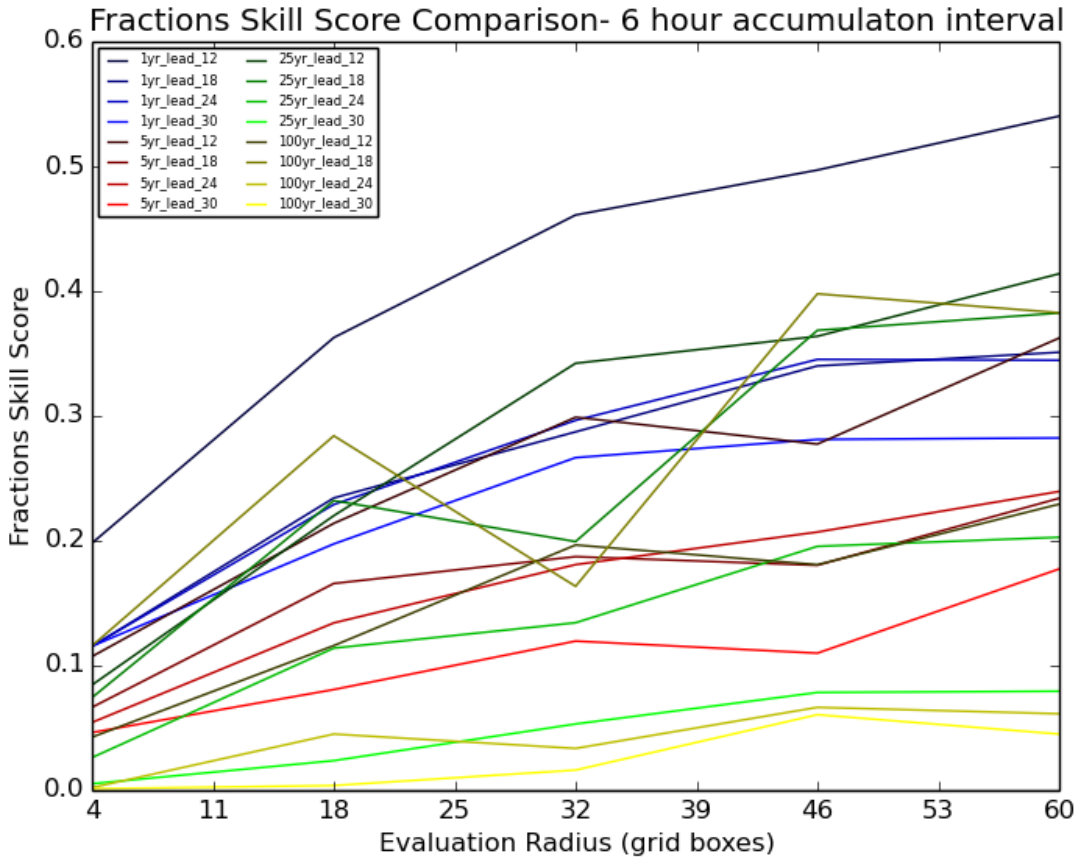

Figure 3.14: Aggregated Fractions Skill Scores for the NSSL-WRF for the 6-hour accumulation interval for the 1-, 5-, 25-, and 100-year return periods. Verification is performed over the 09 June 2009-30 August 2014 period. Forecasts taken from the 00Z initialization; ergo, lines indicated in the legend to have a lead of 12 correspond to the 12-18Z forecast period, leads of 18 to the 18-00Z period, 24 to the 00-06Z period, and 30 to the 06-12Z period.

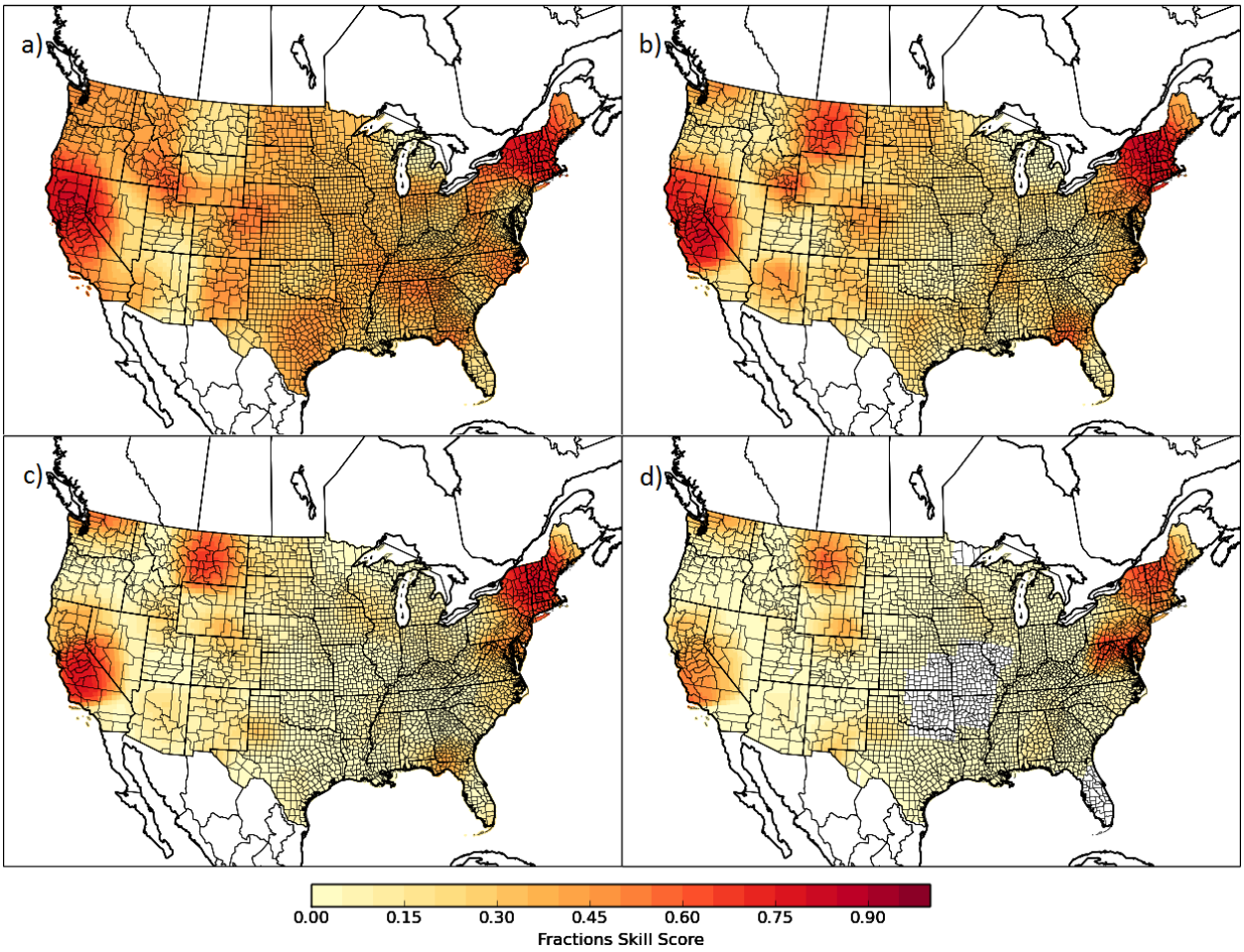

Figure 3.15: Gridded Aggregated Fractions Skill Scores for the NSSL-WRF for 6-hour precipitation forecasts aggregated over each of the 12-18, 18-24, 24-30, and 30-36 hour forecast periods for the 00Z model initialization. Panel (a) corresponds to verification on the 1-year return period thresholds, (b) to the 5-year return period threshold verification, (c) to 25-year return period verification, and (d) to 100-year return period verification. Verification is performed over the 09 June 2009-30 August 2014 period. Fractions Skill Scores on plots shown correspond to an evaluation radius of 40 grid boxes on the Stage IV HRAP grid.

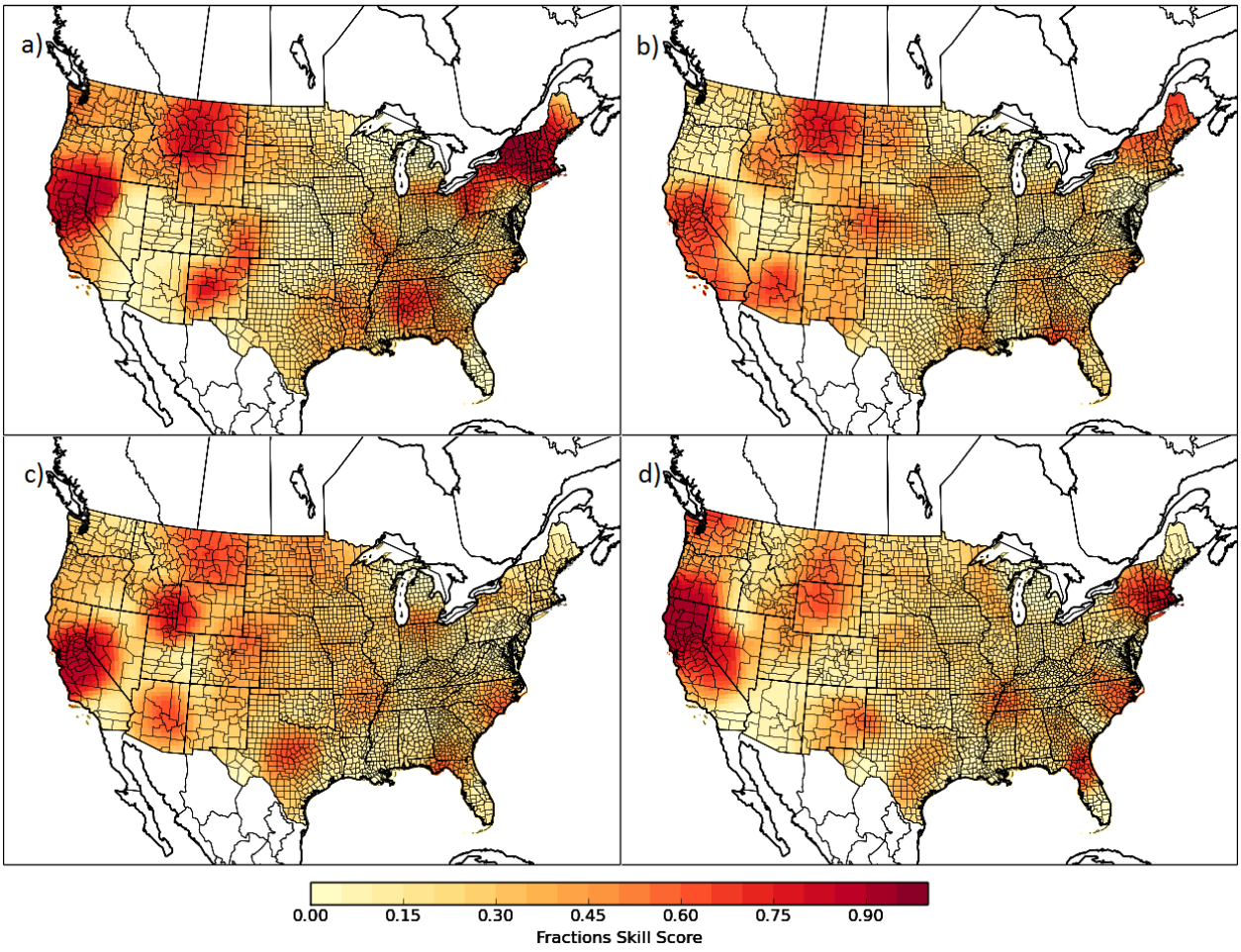

Figure 3.16: Aggregated Fractions Skill Scores for the 00Z initialization of the NSSL-WRF for 6-hour accumulated precipitation forecasts verified for 2-year return period thresholds. Panel (a) corresponds to verification over forecast hours 12-18 (12-18Z), (b) to 18-24 (18-00Z) hour forecasts, (c) to hours 24-30 (00-06Z), and (d) to hours 30-36 (06-12Z). Verification is performed over the 09 June 2009-30 August 2014 period. Fractions Skill Scores on plots shown correspond to an evaluation radius of 40 grid boxes on the Stage IV HRAP grid.

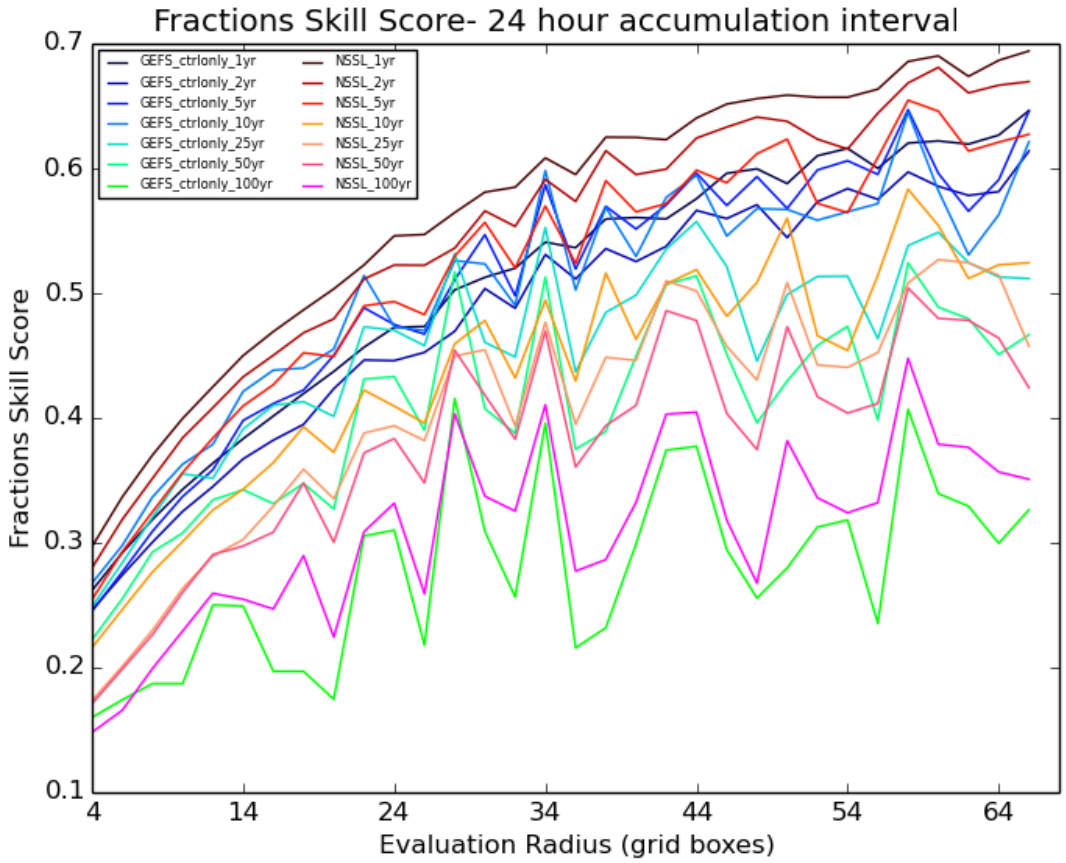

Figure 3.17: Aggregated Fractions Skill Scores for the GEFS and NSSL-WRF for the 24-hour accumulation interval for the 1-, 2-, 5-, 10-, 25-, 50-, and 100-year return periods. Verification is performed over the 09 June 2009-30 August 2014 period. Forecasts taken from each model's 00Z initialization.

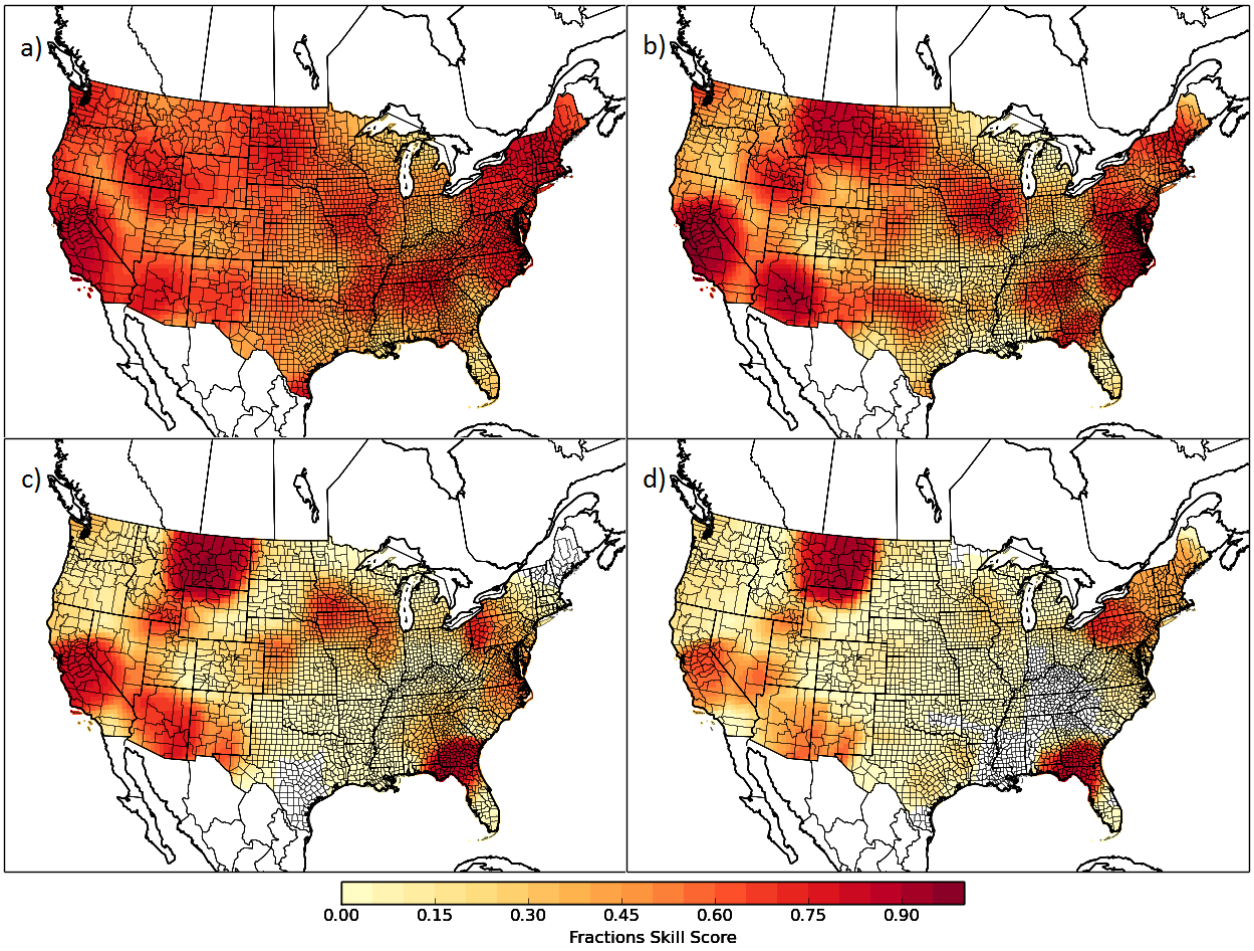

Figure 3.18: Gridded Aggregated Fractions Skill Scores for the NSSL-WRF for 24-hour precipitation forecasts from the 12-36 hour forecasts of the 00Z model initialization. Panel (a) corresponds to verification on the 1-year return period thresholds, (b) to the 5-year return period threshold verification, (c) to 25-year return period verification, and (d) to 100-year return period verification. Verification is performed over the 09 June 2009-30 August 2014 period. Fractions Skill Scores on plots shown correspond to an evaluation radius of 40 grid boxes on the Stage IV HRAP grid.

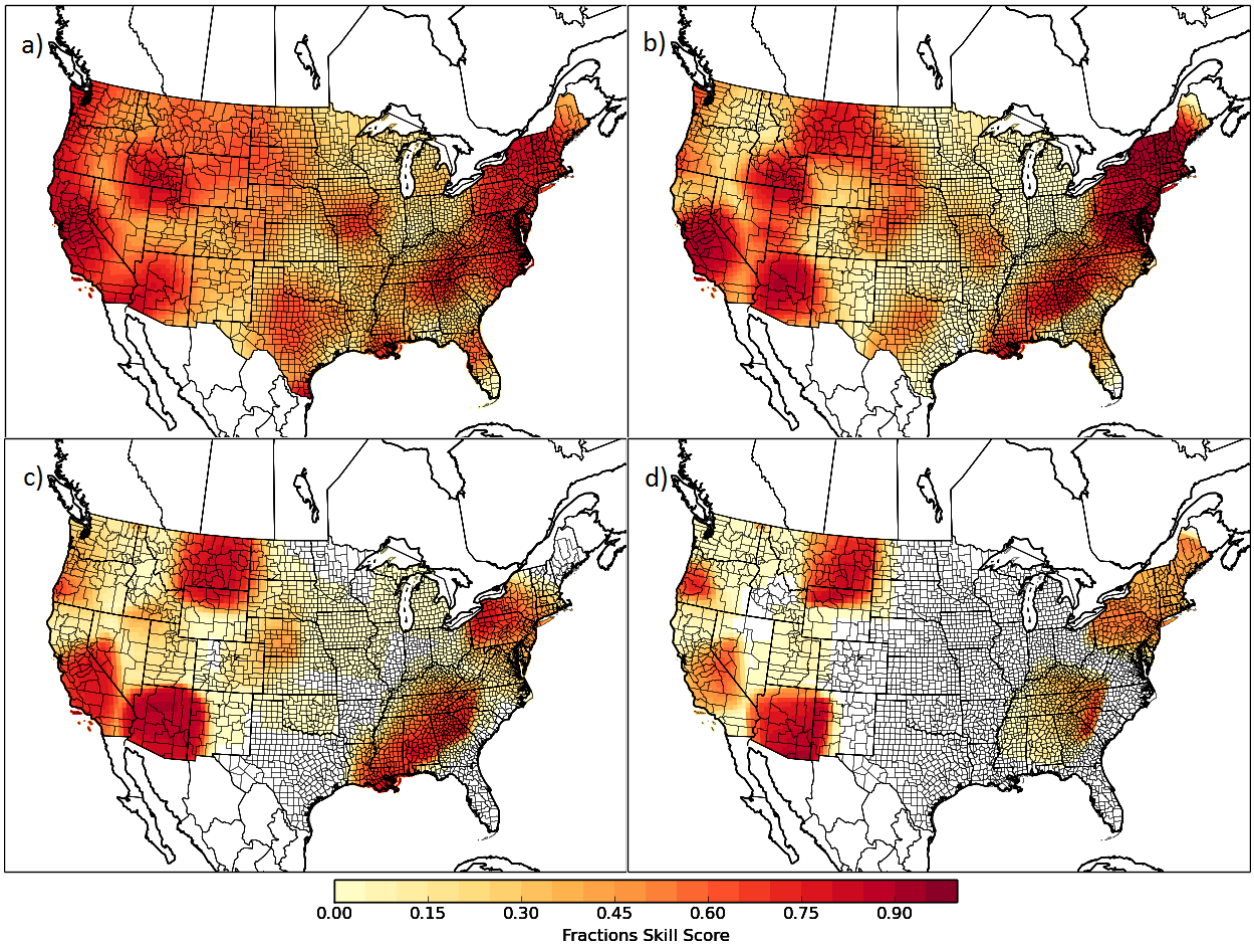

Figure 3.19: Same as Figure 3.18, but for the GEFS.

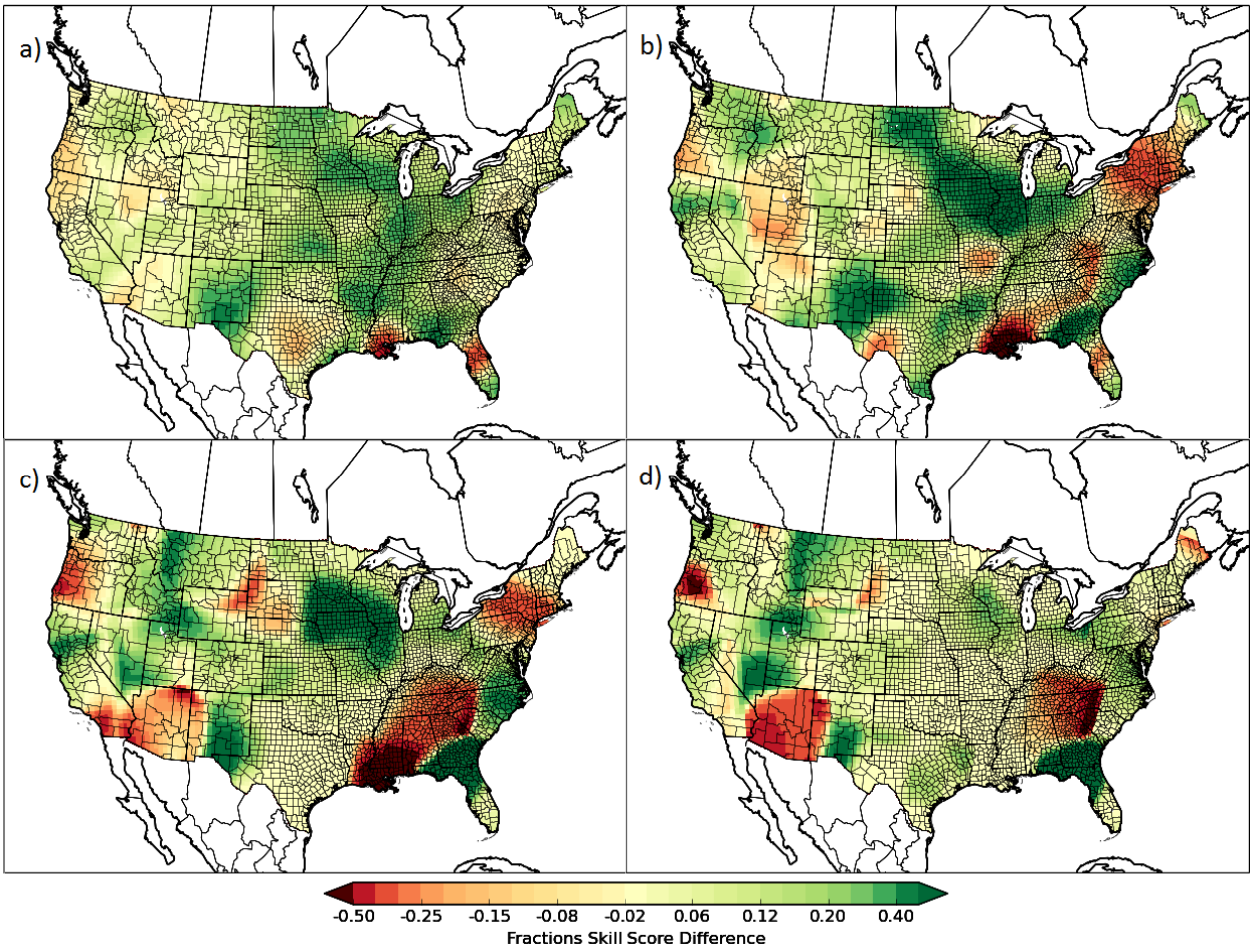

Figure 3.20: Same as Figure 3.18, but instead shows the difference between NSSL and GEFS performance over the verification period. Greens indicate that the NSSL-WRF performed better over the region, while reds indicate that the GEFS performed better.

3.3 Discussion & 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. CONUS-wide RP-threshold grids for 6- and 24-hour AIs were assembled from existing observational estimates for RPs between 1 and 100 years. 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 forecast 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-year period forecasted an order of magnitude more events than were actually identified via Stage IV precipitation analysis. Of the four 6-hour accumulation periods centered about 00Z, the NAM-NEST was found to exhibit the largest bias during the convective initiation period of 18-00Z, and the least in the minimally convective 12-18Z period. Other CAMs- the NSSL-WRF and HRRR- exhibited similar tendencies, also tending to overforecast extreme events from 1-year to 100-year 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-hour AI, the NSSL-WRF was found to exhibit rather similar extreme precipitation characteristics to those seen in the 6-hour AI. The coarser GEFS/R, however, had much different characteristics than the CAMs; GEFS/R produced 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 regards to model skill, models are unsurprisingly found to make most skillful prediction of the less extreme (lower RP) events, with the worst skill typically observed in forecasting for the rarest of event thresholds. For 6-hour periods, the 12-18 hour period stood out as the most skillful period for the NSSL-WRF, 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-hour AI; at low RPs, locally higher skill was generally seen in both models in the west and Mid-Atlantic/New England, and the NSSL-WRF demonstrated superior forecast skill over most of CONUS, particularly over the Great Plains and

Midwest, and 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. In a bulk sense, however, the NSSL-WRF was still found to verify statistically significantly better than the GEFS/R, even at high RPs.

4 Developing Model Precipitation Climatologies

4.1 Methods

Fitting model RP thresholds (RPTs) is performed here through a many step process. As in chapter 3, all model data is first regridded onto the same grid- the Stage IV HRAP ~4.75 km grid using first order conservative regridding. Using the model's threshold training record, RSDs are fit point-by-point to the model precipitation record. Three types of distribution fitting- AMS, PDS, and DF-FDS- are applied and assessed. Throughout this study, cross validation is used to assess algorithm validity. The verification period is split into four consecutive chunks; three of the four chunks are used for model training, and the remaining quarter is used for verification. This process is repeated four times in order to apply cross-validation verification over the entire verification period.

This study applies model climatology fitting methods to two models: 1) the NSSL-WRF, and 2) the control member of GEFS/R. There are three periods to note. The verification period for both models is 09 June 2009 to 30 August 2014. However, GEFS/R has a model precipitation record, or threshold training record- extending back to 01 December 1984. This also allows GEFS/R training to take advantage of the extended Stage IV period: 01 January 2002 to 08 June 2009. The exact use of this extended period will be elucidated below. The verification period, which is also the NSSL-WRF threshold training period, amounts to 1908 days; ¾ of this- the amount used for threshold training in cross-validation- results in 1431 days, or just under four years. The challenge is how to use such a relatively short training record to extrapolate events more than an order of magnitude rarer than the training record length, with no especially rare events occurring at all at many locations in the record.

AMS, PDS, and DF-FDS RP thresholds are fit for the 2-, 5-, 10-, 25-, 50-, and 100-year RPs for ten RSDs: Exponential (EXP), Gamma (GAM), Generalized Extreme Value (GEV), Generalized Logistic (GLO),

Generalized Normal (GNO), Generalized Pareto (GPA), Gumbel (GUM), 4-parameter Kappa (KAP), Pearson Type 3 (PE3), and Weibull (WEI). Parameter estimation for all distribution fits is performed using the Method of L-Moments (MoLM). A comparison between AMS, PDS, and DF-FDS fits is shown in Figure 4.1 for some select locations. Given that the predictand of interest of in this study, daily exceedance forecasts with respect to average recurrence intervals (ARIs) or return periods as opposed to annual exceedance probabilities (AEP), PDS/DF-FDS based analysis more directly addresses the pertinent forecast question when compared with AMS analysis, which would require numerical correction to the PDS framework. From this, and by manual subjective inspection of the distribution fits (presented in part in Figure 4.1), PDS-based fits were found be the most appropriate and sensible fits to model QPFs, and were selected for further investigation (DF-FDS fits were found to be very unrealistic for many RSDs at many locations). Two aspects that were explored in further detail were the PDS cutoff threshold and block size, traditionally one year in association with AMS applications. Having only four years of training data in the case of the NSSL-WRF makes the one-year block size rather problematic, as distribution fits may be based on as few as four data points. To alleviate this problem, a block size of ½ year was explored, with every eighth day being included in a given block so as to alleviate block maxima being biased by inclusion or exclusion of a particular climatologically favored season. This would allow a minimum of eight data points in each distribution fit, ostensibly appreciably reducing model variance, while also allowing for a potentially much more realistic estimate of the 1-year RPTs. Ultimately, the ½-year alternating block was selected for the NSSL-WRF data and the traditional 1-year annual block was selected for GEFS/R distribution fitting; this distinction was made based on the 4-year threshold training period for the NSSL-WRF vs. the 28+ year period for the GEFS/R. The inclusion cutoff threshold for each PDS series was the minimum block maximum (*block minimax*) using each model's respective blocks.

In spite of these efforts to reduce fit *variance* in the NSSL-WRF, the combination of the short data record and convective allowing capability producing local, small-scale convective precipitation features results

in unrealistic model RPT estimates. The latter problem was combatted by applying threshold smoothing. Smoothing RPT fields requires a balance between eliminating spurious highs and lows associated with where convective cells happened to be simulated in the model with the legitimate differences in the climatology of the model. This is especially true in areas of complex terrain where the climatology may change sharply and significantly over short distances. A smoothing method was designed in an attempt to balance these two goals: progressive point by point smoothing was applied such that no gradient between adjacent grid points ever exceeds five times the gradient of the respective gradient witnessed in the observationally-derived Atlas thresholds for the corresponding RP. This preserves realistic gradients in the RPTs in association with complex terrain, since those gradients appear in observational thresholds as well, while eliminating the spurious features associated with coincidental placement of convective cells. Additionally, the challenges presented by the short threshold training data record are alleviated through a modified regionalization technique. Owing to the short data record, not only may extreme events not occur at a given grid point, but may not occur anywhere in the vicinity of such a grid point. Deducing extremes in such a situation then requires either exceptional extrapolation from common events, or use of information about extreme events occurring at remote locations. The extrapolation required in the former case is not considered to be practically obtainable; instead, an assumption is made to allow sensible use of remote extreme precipitation events when possible. Namely, distribution fits at two locations that are similar in the observational thresholds are assumed here to also be similar at the corresponding locations in the model thresholds. For each grid point, a list of grid points with similar distribution parameters (measured here as within 2 mm for the 1-year RPT with tolerance increasing progressively with increasing RP up to a tolerance of 6 mm for the 100-year RPT) is constructed. Often, the vast majority of these points are local to the grid point in question, but occasionally spatially distant points satisfy these criteria as well. This assumption, while it may not hold perfectly, allows for a method to artificially increase sample size and thereby

reduce the variance of the algorithm. For each grid point, the fits corresponding to each similar grid point are averaged to form RPT estimates.

One metric to quantify the algorithm's variance is to plot the coefficient of variation (CoV) of the model climatological RPTs between the four individual cross-validation threshold trainings. This is shown in Figure 4.2. CoVs for the NSSL-WRF model RPT estimates for the 100-year RP from the EXP distribution fitted based solely on the local point data (2a) are generally quite high with many locations experiencing a CoV in excess of 0.2; this is reduced considerably through the smoothing process (2b), and reduced even further through the similar points regionalization (2c). The smaller CoVs, a measure of mean-relative variability, indicate more stable RTP estimates that are less sensitive to the subsample of data on which they are trained, suggesting an algorithm less prone to overfitting and thus hopefully experiencing less generalization error. In comparison, the GEFS/R model climatology CoVs with no smoothing or regionalization are shown in Figure 4.2d. As can be seen comparing 2c and 2d, the CoVs are lower than those of the NSSL-WRF even after smoothing and regionalization, illustrating the side effect of the incapability of the underlying dynamical model to resolve convective elements and, perhaps more significantly, the added benefit of the greatly extended- seven to eight times longer- threshold training period.

Once candidate RPT estimates are obtained through RSD-fitting the model precipitation data, it is next explored to what extent, if at all, the fitted model climatological thresholds may be applied towards improved deterministic predictions of locally extreme rainfall. The question is not trivial since there exists a disconnect between the RPT estimates and the prediction thresholds that maximize model skill. The RPT estimates seek to accurately address the question: "How rare is it for the given model to forecast a precipitation accumulation of X at location Y?", or more directly "What is the minimum precipitation accumulation forecast X at location Y in the model in order to occur with a long-term

average frequency of once every Z-years?" Even a perfect estimate of this amount doesn't necessarily correspond to the threshold which maximizes skill at predicting observed events of the same frequency, however. A coarse model like the GEFS/R, for example, may fail to forecast hardly any of the observed extreme events in regions such as the US Great Plains where extreme precipitation events are almost exclusively convective. The heaviest accumulated precipitation forecasts over this region may all be associated with synoptic systems, with forecast associated with the highest precipitation totals in reality receiving very small accumulation forecasts. In this case, using the true model 100-year RPTs may yield very poor forecast skill, since the effect will be to incorrectly forecast 1-5 year RP event-creating synoptically driven systems as 100-year events, while still failing to forecast the true 100-year events, which are convectively driven, as extreme events at all. For this reason, it is not guaranteed that good RPT estimates will enhance forecast skill, but given the great value of even small improvements in forecast skill in the extremes, the question is still very worthy of investigation.

Given this disconnect, why use RSD fits for determining the most skillful thresholds at all? Considering the events being forecasted are often an order of magnitude or more rarer than the data record length, it is essential to preserve an appropriate extrapolation from the common events- those one can reasonably anticipate to observe within the training record- to the rare events that one can not anticipate observing in the training record. RSD fits provide a theoretically sound relationship from the low RPs to the high RPs and by selecting an RSD, rather than individual RPTs, for a given location, one preserves a logical relationship between the RPT estimates across the RPs. There is also substantial research (e.g. Coe and Stern 1982, Wilson and Toumi 2005, Bonnin et al. 2004, Hershfield 1961, Miller et al. 1973) in the literature exploring appropriate RSD fits to observed precipitation data. Different studies reach different conclusions about the most appropriate RSD to use for fitting accumulated precipitation, and a synthesis of the literature suggests that it is likely that different RSDs are appropriate over different regions of CONUS.

Recognizing these properties, each RSD fit, in addition to the observationally-derived Atlas thresholds, is considered a candidate set of RPTs; the goal is to select the RPT set (RSD fit) which maximizes forecast skill. In order to do this, each of the four sets of RSD fits corresponding to the four cross-validation sections is validated over the three quarters in which they were trained. In the case of the GEFS/R, since the necessary model data is available, the extended verification period- from 01 January 2002 to 08 June 2009- is employed for this stage as well. The local Fractions Skill Score (FSS) is computed at each grid point aggregated over each quarter. Then, for a given cross-validation RSD set, the RSD fit which maximizes the global FSS is selected:

$$DIST(y,x) = k\left(\text{argmax}_k\left(1.0 - \frac{FBS_{TOT} + FBS\left(DIST_k(y,x)\right)}{FBSWORST_{TOT} + FBSWORST\left(DIST_k(y,x)\right)}\right)\right)$$

It should be noted that this is distinct from simply selecting the distribution corresponding to the maximum local FSS. Assume that, absent the global FSS contribution from point (y,x), the FBS_TOT from other points is 79 and FBSWORST_TOT from other points is 99. One distribution at (y,x) produces a verification with a local FBS of 1 and local FBSWORST of 1, for a local FSS of 0. A second distribution produces a verification with a local FBS of 10 and FBSWORST of 11, for a local FSS of 0.091. But the global FSS by inclusion of the former distribution is 1.0 – (80/100) = 0.2, versus the latter distribution, which has a higher local FSS, yields a lower global FSS: 1.0 – (89/110) = 0.191.

This process was conducted at each point to determine locally most skillful RSDs and, by association, RPTs for all RPs of interest. These are then validated for each quarter of the verification period applying the identified best RSDs based on the thresholds derived and tested without the use of the same validation quarter. So doing provides a conservative estimate for how well the method will extrapolate to future forecasts beyond the validation period; the estimate is likely 'conservative' because the amount of training data used here in cross-validation is considerably less than what would

be available for real-time forecasting, and it is believed that increased training data may substantially improve algorithm performance. Many algorithmic details were also determined via cross-validation verification. The RP or RPs to use for FSS maximization and best RSD fit are examined; verification may be performed over a single RP, over all available RPs, or over some chosen subset. The FSS evaluation radius to use may also be tuned; similar to the RP selection, a combination of evaluation radii may also be applied, with the best FSS being determined based on an average of results corresponding to verification over a single evaluation radius. It can also be noted by inspection of the quantitative Stage IV event values in Figure 3.13 that the number of observed events for each RP over the verification period is roughly one half to two fifths what one would expect to observe *a priori* based on the number of grid points, the length of the verification period, and the frequency definition of the return period. This can likely be primarily attributed to a combination of two factors. First, Stage IV analysis is gridded, and does not directly correspond to point estimates. The smoothing inherent in upscaling from point observations to a 4.75km grid results in less occurrences of extreme precipitation amounts, particularly in instances where extreme precipitation is occurring at highly local spatial scales. Second, the limited radar coverage and sparsity of rain gauges in areas of complex terrain- particularly in the intermountain west- lead to an underestimate of the number of observed events in those areas. Accordingly, model-derived RPTs for a given RSD fit may more appropriately be taken from the threshold corresponding to an RP approximately twice the given observational RP. These thresholds will be termed “distribution offset thresholds”, and will be abbreviated with an ‘O’ suffixing the RSD abbreviation. The inclusion and exclusion of these thresholds from consideration is explored. In order to avoid an overfitting model, a tunable ‘regularization’ term was added to the model. The idea behind this term is to not attempt to adjust in accordance with a small signal that may be attributable to noise in the sampling data. As such, applying the regularization term, the algorithm will not suggest an adjustment at a given point unless the selection of the distribution improves the global FSS when compared against the use of the Atlas

threshold at that point by at least a specified regularization threshold R. The last major factor explored in RSD selection is the use of regions. The short data record makes determination of the most appropriate RSD at a single point rather difficult to accurately discern. Use of neighboring points within a similar geographic/topographic/meteorological region allows for an expansion of the data record used for determine the most appropriate RSD fit. Applying a single RSD fit to an entire region with similar meteorological characteristics also makes sense in the sense that the true distribution family of precipitation is likely the same for all points within a given region, but not necessarily across regions. Results for using regions as opposed to point-by-point analysis are compared and evaluated. CONUS regions are subjectively broken down into 27 regions exhibiting similar climatological characteristics, as depicted in Figure 4.3[3].

4.2 Results

An example of RSD fits near some select cities scattered across CONUS is presented in Figure 4.4 for the GEFS/R and 4.5 for the NSSL-WRF. Not too surprisingly, the observational thresholds are mostly much higher than the model-fit precipitation climatologies for GEFS/R; that is, less precipitation is required to be forecast in GEFS/R for the same frequency of occurrence. The one exception for the cities appearing here is near Phoenix, AZ (4.4e), where a few of the fitted distributions do exceed the Atlas thresholds. It is of note that the thresholds for all RSDs are lower even in Seattle (4.4a), where most locally extreme precipitation is stratiform and caused by large-scale forcing. The characteristics of the RSD fits exhibit some similarities across regions and even across models. All of the distributions produce very similar RPTs for the low RPs- 5-years or less. It should be noted that the values to which model RSD fits converge at low RPs often differs significantly from the Atlas-based thresholds (e.g. 4.4c). At the higher RPs, the distributions' RPTs diverge, and thus the choice of distribution becomes more

[3] The use of similar points, as described with threshold derivation at the beginning of this chapter, was considered too computationally expensive to realistically examine through cross-validation procedures here.

significant. The GLO and GEV distributions unequivocally yield the highest and second highest RPTs, respectively, at the higher RPs among the RSD fits. Often they are very similar (see 4.5b, 4.5f), though they do diverge in other places. On the other end, the GAM distribution almost always produces the lowest RPT estimates, followed by the GUM distribution, though in Phoenix for the NSSL-WRF (4.5e), GUM produced lower estimates than GAM. The behavior of the KAP distribution was perhaps the most erratic, but tended to produce the third lowest estimates, though occasionally at some higher thresholds (4.5c, 4.5f) was the lowest, and other times was among the highest (4.5e). The remaining five distributions- EXP, GNO, GPA, PE3, and WEI- tend to produce middling estimates. Among these, GNO and GPA tend to produce higher estimates, and EXP tends to produce the lowest RPTs. All distributions, with the possible exception of KAP, tend to produce precipitation/frequency curves of approximately similar shapes; KAP occasionally produces very sharp precipitation limits, where its estimates at low RPs follow a similar path to the "high estimate" thresholds, but at some point, frequency begins increasing rapidly with precipitation nearly constant (4.5c). In contrast, the Atlas thresholds often exhibit a very different shape, and RPTs from Atlas cross the model-derived RPTs at some high RP. Minneapolis (4.5c) and Houston (4.5f) in the NSSL-WRF illustrate examples of this.

Figure 4.6 provides CONUS-wide graphical comparisons of the distributions at the 100-year RP for the GEFS/R, and using point-by-point threshold estimates, absent smoothing or regionalization. One can see the particular reduction in model-climatological thresholds relative to the observational ones in the GEFS/R in the convective regions of the country, with the most and darkest reds seen to the east of the Rockies over the plains and Midwest. There appears to be the most variation in thresholds along the Atlantic coast, with the GAM distribution producing thresholds roughly 100mm or more lower than the observational thresholds along much of the coast, while the GLO is actually notably higher than Atlas in many scattered places along the coast. The NSSL-WRF thresholds appear in Figure 4.7 for the same RP. Most striking is the highly erratic nature of the RPT estimates using only point by point methods. One

location has an RPT threshold estimate 100 mm higher than Atlas; go 50 km away, and the model-estimated threshold is 100mm lower, and another 50 km and it's 100mm higher again. This unrealistic behavior is surely influenced by the location of convective elements in various forecasts throughout the training record. Figures 4.8 and 4.9 illustrate the corresponding smoothed thresholds for GEFS/R and NSSL-WRF, respectively. The effect of the smoothing on the GEFS/R fields may be seen comparing Figures 4.5 and 4.7, but the differences are not especially glaring. For the NSSL-WRF, however, the effect of smoothing is very considerable, resulting in a field of estimates that seem much more plausible, and largely uninfluenced by individual connective elements appearing in the training data.

In cross-validation, due to the very different characteristics associated with the two underlying dynamical models, it was judged that the two could very realistically have different optimal algorithmic parameters, and as such, the algorithm was tuned separately, rather than jointly, on each model. With respect to RPs to use for RSD selection, in general, middling RPs among those studied here tended to produce the best resolution at identifying skillful RSD fits. At low RPs, the RSD fit RPTs tended to be very similar, or at least clustered between the offset and non-offset fits, meaning all of the fits verify similarly at those thresholds and the method then lacks the resolution to distinguish the differences between the various fits. At the highest RPs, solutions tend to be very high variance due to the small sample size, deciding thresholds based on the performance of the underlying modeling system in a single event. Middling RPs tended to balance these challenges, as there is sufficient separation among the RSD fits to yield different verification results, but also sufficient events in the verification record to avoid being unduly swayed but model performance for a single event. For both models, the use of just the 10-year RP produced the best cross-validation results. Because GEFS/R thresholds were based on 1-year chunked PDS fitting, 1-year RPTs were not considered reliable and were not used in this study; however, the NSSL-WRF used a half-year block size, allowing for examination of the 1-year RPTs. Evaluation radii of 10, 20, 30, 40, and 50 grid boxes were examined, as well as using a weighted average of all five

evaluation radii. The use of only the 50 grid box radius produced the best cross-validation results for both models. No non-zero regularization threshold was applied in the NSSL-WRF; regularization improved GEFS/R results, and were maximized at a threshold of a mean global FSS improvement of 5e-7 per grid point. Both NSSL-WRF and GEFS/R included both offsetted thresholds and regional, as opposed to local, RSD applications. Results were most sensitive to RP set choice, and secondarily to evaluation radius; effects of other choices tended to be of second order importance.

The identified most skillful RSD fits for each iteration of cross-validation for the NSSL-WRF appears in Figure 4.10. Panel (a) is trained and evaluated from 29 September 2010 to 30 August 2014 and applied to 9 June 2009 to 28 September 2010, panel (b) is trained using 9 June 2009 to 28 September 2010 and 19 January 2012 to 30 August 2014, (c) uses 9 June 2009 to 18 January 2012 and 10 May 2013 to 30 August 2014, and (d) uses 9 June 2009 to 9 May 2013 and is applied to the final quarter. Since the forecast predictand is based on the Atlas thresholds, it is not terribly surprising that the Atlas thresholds are most often identified as the best verifying model. For an unbiased model, though one would expect different distributions to occasionally be identified as most skillful due to sampling noise, in expectation, one would always anticipate the Atlas thresholds being identified as the RSD fit of choice. The second most employed RSD fit is the highest threshold, the offsetted GLO, or GLOO distribution. It makes some sense that this is often identified, as GLOO almost uniformly produces the highest thresholds. When no events are observed, the highest thresholds will strictly dominate lower ones, since contribution to global FSS is then solely determined by false alarm rate. The one region that is consistently identified as needing adjustment is the Mid-Atlantic region. In the first, third, and fourth segments, the non-offsetted EXP distribution is identified as maximizing forecast skill. For the second segment, where Tropical Storm Lee is absent from the training data, higher thresholds associated with the GNO distribution are selected instead. Many of the adjustments from Atlas appear in three out of four segments, excluding the one segment where a major extreme rainfall occurred in the region.

Examples of this include the eastern Colorado flooding of September 2013, which occurred during segment 4; the other three segments adjust to higher thresholds in association with the GNOO or GLOO fits. The Montana flooding of August 2014 is another example; Atlas thresholds apply to this region in segment 4, but higher GPAO or PE3O thresholds are deemed best when this event is included for the other three segments. Montana did experience two events of significance, the other occurring during segment 1 (spring 2010); it is of note that this did not result in an adjustment of thresholds in segment 4. Another location of interest is in the southeast US in the vicinity of the Tennessee valley. This region was experienced several major events during the verification record: the historic southeast floods of September 2009 (Segment 1), the similarly impactful flooding in and near Nashville in May 2010 (Segment 1), Tropical Storm Lee during September 2011 (Segment 2), and several slightly less impactful events. The results of verification over segment 1 resulted in an adjustment to the WEIO fit for segment 2 and the PE3O for segment 4. Another area of note is eastern Florida, which is adjusted to the GLOO fit in all but segment 3 in association with Tropical Storm Debby, which affected the region during that time.

Bulk cross-validation verification for the NSSL-WRF is presented in Figure 4.11. The impact of using the model thresholds selected here is negligible for the 1- and 2-year RPs. The change is generally positive, but by less than 1% and the change is not statistically significant. At the 5-year RP, a 2-3% improvement in aggregated FSS is observed at all evaluation radii; at most of the evaluation radii, this improvement is not found to be statistically significant, but it is significant at a few of them. At the 10-year RP, thresholds from the algorithmically determined best RSD fits enhance forecast skill by 3-5%, with more skill improvement witnessed at the lower evaluation radii. Here, the improvements are found to be statistically significant at all evaluation radii. The improvement is even larger at the 25-year RP- approximately 5-6% depending on the evaluation radius chosen- but due to the decreased number of events at this RP, the uncertainty associated with skill comparisons corresponding increases, and this

improvement is not statistically significant. The 50-year RP, the highest evaluated for the NSSL-WRF, saw a lesser improvement over the Atlas thresholds than compared with the 25-year RP, with 1-3% improvements observed. With the very large error bars in association with the rarity of the event, none of these improvements were found to be statistically significant.

Lastly for the NSSL-WRF graphical CONUS-wide verification is presented in Figure 4.12. At the 2-year RP (4.12a), no major differences are seen, but slight improvements in skill are observed in the upper Mid-Atlantic/southern New England and over much of the high plains, but is degraded in Virginia and vicinity, over much of the Ohio Valley, and over eastern Colorado and vicinity. At the 5-year RP (4.12b), the Mid-Atlantic/New England area skill enhancement is substantially further improved, and new areas of improvement are seen in Louisiana from northern Georgia through far southwestern Virginia. The degradation in forecast skill seen in Virginia and vicinity for the 2-year RP is greatly alleviated for the 5-year RP verification. At the 10-year RP, four major regions of skill improvement are seen: 1) all of the Mid-Atlantic corridor and southern New England; 2) the southeast, particularly Georgia, the Carolinas, and Tennessee; 3) the high plains in the vicinity of South Dakota; and 4) northern California. On the negative side, forecasts for the Ohio Valley and Colorado Plains regions are significantly harmed by applying the identified best RSD fits. Finally, for the 50-year RP, major improvements are seen in the southeast region and over West Virginia, southern Pennsylvania, and western Maryland. Moderate improvements are also retained over northern California. However, FSS degradations are seen along parts of both the Gulf Coast and Atlantic Coast, limiting the potential improvement by applying the identified best RSD fits.

The identified most skillful distributions from cross-validation for the GEFS/R model appear in Figure 4.13. Unlike the NSSL-WRF best fits, which varied considerably across cross-validation segments, the identified adjustments here are identical across segments, in part due to the regularization. Many

fewer adjustments are seen in general, with only one region- the Mid-Atlantic- identifying model-derived thresholds as verifying more skillfully than the Atlas thresholds. Here, the EXPO fit was found to be the most skillful fit over each validation period.

Figure 4.14 presents bulk GEFS/R verification using the identified best thresholds alongside the Atlas threshold based verification. 2-3% improvements are observed at the 2-year RPs over all the evaluation radii examined; all of these improvements are found to be statistically significant. At the 5-year RP, however, larger improvements of 3-8% are observed with the application of the model RSD fit RPTs, with the larger improvements at smaller evaluation radii; these are also found to be statistically significant. Like in the NSSL-WRF verification, the model RPTs improve over the Atlas thresholds even more at the 10-year RP, with 12% improvements decreasing to approximately 6% at the largest evaluation radii. In spite of increasing error bars in association with the decreased sample size, these improvements are largely found to be statistically significant. Though the 25-year RP sees the largest skill improvements, with mostly 10-25% enhancements over the Atlas thresholds, the uncertainty associated with the sample size renders the results not statistically significant. Though noisy, similar but still insignificant improvements of mostly 10-25% are seen at the highest RPs, 50- and 100-years. This behavior at the higher RPs is consistent with what is seen in the NSSL-WRF.

Regional FSS changes with the implementation of the model best fits is depicted in Figure 4.15. Since the only threshold changes (see 4.13) occurred in the Mid-Atlantic/southern New England region, this is the only region that experiences non-zero skill changes. Slight improvements are seen in forecast skill at the 2-year RP in Virginia, increasing to moderate improvements over West Virginia and eastern Ohio. Conversely, moderately negative changes are seen over much of New York and southern New England. The trend continues for the 5-year RP, with two major areas of improvement seen over the Virginia/North Carolina coast and in northeastern Ohio complemented by continued degradation in

forecast skill to the north. The pattern is similar but amplified at the 10-year RP, except with forecasts over much of Pennsylvania now harmed by the switch to model RPTs. By the 50-year RP, very large FSS improvements are seen for essentially the entire modified region from Maryland south, with moderately worse performance over much of Pennsylvania.

Final application of the developed algorithms over the entire verification period was applied, and a portion of those results appears in Figure 4.16. The final NSSL-WRF verification (4.16a) identifies the GLO distribution as the most skillful fit over the Mid-Atlantic region, and the GLOO as the best distribution over the northern Great Plains and intermountain northwest. The GEFS/R verification (4.16b) is nearly identical to all of its cross-validation components, with the only adjustment from Atlas thresholds being the PE3 and WEI implementation in the Mid-Atlantic region. The corresponding changes to the model precipitation thresholds at the 50-year RP appear in Figure 4.16c and 4.16d for the NSSL-WRF and GEFS/R, respectively. It is evident that the NSSL-WRF adjustments have the effect of raising the thresholds moderately in the Montana vicinity, and raising the thresholds slightly in the Mid-Atlantic. The GEFS/R, in contrast, corresponds to a slight downward adjustment in the Mid-Atlantic thresholds. These are the grids that would be applied towards evaluating the utility of the algorithm on a test sample.

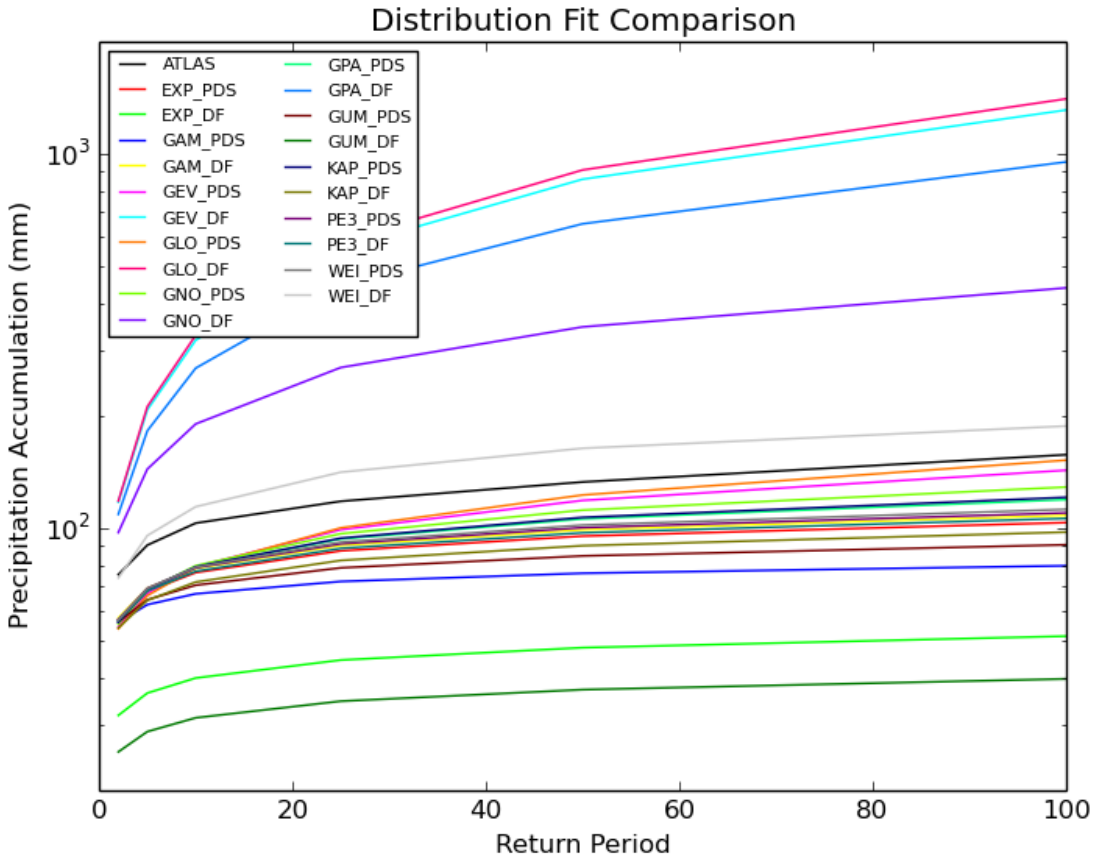

Figure 4.1: An example at an arbitrary point comparing DF-FDS and PDS fits to GEFS/R data using 01 December 1984-09 May 2013 data. Precipitation accumulations are plotted on a logarithmic scale as shown.

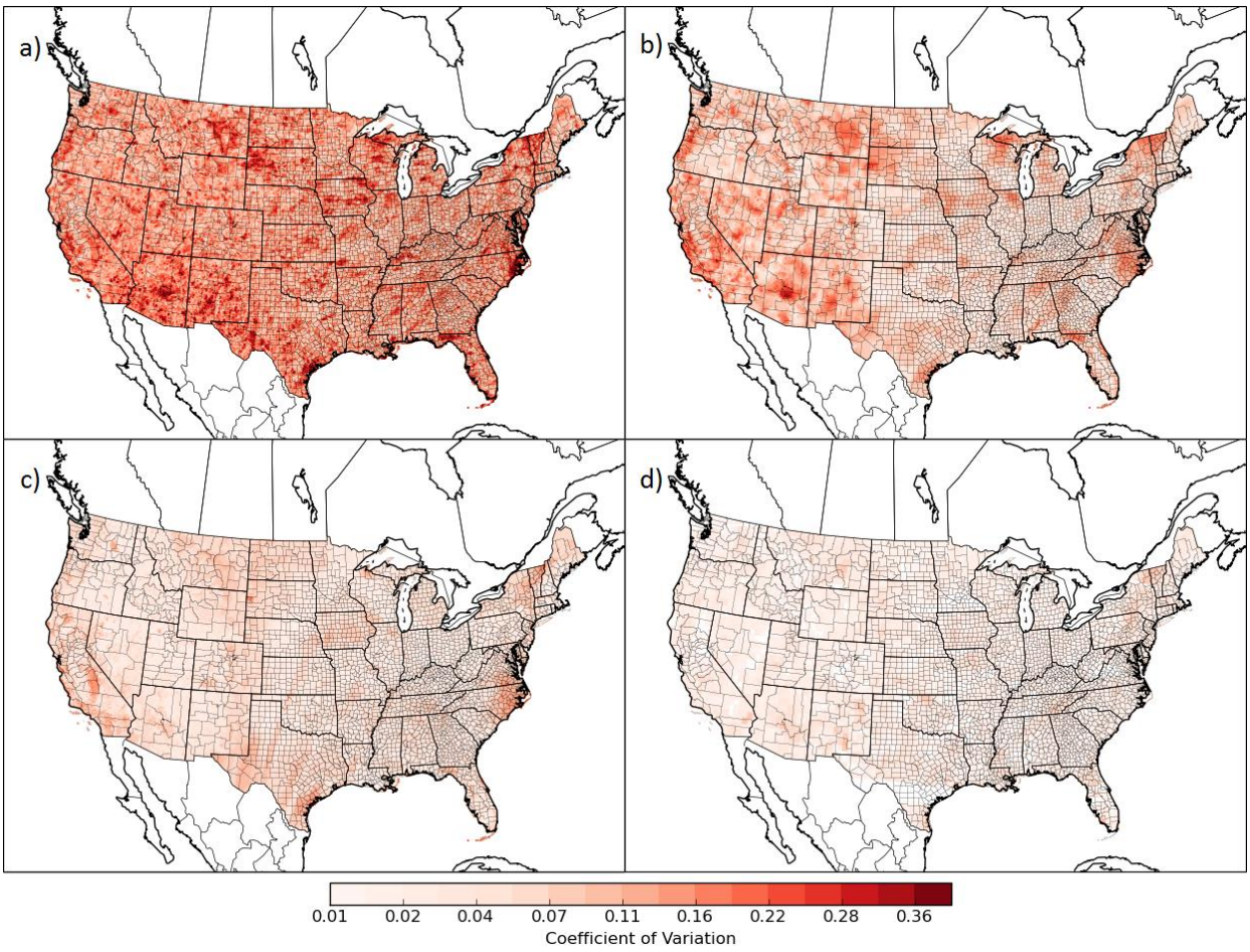

Figure 4.2: Coefficient of variation obtained by comparing the cross-validation RPT estimates for the 100-year RP from the EXP distribution. Panel (a) corresponds to estimates for the NSSL-WRF obtained from point-by-point fitting; panel (b) corresponds to the estimates after application of the smoothing procedure discussed in-text; panel (c) corresponds to the estimates after additional regionalization step discussed in-text; and panel (d) corresponds to GEFS/R estimates without smoothing or regionalization.

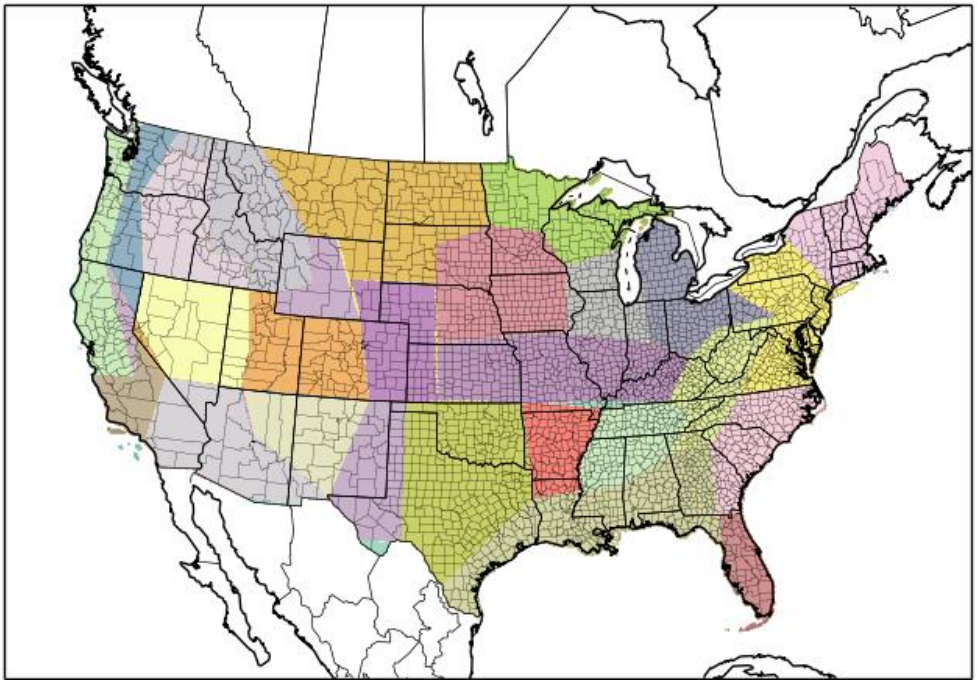

Figure 4.3: Plot of different regions used for discerning best regional RSD fits; each color indicates a distinct region.

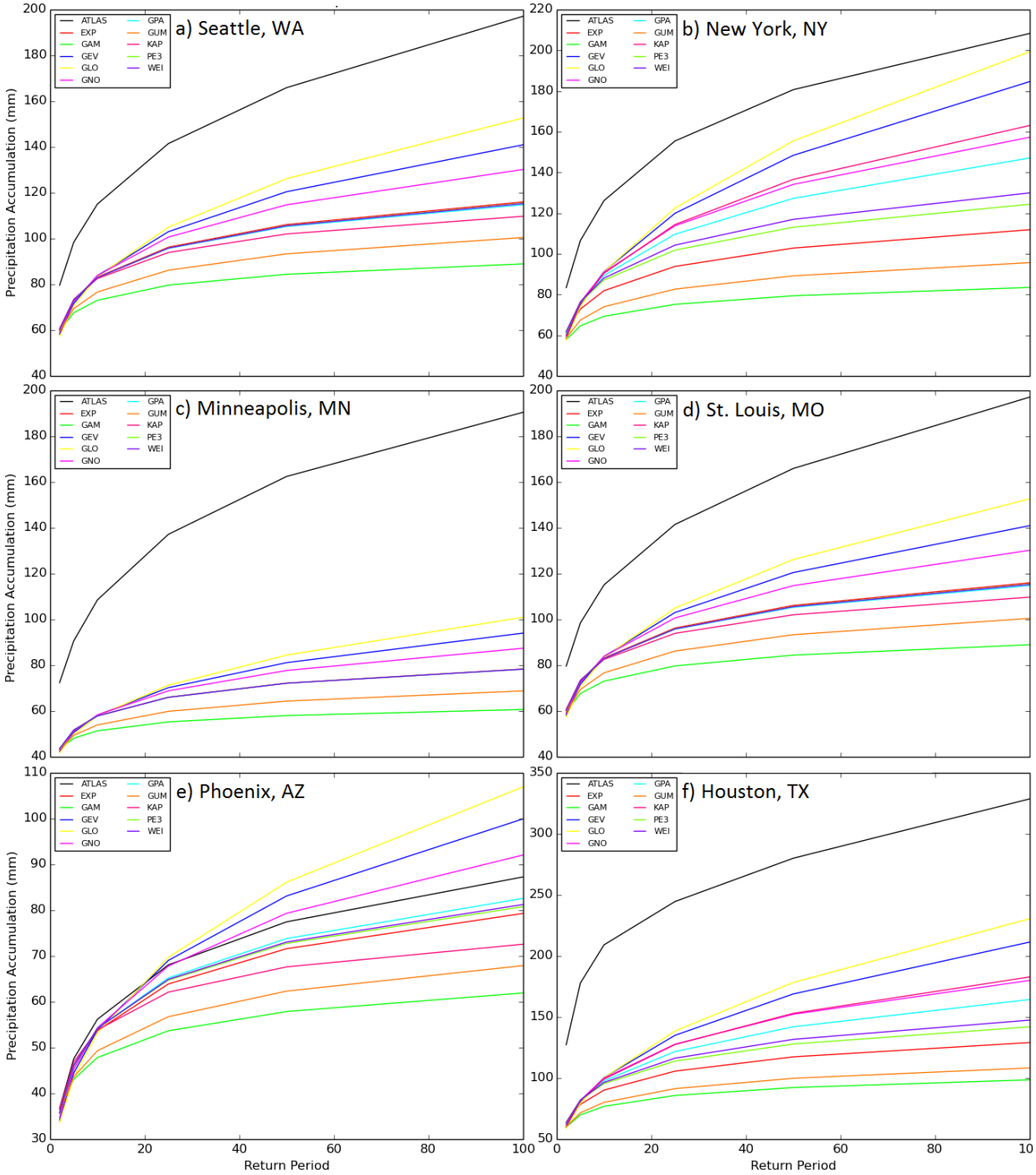

Figure 4.4: GEFS/R fits for grid points near select cities around CONUS. Distribution fits are from PDS-derived thresholds for GEFS/R control run thresholds for initializations from 01 December 1984 to 09 May 2013.

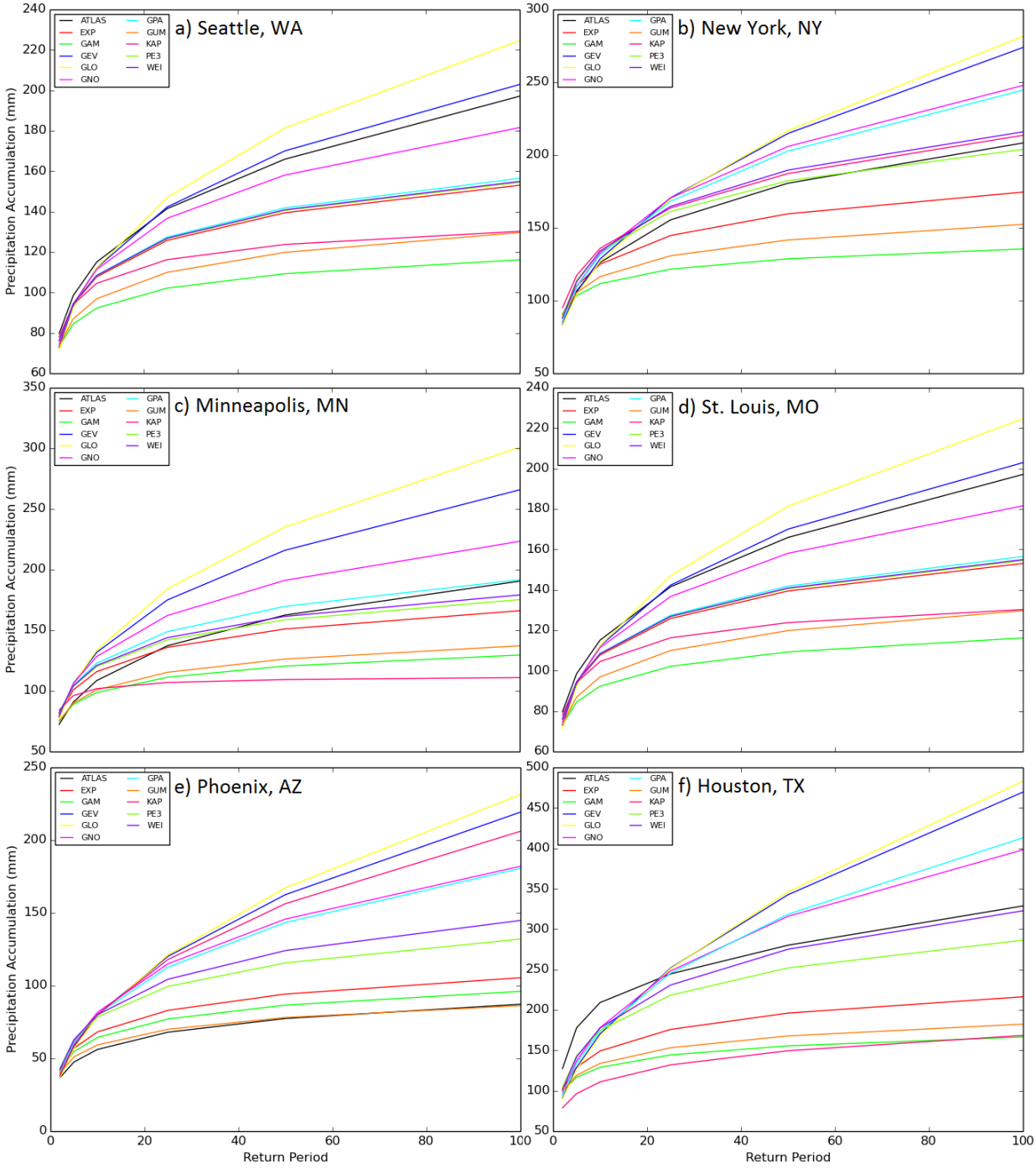

Figure 4.5: NSSL-WRF fits for grid points near select cities around CONUS. Distribution fits are from PDS-derived thresholds for initializations from 09 June 2009 to 09 May 2013.

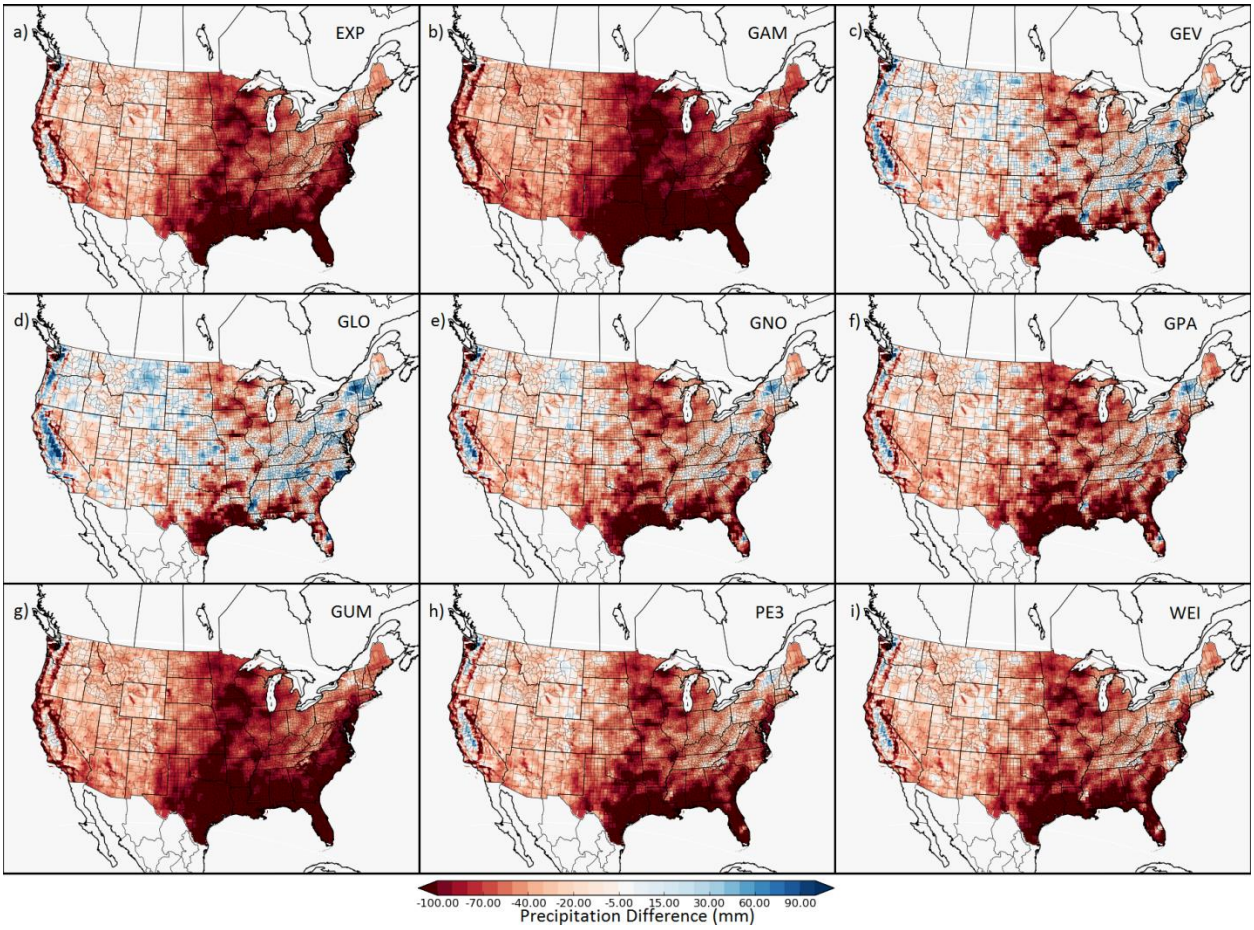

Figure 4.6: RPTs at the 100-year RP for various RSD fits for GEFS/R smoothed using PDS fits performed point-by-point on the point precipitation data. Panels (a)-(i) provide the EXP, GAM, GEV, GLO, GNO, GPA, GUM, PE3, and WEI fits relative to the Atlas thresholds at the same RP, respectively. Fits correspond to those excluding the 10 May 2013-30 August 2014 portion of the verification period.

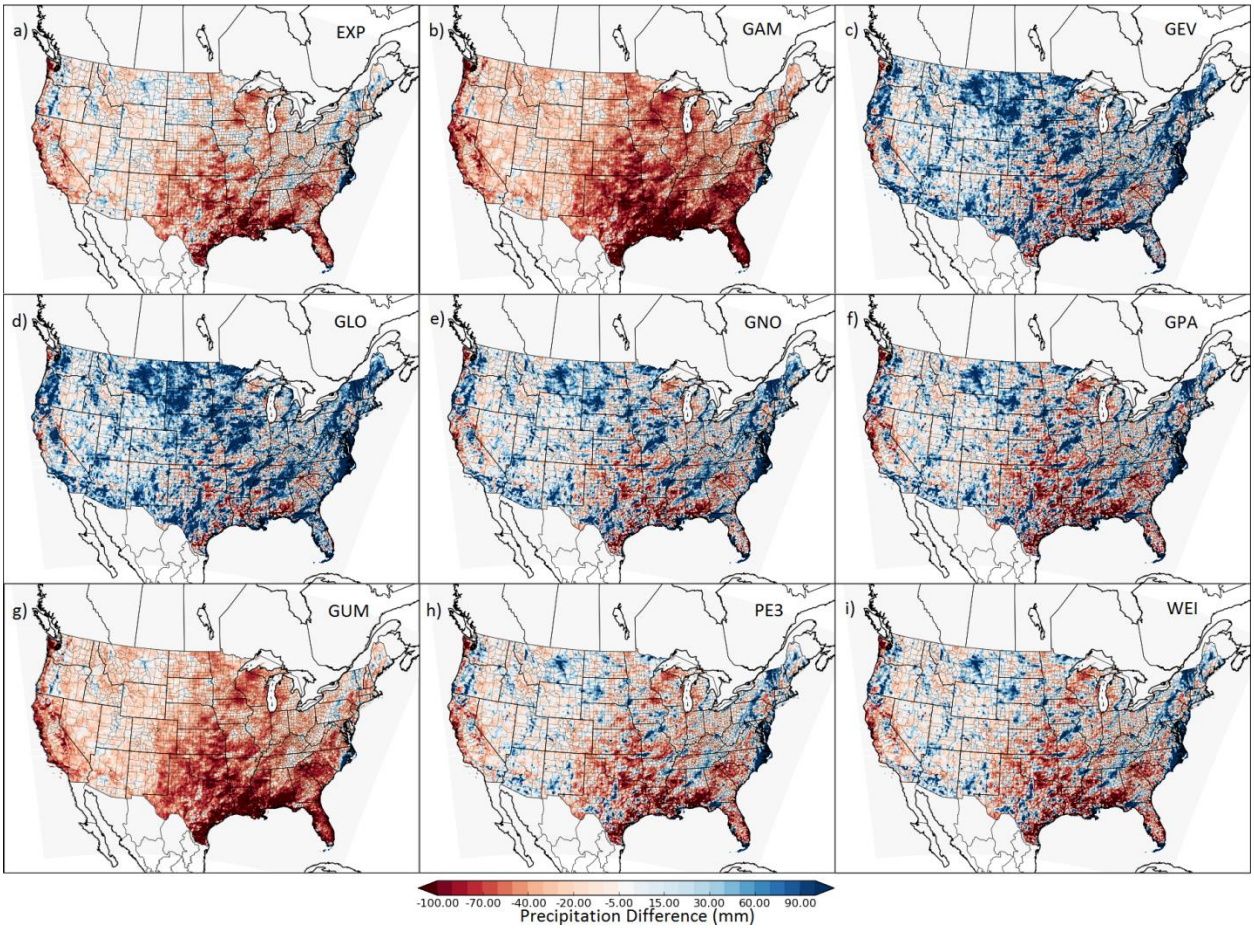

Figure 4.7: Same as Figure 4.6, but for the NSSL-WRF.

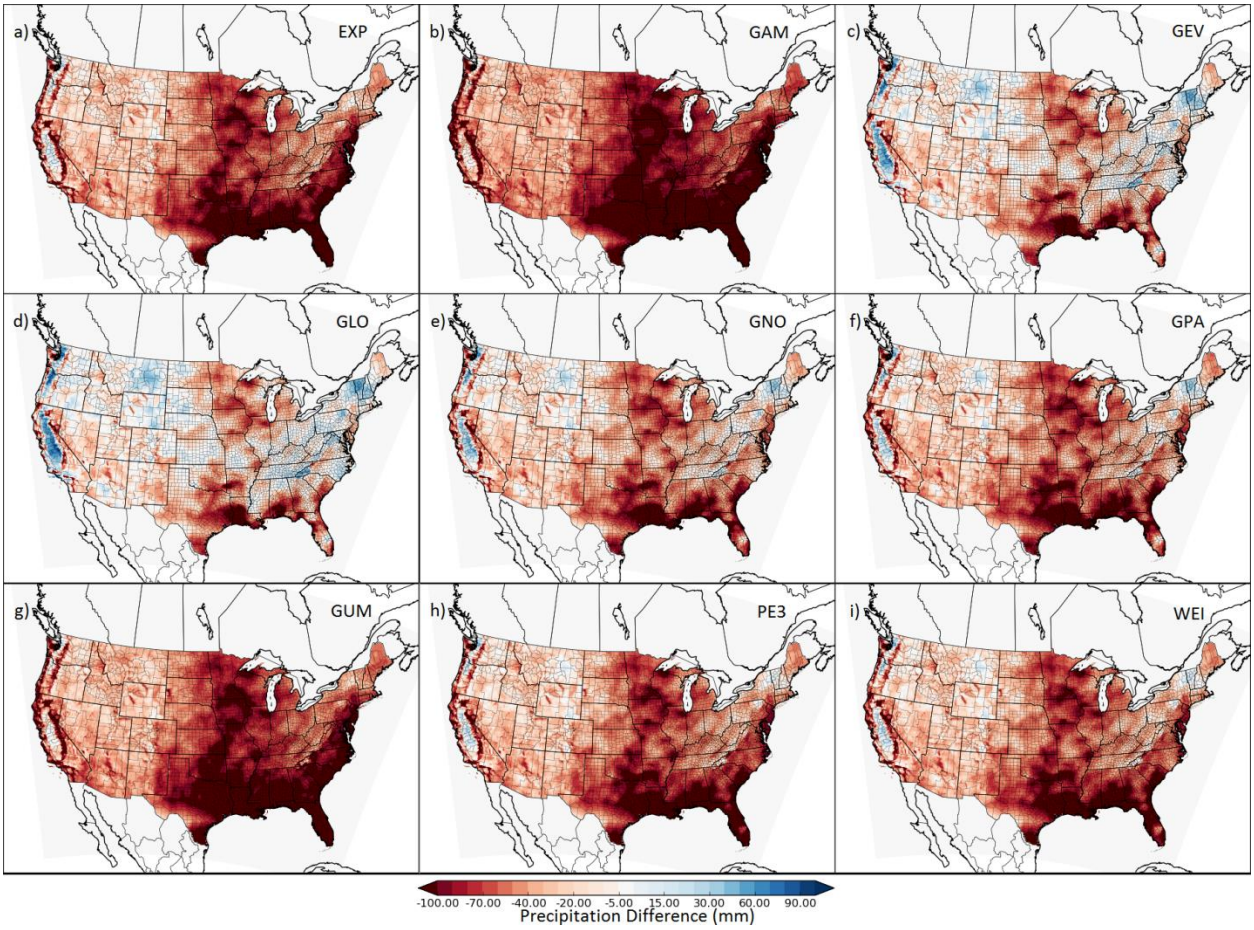

Figure 4.8: RPTs at the 100-year RP for various RSD fits for GEFS/R smoothed using the algorithm described in the text. Panels (a)-(i) provide the EXP, GAM, GEV, GLO, GNO, GPA, GUM, PE3, and WEI fits relative to the Atlas thresholds at the same RP, respectively. Fits correspond to those excluding the 10 May 2013-30 August 2014 portion of the verification period.

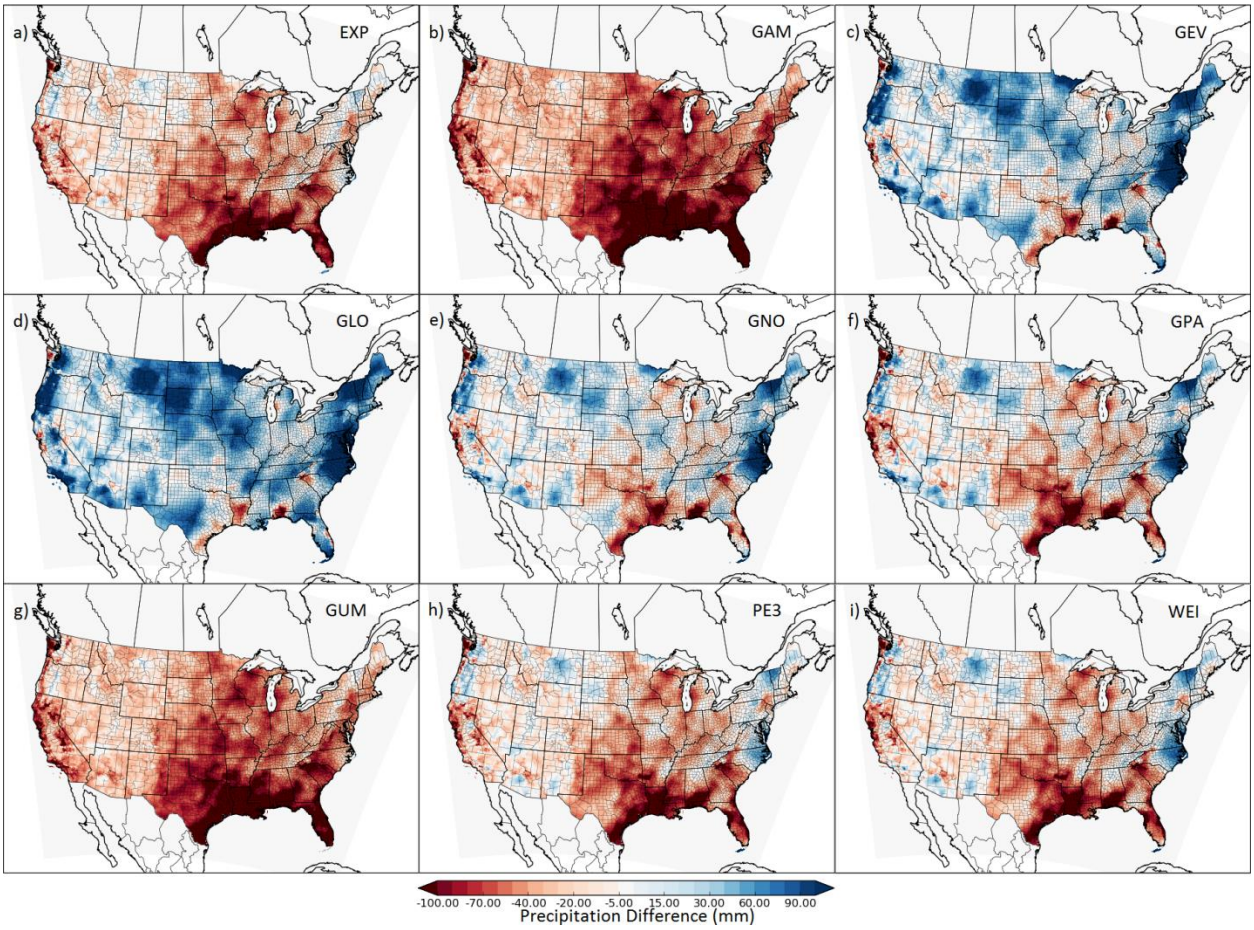

Figure 4.9: Same as Figure 4.8, but for the NSSL-WRF.

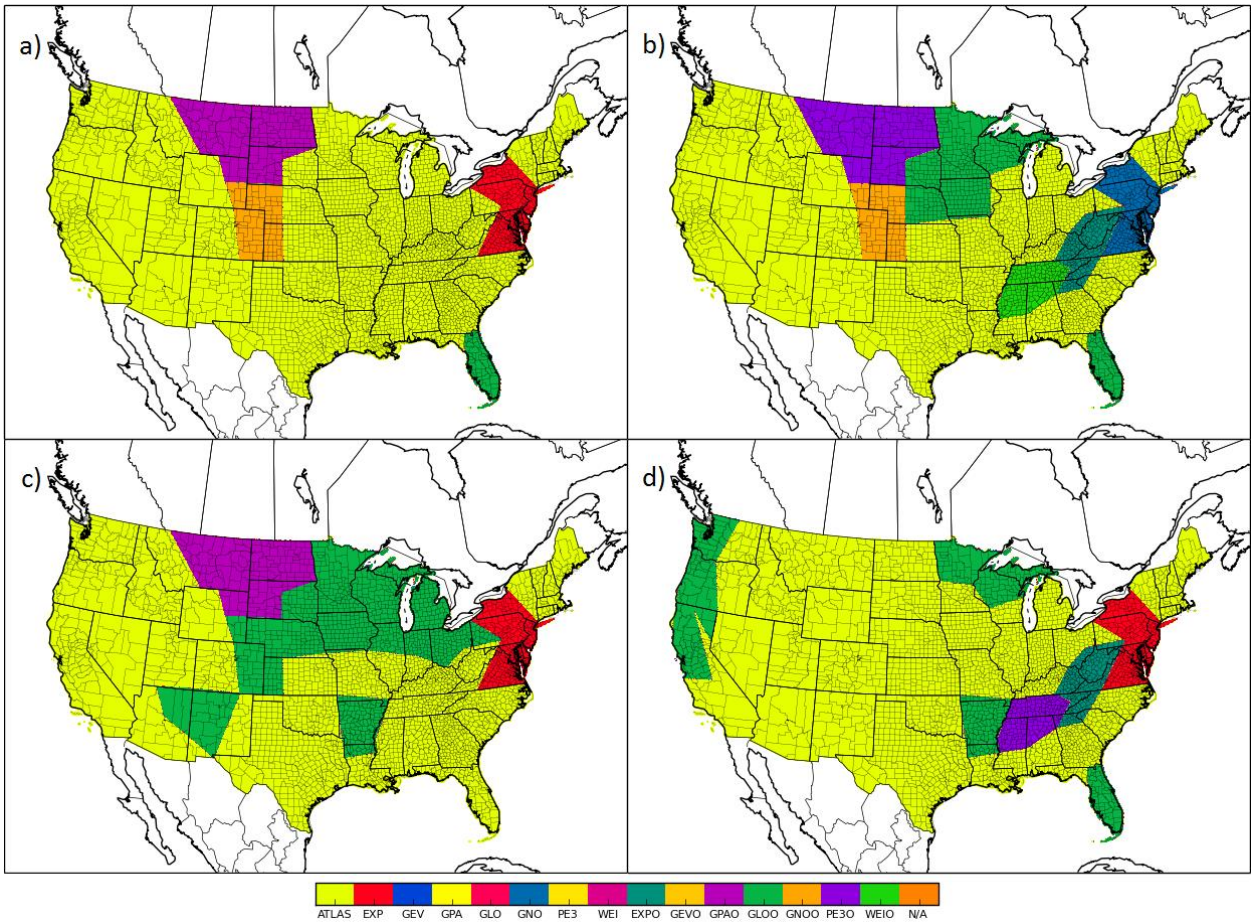

Figure 4.10: Identified best distributions for the NSSL-WRF for the four cross-validation trainings. Panel (a) corresponds to the best distributions identified for the first quarter of the verification period using thresholds derived from the second, third, and fourth quarters and evaluated over those same quarters. Panel (b) is the same as panel (a), except for the second quarter; panel (c) identifies best distributions for the third quarter, and panel (d) for the fourth. 'ATLAS' implies the local observationally-derived thresholds were deemed the most skillful. Distribution names with an 'O' appended are offset by a factor of two, meaning, for example, that the 50-year RPT estimates from that distribution are used for the 25-year verification. 'N/A' would indicate that the choice of distribution was immaterial for that location.

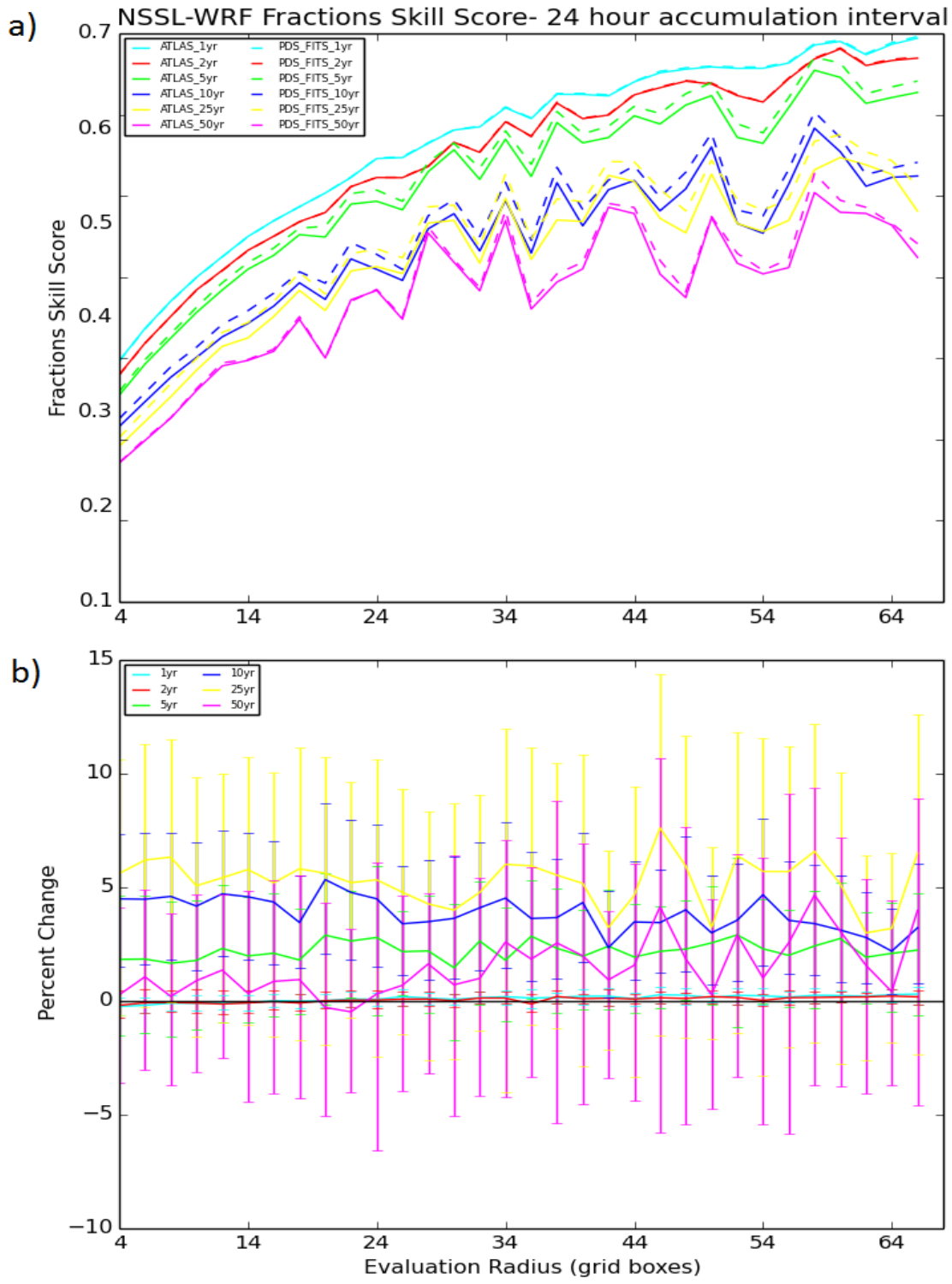

Figure 4.11: NSSL-WRF Fractions Skill Score verification for 1-50 year return periods. Top panel shows actual FSSs; solid lines depict verification using the ATLAS thresholds, while dash lines correspond to verification against the model climatology derived thresholds. The bottom panel indicates the percent change in FSS, as a function of evaluation radius, by applying the model climatology thresholds. Error bars are 90% confidence bounds obtained by bootstrapping.

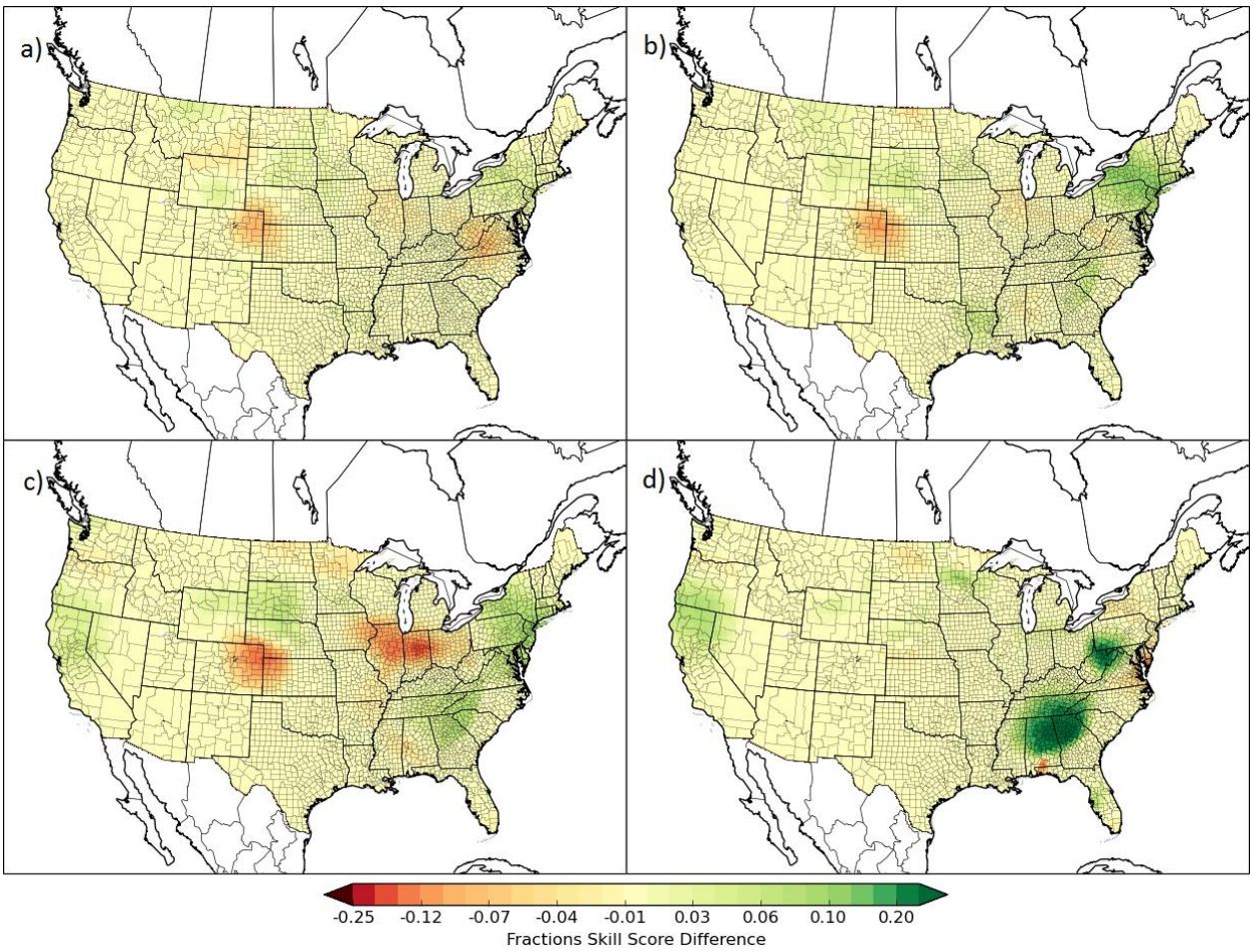

Figure 4.12: Graphical representation of NSSL-WRF FSS verification by applying model climatology RPTs. Differences are with respect to ATLAS threshold-based verification; greens indicate local improvement by switching to the model thresholds, reds indicate degradation. Panel (a) corresponds to the 2-year RP, panel (b) to the 5-year RP, panel (c) to the 10-year RP, and panel (d) to the 50-year RP. All plots correspond to verification using an evaluation radius of 40 grid boxes, and over the entire verification period, from 09 June 2009-30 August 2014.

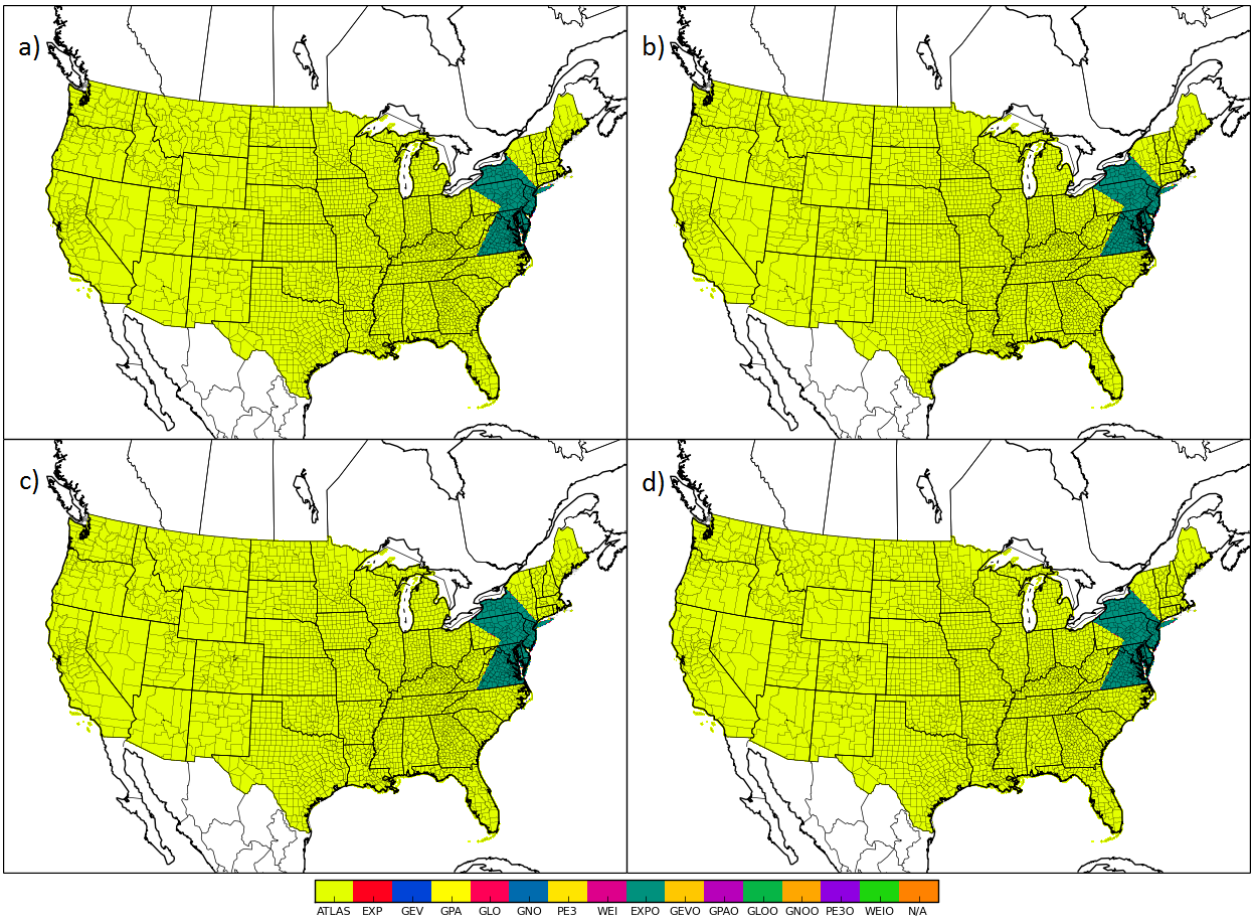

Figure 4.13: Same as Figure 4.10, except for the GEFS/R.

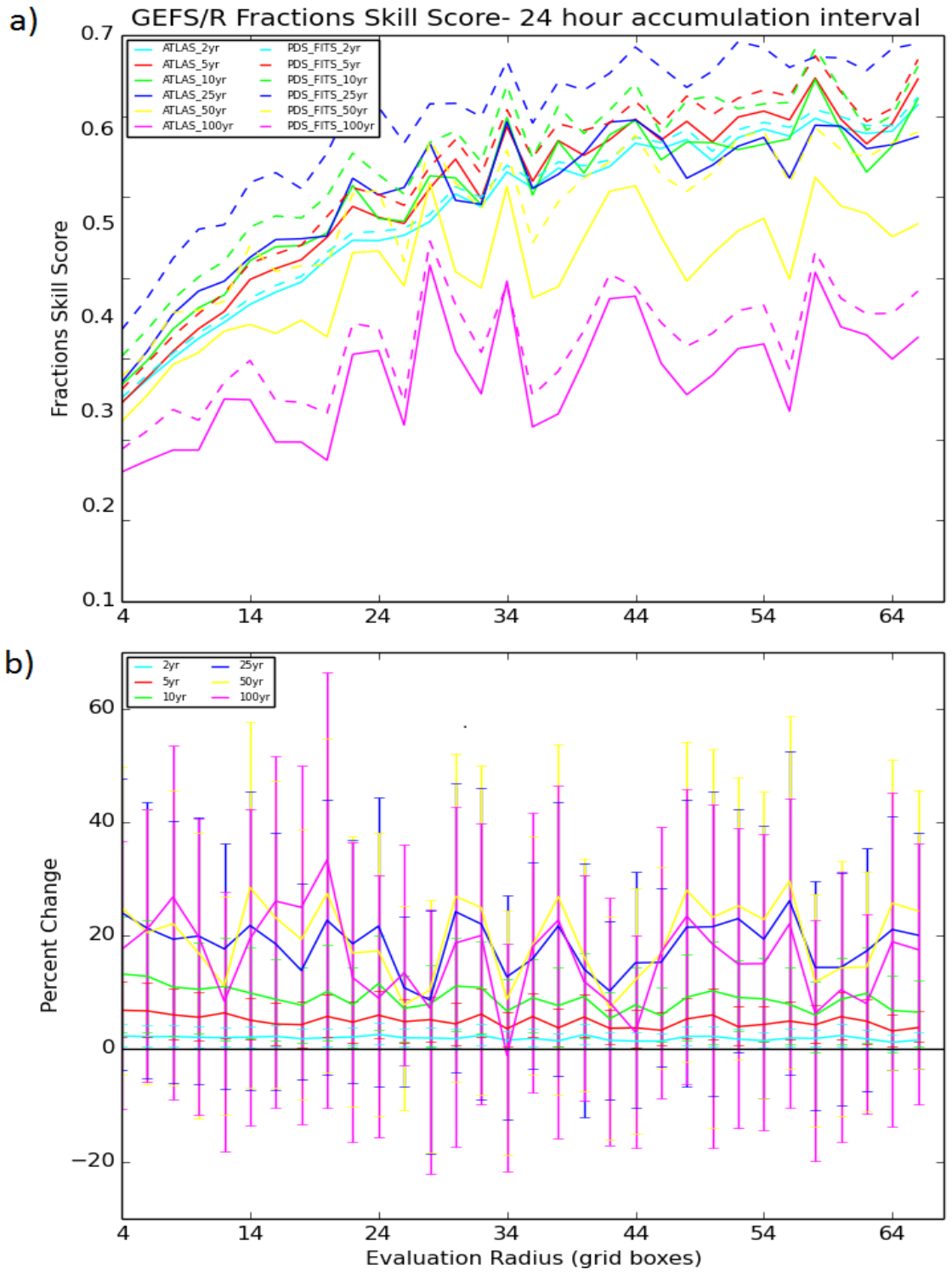

Figure 4.14: GEFS/R Fractions Skill Score verification for 2-100 year return periods. Top panel shows actual FSSs; solid lines depict verification using the ATLAS thresholds, while dash lines correspond to verification against the model climatology derived thresholds. The bottom panel indicates the percent change in FSS, as a function of evaluation radius, by applying the model climatology thresholds. Error bars are 90% confidence bounds for the skill difference obtained by bootstrapping.

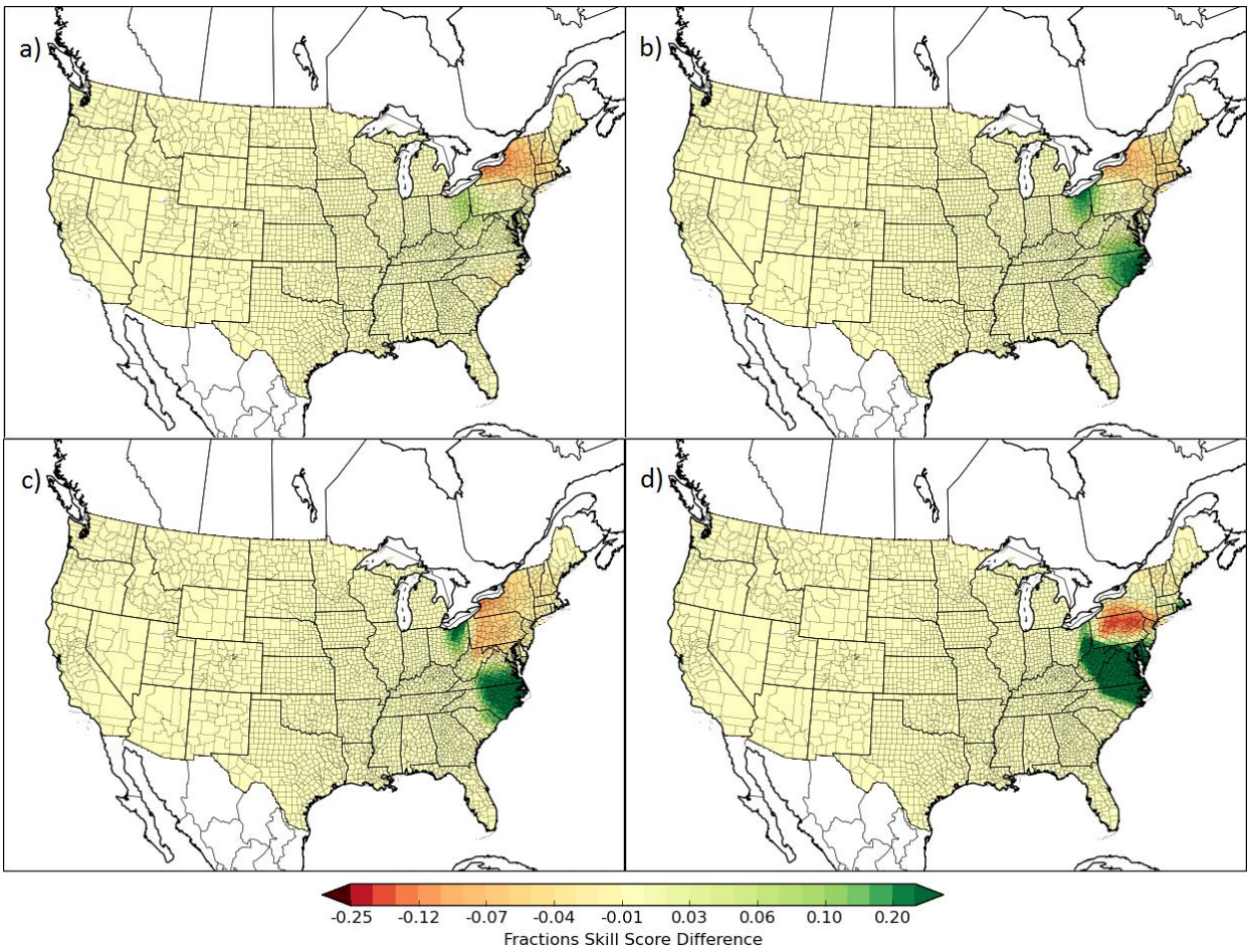

Figure 4.15: Graphical representation of GEFS/R FSS verification by applying model climatology RPTs. Differences are with respect to ATLAS threshold-based verification; greens indicate local improvement by switching to the model thresholds, reds indicate degradation. Panel (a) corresponds to the 2-year RP, panel (b) to the 5-year RP, panel (c) to the 10-year RP, and panel (d) to the 50-year RP. All plots correspond to verification using an evaluation radius of 40 grid boxes, and over the entire verification period, from 09 June 2009-30 August 2014.

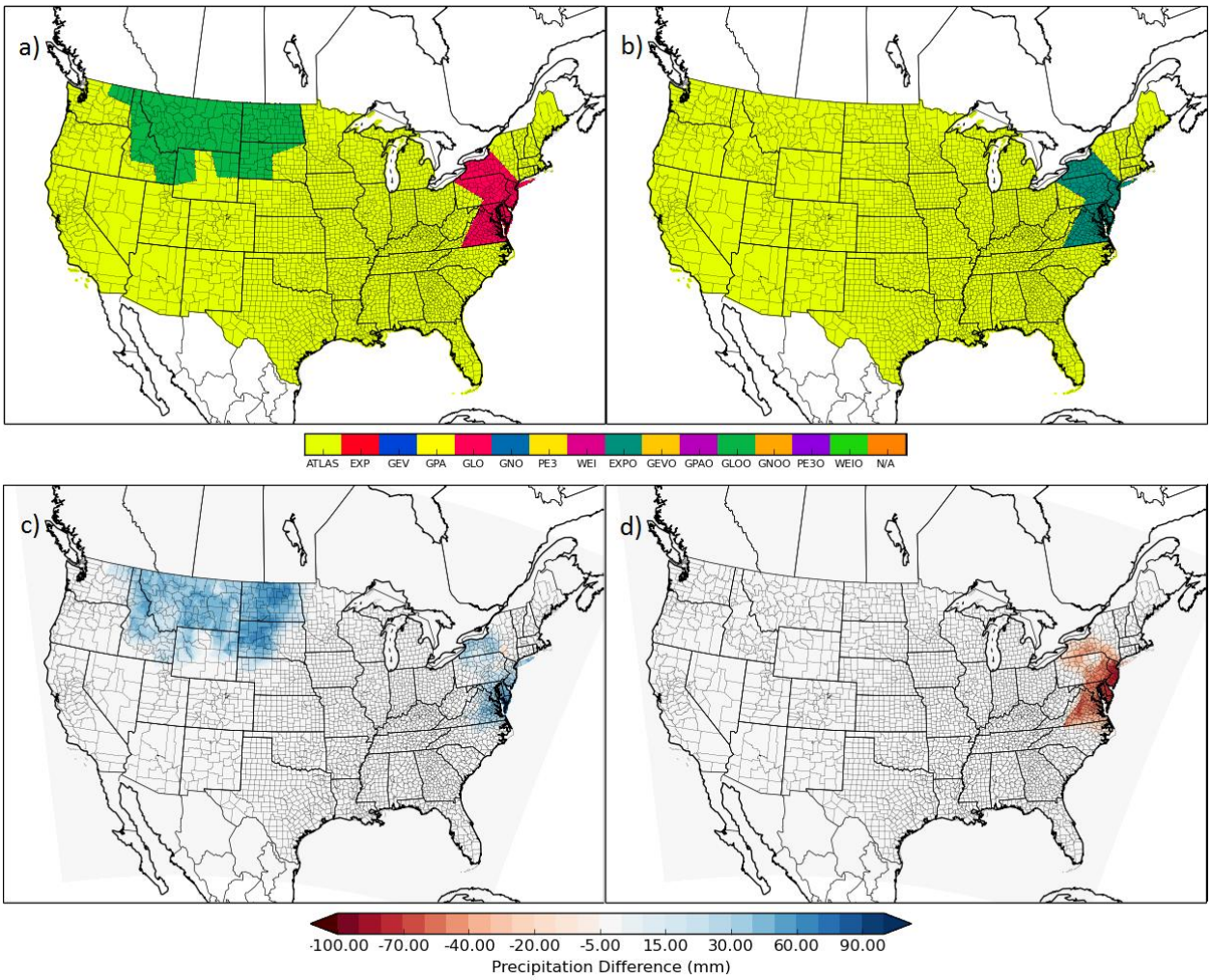

Figure 4.16: Final identified best RSD fits and corresponding thresholds trained and evaluated over the full verification period. Panel (a) corresponds to NSSL-WRF verification best RSD identification, and (b) the same for the GEFS/R. Panels (c) and (d) correspond to the precipitation threshold comparison vs. Atlas thresholds at the 50-year return period for the NSSL-WRF and GEFS/R, respectively.

4.3 Discussion & Conclusions

Applying model precipitation climatologies toward locally extreme rainfall forecasting was able to improve forecast skill over simply using the true predictand thresholds derived from observations. The results were especially robust at medium RPs of 5 and 10 years, where the improvement was generally found to be statistically significant. Larger improvements were generally seen in the GEFS/R, but the

uncertainty was also higher in association with the GEFS/R not being able to resolve many extreme events reducing the effective extreme precipitation event sample size for that model.

There is strong evidence to suggest that the validation of the algorithm presented here suffers from insufficient model and verification data. The region that experienced the most unique extreme rainfall events is almost undoubtedly the Mid-Atlantic and New England, with an anomalous number of tropical cyclones impacting the region during the verification period. This is also one of the areas where the biggest average improvement in skill is observed. The methodology used here does extrapolate thresholds for very rare events from the characteristics and distribution of more common precipitation forecasts, but it is very difficult to discern differences in the skill of using particular thresholds in forecasting when there is no relevant event to verify against. Often times, the higher thresholds end up being deemed most skillful, since when there are no events in the record, the highest thresholds are the best since they produce the fewest false alarms. From the northeast US, we see that when there are multiple events to compare against, thresholds are often able to be sensibly adjusted from the Atlas thresholds to enhance model predictive skill. The extrapolation from distant points to local verification struggles between being too small to bring in new useful forecast information, and being too large such that the local verification results are no longer representative of the true model behavior at that location. Unfortunately, model data will almost always fall well short of this data length objective, and the verification period length issue is likely to remain in future work in this realm. However, GEFS/R data has a uniquely long consistent model data record- now over 30 years. Stage IV precipitation analysis was used for verification in this study, which limited even the extended verification period back to only 01 January 2002. There are coarser precipitation analyses, such as the Climate Prediction Center's Unified Precipitation Analysis, provided at 0.25°-0.25° grid spacing. This is much coarser than the Stage IV analysis, the grid on which Atlas 14 thresholds are calculated, which may render verification of extreme precipitation using this dataset unreliable and inaccurate. However, if verification can be

reliably obtained with a coarser dataset (Stage IV is not without its own major problems), then using this entire record may be able to significantly enhance identification of the most skillful thresholds. This may be a fruitful avenue of future research to pursue, though it may suffer from a lack of generalizability in operations, since no other model has a constant data record nearly as long. It would also be worth exploring regionalization in the threshold selection phase to appropriately enhance local verification with threshold-specific model performance at more distant venues where possible.

The verification period length is also too short relative to the event frequency to be able to draw many firm conclusions about algorithm skill. At the high RPs, error bars are so large that 15-20% improvements seen at the 25-year RP in the GEFS/R are insufficient to yield statistical significance; there simply are not enough events to be able to tell, in light of the fact that changing the thresholds does not uniformly enhance forecast skill. A longer verification record could appreciably reduce the skill uncertainty and allow a more definitive acceptance or rejection of the proposed algorithm for high RP exceedance forecasting. At its face, a 10+% improvement in deterministic forecast skill is quite encouraging, and could easily add tremendous value to end users. The statistically significant improvements seen for the 5- and 10-year RPs for both the GEFS/R and NSSL-WRF, in addition to the act that positive improvements are seen for the vast majority of RPs and evaluation radii provide further evidence that the algorithm can enhance locally extreme precipitation forecast skill.

Several additional avenues of further research in this area persist. It is certainly of interest to examine the applicability of the technique proposed here to the 6-hour accumulation interval, which has not been explored here. Additionally, the GEFS/R's forecast skill suffers primarily from its inability to resolve many heavy precipitation features, and, from chapter 3, the NSSL-WRF's biases were relatively tame relative to some other CAMs. Perhaps model climatologies can add more skill, even with a shorter training record, applied to a model without the inherent deficiencies of the GEFS/R, but with more

pronounced biases than the NSSL-WRF, such as the NAM-NEST. This would be of interest for further research as well.

5 Techniques for Locally Extreme Rainfall Post-Processing: Model Development and Training

5.1 Methods

This investigation will explore several classes of algorithms for using raw NWP guidance to generate a probabilistic forecast. The first class will be termed Naïve Algorithms (NIVAs), so termed because they assume no knowledge about the characteristics of any of the ensemble members used to generate the FPs; they simply take each forecast at face value. Most ensemble-based probabilistic guidance generated in real time is generated by NIVA methods. Making no specific assumptions about ensemble members has both benefits and drawbacks. It allows the use of a very large amount of model guidance, and thus a large number of ensemble members, to inform the FPs, since no prior training or evaluation of ensemble members is required and thus models can be used even if they're new or recently modified. As discussed in Section 2.3, a large number of ensemble members may be required to appropriately gauge the extreme PDF tail. They're also simple to apply and interpret, and since no model training is required, NIVA application is computationally inexpensive. On the negative side, NIVAs make no attempt to correct for model biases, and so systematically biased input will yield biased, unreliable FPs. Since NIVAs make no effort to discern model skill, the addition of poor, unskillful guidance will substantially degrade the performance of the output FPs. On a related note, the FPs will also be overly influenced by redundant forecast data. This may appear to be more of a theoretical issue than a practical one, but many operational EPSs are only IC-perturbed and systematically underdispersive, resulting in member solutions that are inappropriately correlated. Using many members from such an EPS to inform these FPs may thus overemphasize EPS-based solutions compared to deterministic runs which may be equally or more skillful. No direct calibration of FPs may result in

them being unreliable (see Sections 2.3 and 2.7). These algorithms also inherently make no effort to discern relationships between other forecast variables and observed precipitation, and useful information may be contained in these relationships.

The second FP-generation algorithm class will be termed Intermediate Algorithms (INTAs). These algorithms operate like NIVAs, except that ensemble members are weighted based on their historical performance. Single weights may be assigned to each ensemble members, or weights may be allowed to vary based on location and/or potentially time (of year or, if applicable, of day). This approach has the capability of alleviating many of the aforementioned problems associated with NIVAs. Less skillful models may be assigned lower weights so as not to degrade FP estimates; further, redundant (highly correlated) members may be assigned lower weights so that the forecast information each contains is not overrepresented. Not all problems are solved, however: there is still no mechanism in place to address model bias, and no FP calibration in place to ensure forecast reliability. New issues are introduced as well. Weighting members requires estimates of model skill, which requires historical model data to verify. This limits the amount of forecast guidance that can inform the FP estimation, since new models, models that have recently been changed and have acquired new biases and skill characteristics, and models with limited or no access to historical model data become difficult to use. Weights may ostensibly be assumed or crudely estimated, but poor performance estimates or assumptions can result in worse results than having assumed no weights at all.

The last FP-generation algorithm class will be termed Advanced Algorithms (ADVAs). The aim of this class is to directly derive historical relationships between model predictors and the forecast predictand of interest, which in this case is the exceedance probability for various RP thresholds. Unlike NIVAs and INTAs, ADVAs do not take the NWP output at face value, and can theoretically correct for meteorological dependent biases, displacement biases, and other issues in the underlying models

comprising the ensemble members that simpler algorithms lack the complexity to achieve. These methods also much more directly calibrate the FPs, and should ostensibly result in a more reliable PFS than results emerging from either of the previous two algorithm classes. The primary drawback of this method is that deriving these relationships requires a long, consistent, robust record of historical model data; this is especially true here when the forecast predictand is, by definition, a (very) rare event. It is unknown exactly how long of a data record is required to accurately discern such relationships, but it is certainly sufficiently long to, at a minimum, drastically reduce the number of models that can be realistically used, since relatively few models run for several years without changes that significantly impact the necessary predictor-predictand relationships. ADVAs can also be rather computationally expensive to train, at least relative to the other two approaches; however, relative to a single high resolution dynamical model, the cost is still quite low and does not appear to be a substantial consideration in any operational setting. The next subsections will explore some of the algorithmic details that will be explored.

5.1.1 Naïve Algorithms (NIVAs)

5.1.1.1 Point Democratic Voting (PDV)

Consider an ensemble of size *n*. For a given location (y,x) and accumulation period AP specified by forecast lead time (FL) and accumulation interval (AI), a given ensemble member *k*'s QPF may be denoted Q_{yxk}. The threshold required to be met or exceeded for a specified event of interest to be forecasted to occur in the model- in this case, the exceedance of an N-year precipitation RP- may be denoted $\theta_{N_{yx(k)}}$. The subscript k in parentheses is added to indicate that the threshold may optionally vary as a function of ensemble member if Quantile Mapping Bias Correction (QMBC) as described in Chapter 4 is applied to the members. A binary predictand E may then be defined to describe whether an ensemble member *k* is forecasting the event of interest at point (y,x):

$$E_{yxk} = \begin{cases} 1 & Q_{yxk} \geq \theta_{N_{yx(k)}} \\ 0 & Q_{yxk} < \theta_{N_{yx(k)}} \end{cases}$$

The PDV method then simply generates FPs by taking the mean forecast E at each point:

$$FP_{yx}(PDV) = \frac{\sum_k E_{yxk}}{n}$$

PDV is the traditional method of deriving FP's from an EPS; most real-time ensemble FP products are generated using the PDV algorithm. It does have the advantage of being straightforward, intuitive, and very inexpensive. However, as previous authors have elucidated (e.g. Eckel 2003, , there are several theoretical issues with PDV probabilities. Most obviously, the probabilities are unrealistically discretized into n+1 bins, when, in reality, true FPs lie continuously on the interval [0,1]. This problem is exacerbated at small ensemble sizes where each bin consumes a larger proportion of probability space. PDV is known to be overconfident in its probability estimation, biasing to the extremes. For example, when no member forecasts an event to occur, PDV assigns an FP of 0, when in reality, there may very well be a small non-zero probability of event occurrence for any finite-sized ensemble. Lastly, PDV makes inefficient use of forecast information; it only considers the binary relation of QPFs to the critical threshold θ_N without regard to *how close* the QPF lies to the threshold. In general, forecasts farther from the threshold are known to be more confident than those which are close (with respect to a threshold of 10mm, typically a 100mm forecast results in a more confidence in a 10mm exceedance than a 12mm QPF).

5.1.1.2 Point Uniform Ranks (PUR)

Point Uniform Ranks attempts to alleviate some of the glaring issues with PDV by using the quantitative (as opposed to binary) Q-Θ relationship to generate continuous FP estimates and reduce some of the associated FP overconfidence. Instead of just using the fraction of members exceeding Θ, PUR adjusts the PDV FP based on how close the *surrounding members* are to the threshold. Here, the

surrounding members at a point (y,x) are defined to be the ensemble member with the highest Q_{yx} not exceeding $\theta_{N_{yx}}$ and the member with the lowest Q_{yx} exceeding $\theta_{N_{yx}}$. These will be denoted SM_{low} and SM_{high}, respectively.

Before calculating any FPs, ensemble member QPFs are normalized with respect to their local critical thresholds: $Q'_{yxk} = \frac{Q_{yxk}}{\theta_{yx(k)}}$. This is done both to appropriately rescale members in the event that different local thresholds are applied when member-by-member QMBC is applied, but also because the ratio tends to have a better correspondence with the departure from a critical threshold for positive definite variables than would an absolute difference. An event and the surrounding members are accordingly redefined as:

$$E'_{yxk} = \begin{cases} 1 & Q'_{yxk} \geq 1 \\ 0 & Q'_{yxk} < 1 \end{cases}; \; SM'_{low} = \frac{Q_{yxSM_{low}}}{\theta_{N_{yxSM_{low}}}}; \; SM'_{high} = \frac{Q_{yxSM_{high}}}{\theta_{N_{yxSM_{high}}}}$$

Following the PUR goal to adjust PDV FPs based on member proximity to the critical threshold, it should never be the case that the FP(PUR) adjustment results in assigning a probability higher or lower than the surrounding PDV-based probabilities. For example, if two of ten ensemble members forecast an event, resulting in FP(PDV)=0.2, FP(PUR) should never be less than 1/10 or greater than 3/10, as this would imply a different number of members forecasting the event. There are always n+1 possible PDV probabilities, resulting from 0 to n members forecasting the event of interest. According to these principles, FP(PUR) should be FP(PDV)*$\frac{n}{n+1}$ + $\frac{1}{n+1}$*some measure of proximity to SM'_{high} relative to SM'_{low}. In the context of ratios, as SM's are defined, determining proximity implies the use of logarithms rather than absolute difference (10 being equally close to 1 as 0.1, for example). The set of surrounding members (SMs) is said to be *deficient* when no member satisfies the criterion of either the SM_{low} or SM_{high} definition, or equivalently, if all or no member(s) exceeds the critical threshold. When the SMs are not deficient, the PUR FP may be readily assigned:

$$FP_{yx}(PUR) = \frac{\sum_k E'_{yxk}}{n+1} + \frac{1}{n+1}\frac{\log SM'_{high}}{\log \frac{SM'_{high}}{SM'_{low}}}$$

When the SMs are deficient, the same formula cannot be applied, since the latter term on the right hand side is not defined. It is first necessary to define the desired FP behavior in these intervals of probability space. First, it is intuitively desirable that $FP_{yx}\left(PUR|\ \forall_k(Q_{yxk} = 0)\right) = 0$; when no member forecasts any precipitation at all, the FP reduces to 0. In the event of no member forecasting exceedance, it is desirable that as the highest QPF member approaches the critical threshold, the FP approaches $\frac{1}{n+1}$ to ensure no discontinuity in probability space: $FP_{yx}\left(PUR|\ \lim_{\to 1} SM'_{low}\right) = \frac{1}{n+1}$. Similarly, when all members forecast exceedance, as the lowest QPF member approaches the critical threshold, the probability should reduce to $\frac{n}{n+1}$: $FP_{yx}\left(PUR|\ \lim_{\to 1} SM'_{high}\right) = \frac{n}{n+1}$. Lastly, although loosely defined, it is desirable that the change on the intervals [0,1/(n+1)] and [n/(n+1),1] is continuous, monotonic, and proportional as SM'_{low} and SM'_{high} vary between [0,1) and [1,∞], respectively. Right-Skewed Distributions (RSDs) (see section 2.4) may be reasonably and appropriately applied to ensure this smooth transition, and to ensure the 0 FP limit when all QPFs are 0, an RSD defined only for nonnegative values must be employed. The Gamma Distribution meets these criteria and is chosen here. The Gamma CDF is reproduced here in the nomenclature of this chapter from section 2.4.3 for convenience:

$$F_\gamma(x;\ \alpha, \beta) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} = \frac{\int_0^{\beta x} q^{\alpha-1}e^{-q}dq}{\int_0^{\infty} q^{\alpha-1}e^{-q}dq}$$

The Gamma CDF parameters may be estimated using the method of moments (MoM; section 2.4.4.1) and the ensemble member QPFs:

$$\widehat{\alpha_{yx}} = \frac{\overline{Q'_{yx}}^2}{s_{Q'_{yx}}^2} \qquad \widehat{\beta_{yx}} = \frac{\overline{Q'_{yx}}}{s_{Q'_{yx}}^2}$$

Combining all of the above yields the following equation summarizing PUR-based FP estimation:

$$FP_{yx}(PUR) = \begin{cases} \left(\frac{1}{n+1}\right)\frac{1 - F_\gamma(1;\, \widehat{\alpha_{yx}},\, \widehat{\beta_{yx}})}{1 - F_\gamma(SM'_{low};\, \widehat{\alpha_{yx}},\, \widehat{\beta_{yx}})} & \sum_k E_{yxk} = 0 \\ \frac{\sum_k E_{yxk}}{n+1} + \frac{1}{n+1}\frac{\log SM'_{high}}{\log \frac{SM'_{high}}{SM'_{low}}} & 0 < \sum_k E_{yxk} < n \\ \frac{n}{n+1} + \left(\frac{1}{n+1}\right)\left(1 - \frac{F_\gamma(1;\, \widehat{\alpha_{yx}},\, \widehat{\beta_{yx}})}{F_\gamma(SM'_{high};\, \widehat{\alpha_{yx}},\, \widehat{\beta_{yx}})}\right) & \sum_k E_{yxk} = n \end{cases}$$

A schematic comparing the PDV and PUR methods is shown in Figure 5.1 for visualization. Eight ensemble member forecasts are shown; it is evident how the PUR method (appropriately) reduces the forecast's sharpness.

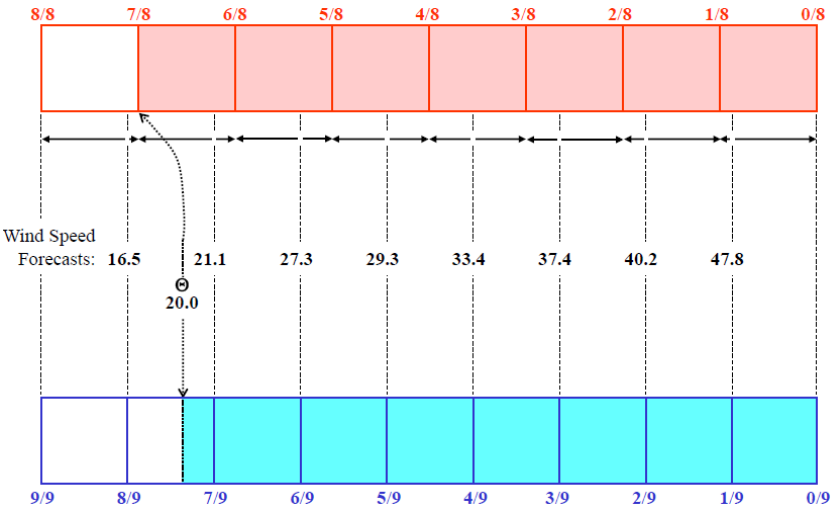

Figure 5.1: Schematic comparison of the PDV and PUR methods. PDV forecast quantities are shown in red, while PUR forecast quantities are shown on the bottom in blue. Adapted from Eckel (2003).

5.1.1.3 Neighborhood Democratic Voting (NDV)

As previously noted, models often have displacement errors in the location of precipitation features. Both PDV and PUR use only model QPFs collocated with the forecast location, and as such, FPs

may be too low when models correctly predict the existence of an extreme precipitation feature, but incorrectly displace it in space. They may conversely be too high over locations where the model solutions forecast an event, since PDV and PUR methods assume complete confidence in the feature location predicted in the model. Nearby, or neighboring, points often have very similar characteristics and due to the unpredictability of extreme precipitation, forecast values at these points can often serve as nearly or even equally good forecasts as the QPF collocated with the forecast point. This notion- using neighborhoods of radius *r* about the forecast location and, in a sense, forming an ensemble of forecasts from a single model run in so doing- has been used effectively in previous applications in the literature (e.g. Theis et al. 2005, Schwartz et al. 2010), and will be explored in more detail here.

Neighborhood Democratic Voting (NDV) is a simple extension of PDV by applying neighborhoods to each of the original ensemble members. NDV probabilities are a function of the neighborhood radius *r*; the choice of *r* can substantially affect FPs and must be tuned with care. Under NDV, an event for an ensemble member may be redefined as:

$$E'_{yxk} = \frac{\sum_{a=y-r}^{y+r} \sum_{b=x-r}^{x+r} E_{abk}}{r^2}$$

And the associated FP becomes:

$$FP_{yx}(NDV) = \frac{\sum_k E'_{yxk}}{n}$$

NDV still produces discrete probabilities like PDV, but the number of FP options increases from *n*+1 to nr^2, often two orders of magnitudes larger as a function of the increase in effective ensemble members. This greatly alleviates the discretization problem and results, at least theoretically, in more appropriately smoothed probability fields that more closely reflect sampling from the true PDF. Like PUR, NDV

theoretically reduces forecast *sharpness* when it is appropriate to do so, and substantially improves forecast *reliability*.

NDV does have a few drawbacks. NDV can incorrectly smooth the FP field when spatial uncertainty on a precipitation feature is low. It can also inappropriately influence FPs by applying forecasts from a neighboring point that dos not exhibit similar characteristics to the forecast location. This can occur in complex terrain, near bodies of water, or when the neighborhood radius becomes too large. Appropriately tuning the neighborhood radius requires performing cross-validation or some similar technique. Doing this requires the availability of historical forecast data, and can be computationally expensive to accurately tune. NDV is also more computationally expensive than PDV in real-time forecasting, but this is negligible in most settings.

5.1.1.4 Neighborhood Uniform Ranks (NUR)

Analogous to NDV, Neighborhood Uniform Ranks (NUR) is a simple neighborhood extension to the PUR algorithm. NUR can theoretically combine some of the individual advantages presented by the PUR and NDV algorithms. However, implementation of the neighborhood extension is not quite as straightforward as with democratic voting; some design choices must be considered. Unlike NDV, whose FP estimates are not altered with the addition of redundant data so long as such data is proportionally sampled (such as taking adjacent grid points which are so correlated so as not to provide sufficient new forecast information), in the limit as number of effective ensemble members gets large, NUR FPs reduce to NDV FPs, thereby eliminating some of the PUR-based advantages. In order to appropriately retain these, it is important that each ensemble member be sufficiently uncorrelated with other members so as to be introducing new information to the ensemble. To accomplish this, in addition to a neighborhood radius *r*, a spacing *s* between sampled grid points may be instituted. This

also must be tuned, most effectively through cross-validation, and can thus be more expensive to implement.

There are three basic approaches to applying neighborhoods to PUR. The first approach applies a neighborhood to each ensemble member individually, forming *n* ensembles of size $\left(\frac{2r}{s}+1\right)^2$. PUR is then applies to each ensemble member individually to form member-by-member FPs. These are then averaged to generate the final FP; this method is termed Deterministic Neighborhood Uniform Ranks (DNUR). The second approach treats each neighborhood point separately, with the point-by-point QPFS of the different models forming the ensemble members of each 'mini-ensemble'. This thus generates $\left(\frac{2r}{s}+1\right)^2$ ensembles of size *n*. Applying PUR to each mini-ensemble yields $\left(\frac{2r}{s}+1\right)^2$ FPs; these are averaged to yield the final FP. This method will be denoted Neighborhood Point Uniform Ranks (NPUR). Finally, the last approach, called Ensemble Neighborhood Uniform Ranks (ENUR), places all candidate forecasts in the same ensemble, forming a single ensemble of size $n\left(\frac{2r}{s}+1\right)^2$ as in NDV. PUR is applied to this ensemble to determine the FP. It should also be noted that ENUR, which operates on the largest size ensemble, is therefore the fastest to converge to NDV.

5.1.2 Intermediate Algorithms (INTAs)

Unlike NIVAs, INTAs relax the assumption that ensemble member solutions are equally likely to verify as truth. This is not a good assumption when ingesting guidance from a vast array of different data sources; models have different resolutions, numerics, physics, and initializations which all affect model skill and can determine whether a model can simulate certain extreme precipitation features at all. As such, developing weights to quantify the discrepancy in predictive power between the models comprising the ensemble has the potential to significantly improve FP-derived ensemble skill. Determining appropriate weights requires only a historical record of model performance for each

member comprising the ensemble; simple binary forecast and observed hit/miss records are all that is required for this training. So, though a longer data record is still advantageous in improving the skill assessment of the ensemble members, in general it is hypothesized that acceptable weights may be obtained using less historical data than is required for ADVAs, discussed below.

A multitude of possible ways exist to weight the ensemble members. It is instructive in designing a weighting scheme to considered desired behavior, particularly in limiting cases. Ultimately, it is desirable that member weights are a function of two factors: 1) how skillful is the model, and 2) how much new, independent forecast information is the member bringing to the PFS? Intuitively, more skillful models should be allocated more weight, and those models which are introducing more new information- information which is not redundant with that provided by another member- should also be given more weight. Considering the limiting cases of member skill, a perfect model should receive all of the weight; if one member is perfect, there is no reason to consider any guidance, as it can only degrade forecast skill. Conversely, a model with no skill should receive no weight, since its forecast by definition has no bearing on the potential verification. With regard to limiting cases for new forecast information, inserting a carbon copy of another ensemble member, which thus introduces no new forecast information to the PFS, should have no effect on the total weight of that model run. Using historical correlation coefficient between ensemble member precipitation time series as a proxy for the new information introduced by an ensemble member, it follows that the insertion of a carbon copy of an existing member (which then have a correlation coefficient of unity) should result in both the original member and its copy having their weights accordingly halved so as to preserve the weight of the underlying model in the ensemble. Applying these concepts gives rise to the following equations for an ensemble member *m*'s weight in an ensemble of size *n*, with α and β exponents tunable via cross validation:

$$W_{corr_{yx_{m_1 m_2}}} = 1.0 - \frac{corr\left(m_1(y,x,:), m_2(y,x,:)\right)^{\alpha}}{2}$$

$$W_{corr_{yx_m}} = \prod_{k=1}^{n} W_{corr_{yx_{mk}}}$$

$$W_{fss_{yx_m}} = \frac{1.0}{(1.0 - FSS_{yx_m})^{\beta}} - 1.0$$

$$W'_{yx_m} = W_{corr_{yx_m}} W_{fss_{yx_m}}$$

$$W_{yx_m} = \frac{W'_{yx_m}}{\sum_{k=1}^{n} W'_{yx_k}}$$

5.1.3 Advanced Algorithms (ADVAs)

ADVAs have the unique ability to directly calibrate FPs and to directly correct for model biases: displacement biases, quantitative biases, and meteorological biases and errors. They can be expensive to train, much more so than NIVAs or INTAs, and accurate calibration and bias correction requires a long consistent data record of both observations and model simulations for each ensemble member. ADVA algorithms can be quite complex and unintuitive. The ADVA techniques examined here are described in detail in Section 2.6; please refer there for details about algorithm design and the parameters that require tuning in each case.

5.2 Results[4]

Figure 5.2 shows FSS results for DV methods at a variety of evaluation and neighborhood radii, and for different ensemble compositions. It should be noted that the 1 member ensemble corresponds to the inclusion of just the NSSL-WRF in the ensemble, with increasing ensemble sizes including

[4] Results for the research described in this chapter are still forthcoming. Cross-Validation training and tuning is incomplete; results are presented to the extent available.)

additional GEFS/R members, beginning with the control member. There are numerous important aspects to glean from inspection of this figure. One apparent observation is the aforementioned downside of NIVAs: excessively adding new information with positive, but lower skill than the existing forecast information will degrade system performance. This is seen here by inspecting FSSs as a function of ensemble composition. Regardless of the neighborhood or evaluation radius examined, the same general trend is observed. Adding the control member to the ensemble improves forecast skill over just using the NSSL-WRF; having no GEFS/R members in the ensemble, the insertion of one adds considerable new forecast information to the ensemble and skill is enhanced in spite of the inferior skill of the GEFS/R control member to the NSSL-WRF. However, adding subsequent GEFS/R members, each of similar skill to the control run (see Figure 5.3), begins to degrade forecast system skill back towards the skill of an individual deterministic GEFS/R member, since the weight applied to GEFS/R is continually increased with the addition of each new member. Thus, when applying a NIVA in a PFS, it is essential that care is applied to avoid needlessly degrading system skill with the addition of redundant or inferior forecast information. At the highest neighborhood radii, even adding one GEFS/R member degrades forecast skill at the lowest evaluation radii (Figure 5.2a).

A second important observation is the impact of applying neighborhoods. The pattern is strikingly similar to the behavior as a function of ensemble composition. Changing from PDV to NDV with a small neighborhood radius, for example 10 grid boxes, considerably enhances PFS skill. However, at some larger radius, the trend reverses, and continuing to increase neighborhood radius harms PFS skill. Among the neighborhood radii examined, for small ensemble compositions, skill was maximized at a neighborhood radius of 20 grid boxes, while for the larger ensembles, a radius of 10 optimized results. There are two likely reasons that higher neighborhood radii are more effective at smaller ensemble sizes. First, here, smaller ensemble size indicates a higher proportion of forecast information coming from the NSSL-WRF, which almost certainly exhibits different spatial bias characteristics than the

GEFS/R. Further, owing to its smaller horizontal grid spacing, resolves more precipitation features, and thus expanding the neighborhood radius yields inclusion of more new forecast information for the NSSL-WRF than the GEFS/R. Secondly, at larger ensemble sizes, the PFS has already included collocated forecasts from new model simulations that are often- at least in expectation- more likely to verify as truth than some forecast value that is never forecasted in collocation with the forecast point. Increasing the neighborhood radius for these large ensembles thus experiences a "saturation effect" of sorts; by virtue of the larger ensemble size, you already have ample and sufficient forecast information to work from, and the inclusion of more distant points that are less likely to verify as truth and thus of inferior skill will inevitably result in a decline in overall skill. At the highest neighborhood radii, NDV actually underperforms PDF as a result of these effects.

Lastly, it is important to note the skill effects as a function of evaluation radius. Because the application of neighborhoods does not actually change the total probability, instead simply acting to redistribute it, the effect of neighborhoods at the large evaluation radii are small since the effects of probability redistribution are small (the redistributed probability still gets summed in the same box). At small evaluation radii, however, where probability redistribution more often affects what probability is included in each evaluation box, the effects of neighborhood application are large. This can be seen in Figure 5.2, where in panel (a) at an evaluation radius of 4 grid boxes, changes from the deterministic NSSL-WRF forecasts are as large as 25%, but at the larger 40 grid box radius in panel (d), changes are no larger than 8%. It should also be noted that optimal neighborhood radius changes as a function of evaluation radius. As noted above, at the lowest 4 grid box evaluation radius (Figure 5.2a), the optimal neighborhood radius was 10 or 20, depending on ensemble size. However, at the 40 grid box radius as seen in Figure 5.2d, higher 30 or 20 grid box neighborhood radii were found to maximize forecast skill. In contrast to NDV, UR does act to directly alter, rather than simply redistribute, probability, and thus the implications for PUR and NUR in contrast to PDV are different than seen with NDV here.

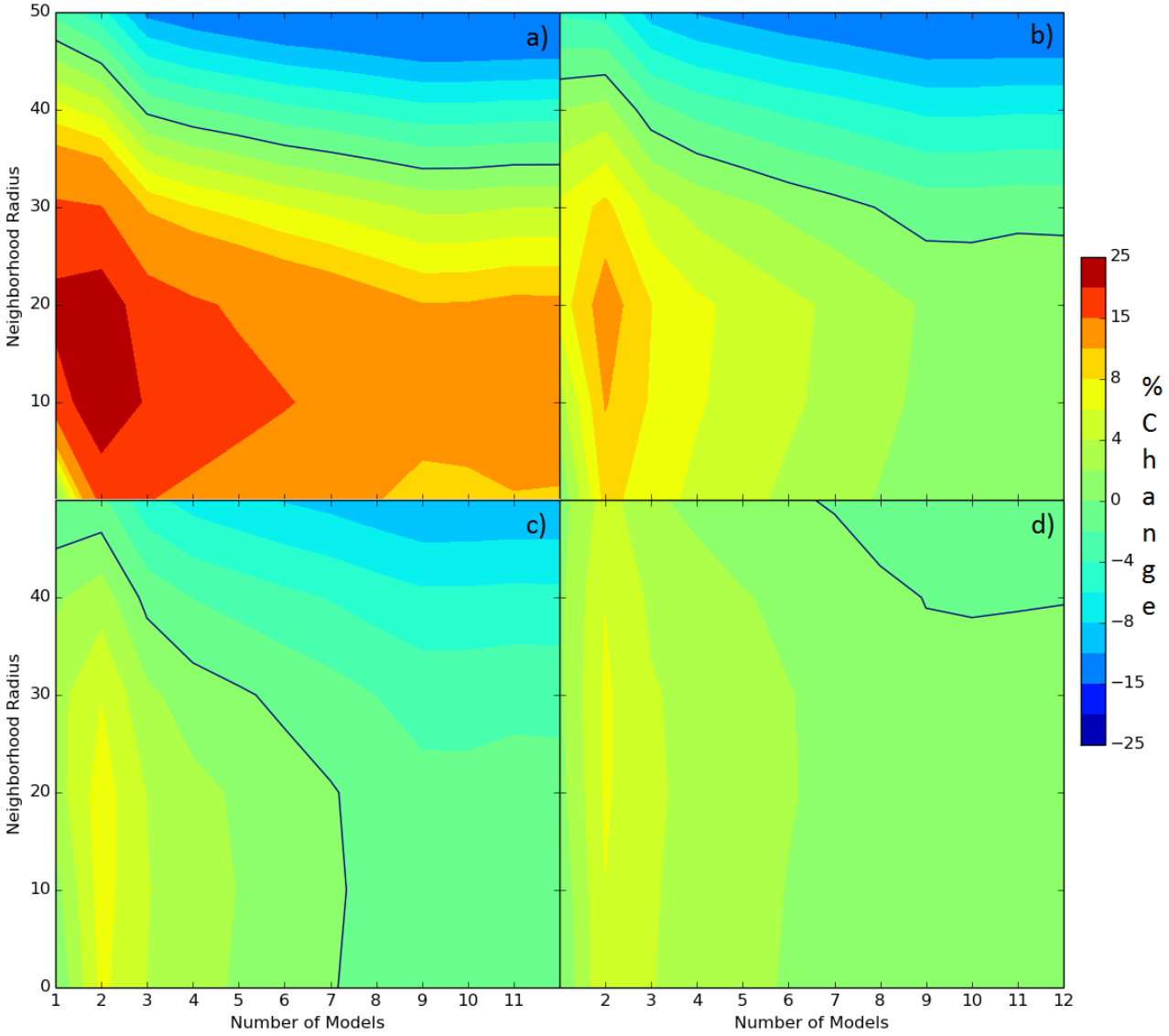

Figure 5.2: Fractions Skill Score differences compared against the deterministic NSSL-WRF, expressed as a percent change. Panel (a) corresponds to a 4 grid box evaluation radius, (b) to a 16 grid box evaluation radius, (c) a 28 grid box evaluation radius, and (d) a 52 grid box evaluation radius. The "1 model" column corresponds to forecasts just based on the deterministic NSSL-WRF model, with increasing model numbers including additional members of the GEFS/R, beginning with the control member. Neighborhood radii on each panel's ordinate axis correspond to neighborhood grid box radius. The dark contour denotes the 0% change line.

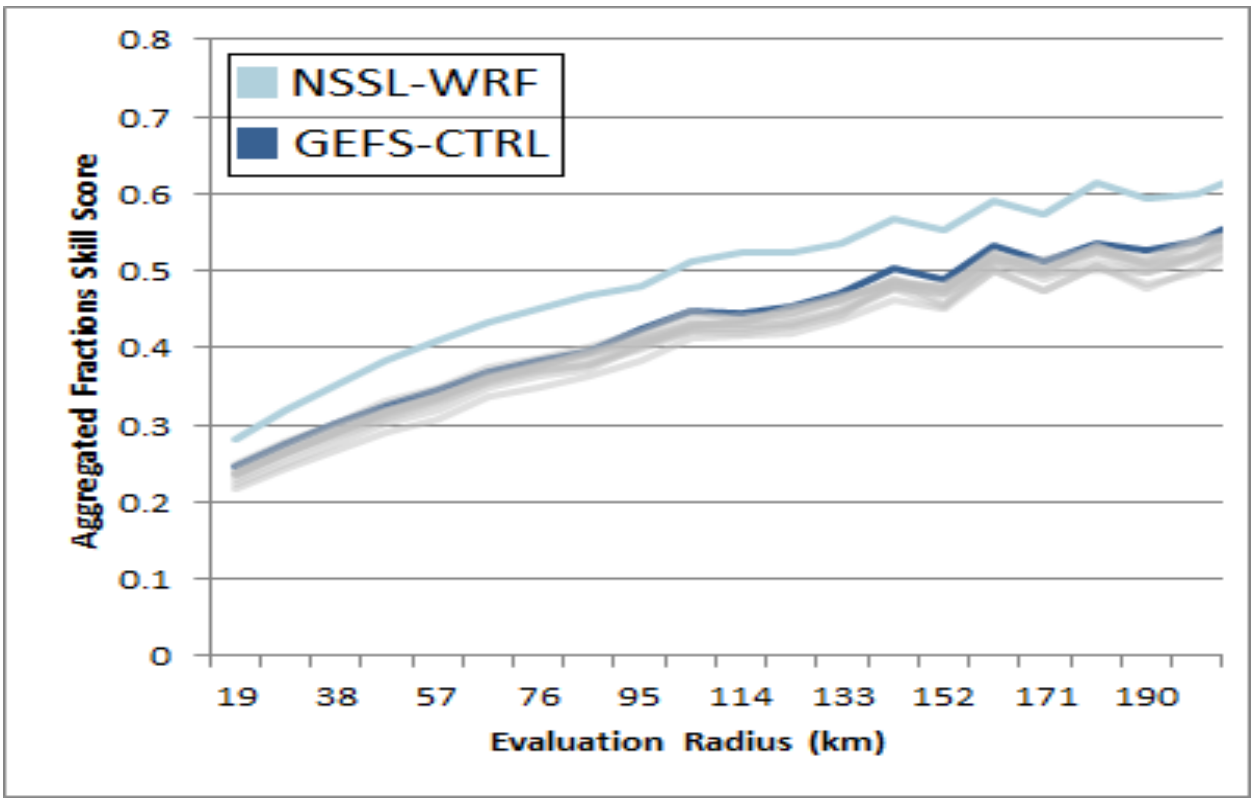

Figure 5.3: Fractions Skill Score comparison of the deterministic models used in this study (also the forecast obtained by PDV on just that one model). The dark blue corresponds to verification on the GEFS/R control, light blue for the NSSL-WRF, and grays correspond to other members of the GEFS/R.

Reliability diagrams themselves do not provide an especially effective mechanism to quantitatively compare many different probabilistic forecast systems; visualization is cumbersome (one line is necessitated for each forecasting method), and comparison is traditionally performed qualitatively. Nevertheless, forecast reliability, is a highly important property of probabilistic forecasts. To aid with reliability analysis and comparison, several summary statistics have been developed.

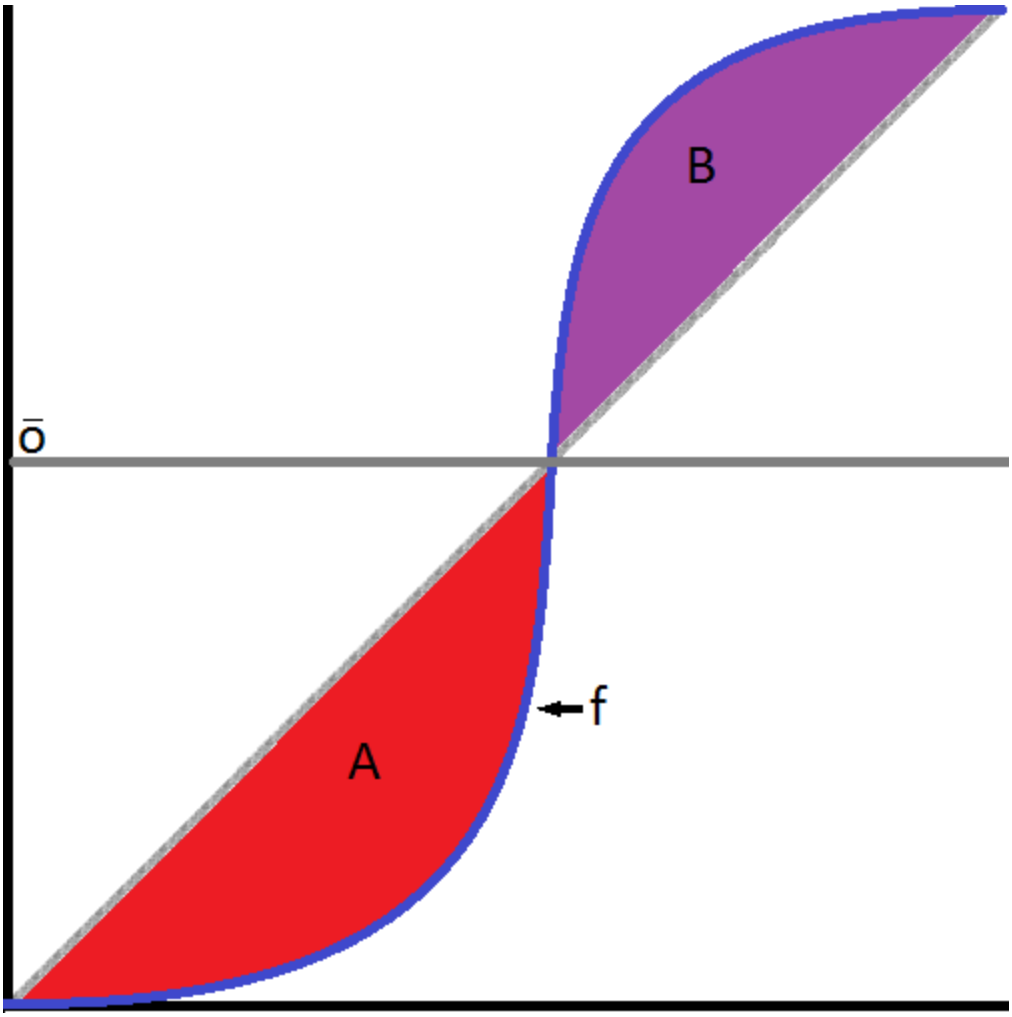

Figure 5.4: An example schematic of a reliability diagram, for reference. $\bar{o}$ corresponds to the climatological frequency, f is the reliability curve, and the light gray line is the one-to-one FP=ORF line.

First, is the reliability skill score (RSS), defined as:

$$RSS = 1.0 - \frac{RS}{RS_{ref}} = 1.0 - \frac{\int_0^1 |f(x) - x| g(x) dx}{\int_0^1 |f_{ref}(x) - x| g_{ref}(x) dx}$$

Where $f_{ref}(x)$ is a reference forecast method, and g(x) refers to the probability density of the forecast system assigning FP x. By definition then, $\int_0^1 g(x)dx = \int_0^1 g_{ref}(x)dx = 1$.

Second, is the reliability bias score (RBS), defined as:

$$RBS = -\int_0^1 (f(x) - x)g(x)dx$$

Last, is the reliability confidence score (RCS), defined as:

$$RCS = -\int_0^1 (f(x) - x)\frac{x - \bar{o}}{|x - \bar{o}|}g(x)dx$$

Of course, we do not have an actual function specifying ORF as a function of FP. Instead, bins are used, and the statistics are approximated by means of a Riemann sum:

$$RSS = 1.0 - \frac{\sum_{i=1}^{n}\left|\bar{f}\left(\bar{x}(i)\right) - \bar{x}(i)\right| G\left(\bar{x}(i)\right)\Delta x(i)}{\sum_{i=1}^{n}\left|\overline{f_{ref}}\left(\bar{x}(i)\right) - \bar{x}(i)\right| G_{ref}\left(\bar{x}(i)\right)\Delta x(i)}$$

Where G() is now the probability mass function and Δx() is the bin width: $\sum_{i=1}^{n} G\left(\bar{x}(i)\right) = \sum_{i=1}^{n} \Delta x(i) = 1$.

$$RBS = -\sum_{i=1}^{n} \bar{f}\left(\bar{x}(i)\right) - \bar{x}(i)\, G\left(\bar{x}(i)\right)\Delta x(i)$$

$$RCS = -\sum_{i=1}^{n} \left(\bar{f}\left(\bar{x}(i)\right) - \bar{x}(i)\right)\frac{\bar{x}(i) - \bar{o}}{|\bar{x}(i) - \bar{o}|}G\left(\bar{x}(i)\right)\Delta x(i)$$

The schematic drawn in Figure 5.4 could be characterized, assuming |A|=|B| and the FP distribution is approximately uniform, as an unbiased, but underconfident; it forecasts FPs too close to climatology ($\bar{o}$) resulting in ORFs corresponding to near-climatology FPs deviating more from it than the FPs. This would be reflected in the statistics, since, assuming the PMF and bins are fairly symmetrical, RBS reduces to an approximation proportional to A + B = 0, indicating no bias, while RCS reduces proportionally to approximately -1*(-1*A + B) < 0, indicating *underconfidence*. The RS, approximately proportional to |A|+|B|, is positive, reflecting that the forecast system is not completely reliable- deviations are observed between FP and ORF. RSS behaves as any skill score- a perfectly reliable

forecast system exhibits an RSS of 1, and the reference by definition has an RSS of 0, with smaller RSS values indicating less reliability. RBS has a sign convention such that positive numbers indicate positive bias (FP > ORF) and negative numbers indicate negative bias. RCS has a sign convention such that positive numbers indicate overconfidence, and negative numbers indicate underconfidence.

Examining the summary statistics for DV forecasts in Figure 5.5, unsurprisingly, the pattern is very different than with PDV. Here the deterministic NSSL-WRF, inherently unreliable due to its binary forecasting capability, is used as the reference forecast. RSS can be seen in Figure 5.5a; in general continuously improved by increasing neighborhood radius. This makes sense: high FPs in large radius NDV indicate very large extreme precipitation features, and owing to less sensitivity to spatial or temporal displacement errors, these will correspondingly have higher ORFs than small-scale precipitation features which happen to be collocated in several models. PDV would not be anticipated to yield especially reliable forecasts unless the ensemble size was both exceedingly large, and all members were sampled from the true forecast PDF. As a function of ensemble composition, the RSS behavior is more like the FSS as seen in Figure 5.2. Improvement in RSS is seen with the addition of the first few GEFS/R members. After this point, at low neighborhood radii, little change is seen with respect to RSS with the addition of subsequent GEFS/R members into the ensemble. At higher neighborhood radii, PFS reliability is actually slightly degraded by more GEFS/R members, as reflected by the reduced RSS. The actual reliability diagrams and their corresponding sharpness diagrams appear in Figure 5.6. All PDV forecasts exhibit the same qualitative behavior: negative bias at the lowest RPs (frequent FP of 0, but occasionally events do occur despite an FP of 0, resulting in a non-zero corresponding ORF), and positive bias at most non-zero FPs. The low end FP behavior holds, but is alleviated, with increasing neighborhood radius in NDV (see Figure 5.6b2). The behavior at high FPs is still fairly monotonic, but is much more pronounced, and at the highest neighborhood radii of 40 and 50 grid boxes, high FPs above 25-30% actually occur at a higher ORF than the FP (Figure 5.6a). Thus, by simply inspecting the reliability

diagrams from afar as in Figure 5.6a, one would anticipate that the RSS should be the highest for middling neighborhood radii, as this yields the best calibration along the FP=ORF line for most probabilities. However, inspecting any of the sharpness diagrams (Figures 5.6c-5.6h), it is, unsurprisingly given the climatological event frequency, evident that the vast majority of forecasts issue an FP of 0, regardless of the forecast algorithm employed. Thus, scrutiny of the ORF when an FP of 0 is issued is of considerable importance, despite being undiscernible from Figure 5.6a. A blowup of the full reliability diagram for smaller FPs is provided in Figure 5.6b, and an even larger zoom to when FP=0 is provided in Figure 5.6b2. From close scrutiny of Figure 5.6b2, one sees that the ORF at FP=0 for the deterministic NSSL-WRF is 3e-4, while it is nearly an order of magnitude lower near 5e-5 for a full ensemble using NDV with a 50 grid box neighborhood radius. A middling radius of 20 yields an ORF around 1e-4, still twice that of NDV50, and quite considerable when noting that this difference accounts for approximately 99% of forecasts, depending on the exact forecasting algorithm.

Bias, as discerned from RBS in Figure 5.5b, is positive for almost all DV forecasts, indicating that, in general, events occur at a lower ORF than the corresponding FPs would indicate. Looking at PDV on the deterministic NSSL-WRF, for example, this is not at all surprising; when an FP of 1 is issued (simply indicating that the NSSL-WRF QPF exceeded the 2-year RPT at that point), the event is only observed to occur about 10% of the time (see Figure 5.6a). This staggering discrepancy more than compensates for the comparatively miniscule negative bias seen at FP=0 (0.9 vs. 3e-4), even despite the greatly increased frequency of FP=0 forecasts compared with FP=1. Even with small neighborhood radii, the positive bias at high FPs is exceedingly large, and ends up dominating the bias despite low frequency of occurrence (see Figures 5.6d-5.6f). By the highest neighborhood radii, despite a change back towards a positive bias at low non-zero FPs (see Figure 5.6b), the now neutral (NDV30, NDV40) to even negative (NDV50) bias at the high FPs results in a near-zero bias, and a small negative net bias is observed for some ensemble compositions for a neighborhood radius of 50. The confidence scores, or RCS, as depicted in Figure 5.5c,

are very similar to the bias scores in these cases. The values do, however, stay positive for all DV composition/radii pairings assessed, indicating ubiquitous overconfidence from DV. Overconfidence here refers to $\begin{cases} FP < ORF \ \ when \ FP < \ \bar{o} \\ FP > ORF \ when \ FP > \ \bar{o} \end{cases}$. This is not a particularly surprising outcome based on theory discussed surrounding Figure 5.1. However, use of neighborhoods does greatly improve the overconfidence, with overconfidence near 0 at the highest radius NDVs.

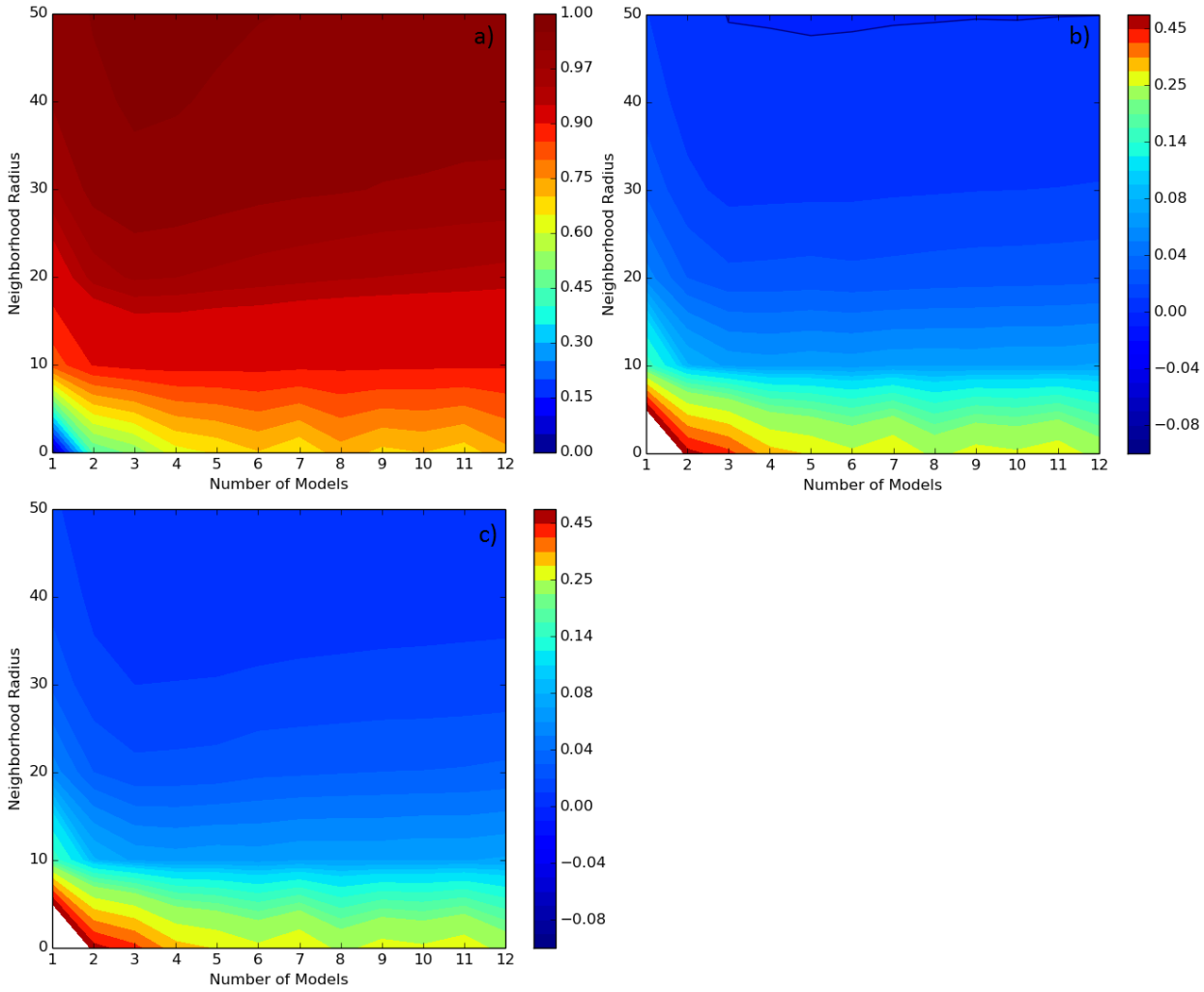

Figure 5.5: Summary statistics for DV algorithms forecast reliability. Panel (a) plots RSS, (b) plots RBS, and (c) RCS; all are described in text. Axes are as described in Figure 5.3. Solid contour in panel (b) corresponds to RBS = 0.

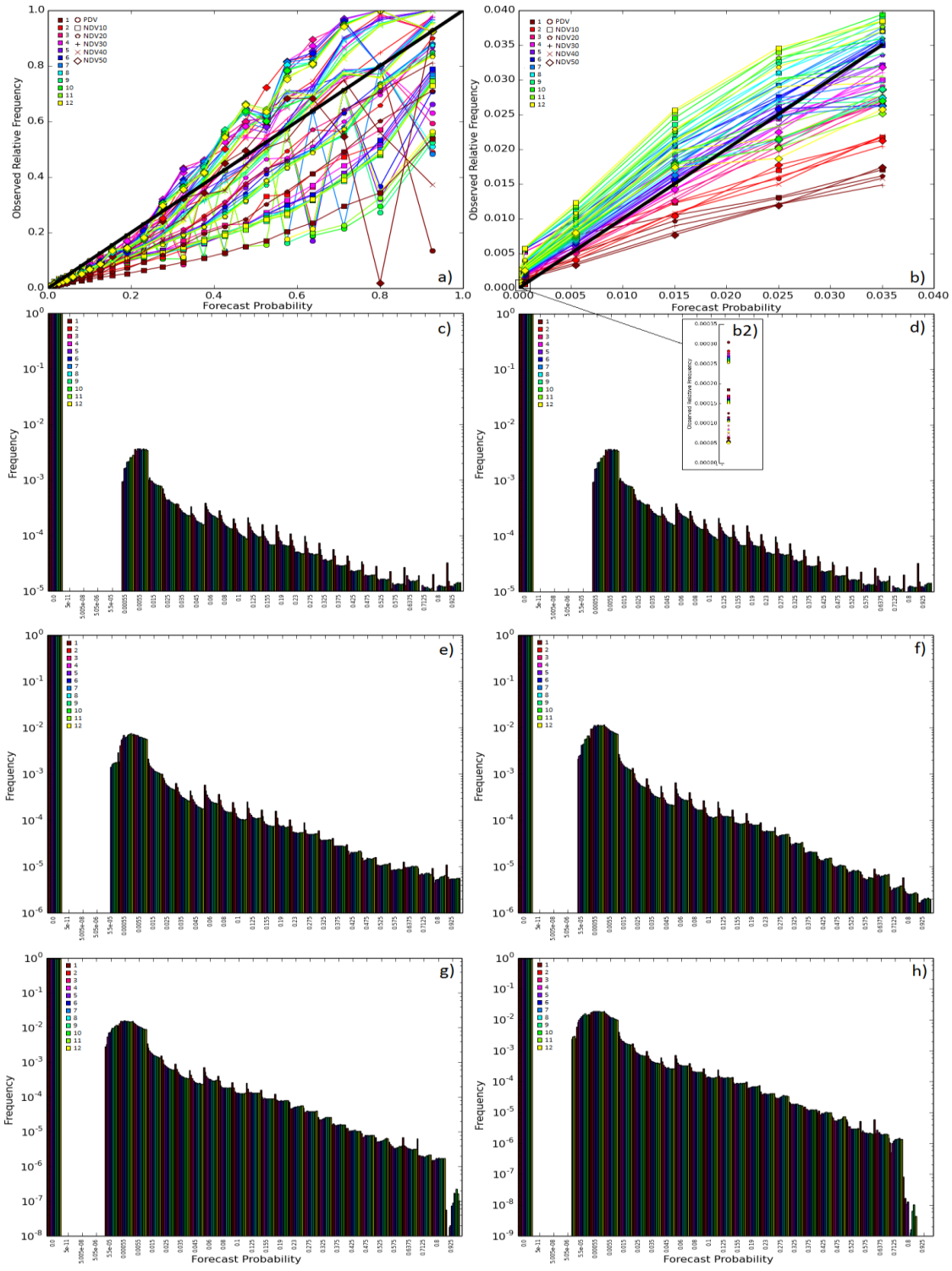

Figure 5.6: Reliability and sharpness diagrams for each of the Democratic Voting forecast methods discussed in text. Panel (a) plots the reliability diagrams, (b) is a version of (a) zoomed for small FPs, and panel (b2) is further zoomed for when FP=0. Panels (c)-(h) are sharpness diagrams, indicating the frequency of the each forecasting algorithm generating a forecast within each specified probability bin. The panels (c)-(h) correspond to PDV, NDV with a neighborhood radius of 10 grid boxes, NDV with radius 20, NDV with radius 30, NDV with radius 40, and NDV with radius 50, respectively.

Finally, forecast value for the DV ensemble configurations are assessed by means of the Value Score (VS) discussed in section 2.7. As with reliability, forecast value can be summarized by means of a Riemann sum of VSs over the spectrum of end user cost/loss ratios. This statistic, termed the summary value score (SVS) is formulated as:

$$SVS = \sum_{i=1}^{n} \overline{VS}\left(\bar{\alpha}(i)\right) G\left(\bar{\alpha}(i)\right) \Delta\alpha(i)$$

Where α is the cost-loss ratio (CLR), n is the number of discrete cost-loss thresholds considered, and G denotes the PMF of all candidate end users in cost-loss space. Here, we will assume all user CLRs are equally likely, meaning that G will be uniformly distributed. Value statistics appear in Figure 5.7, with Economic Value Diagrams (EVDGs, to distinguish from Right-Skewed Distributions) appearing in a log scale in Figure 5.7a, and the SVS statistics appearing in Figure 5.7b. It is evident that users acting based solely on the deterministic NSSL-WRF forecasts would mostly do much better basing decisions off climatology, with users with CLRs above 0.15 and below 0.00035, or roughly one third of the climatological frequency, being negatively affected by use of the forecasts. For no user is the use of the point deterministic forecasts anywhere approaching the optimal score of unity; VS peaks below 0.2 for users with a CLR equal to the climatological event frequency. This is reflected in the SVS with a staggering -4.25 value reported. Adding in GEFS/R members to the aggregate ensemble assists the SVS considerably (see Figure 5.7b), but the overall VS patterns remain similar at the low end of α's. For PDV, where FPs are highly discretized, large discontinuities in VS are observed when crossing new allowed FPs under the PFS. For example, PDV with the NSSL-WRF and GEFS/R control member jumps from a VS of under -6 to nearly 0 when crossing the α=0.5 line (light red circle line, Figure 5.7a). This is where the largest benefit from the addition of new dynamical models into the ensemble is observed: the highly insensitive users for whom protecting against extreme rainfall is nearly as costly as the losses endured

when extreme rainfall occurs and the user is not prepared. The addition of neighborhoods makes a generally larger impact across the gamut of user CLRs, with VS continuously increasing with increasing neighborhood radius for αs below 0.01. At the higher CLRs above 0.01, a different sort of behavior is observed, with the algorithms clustering more by dynamical model composition than by neighborhood radius, diverging again at the highest CLRs where each configuration exhibits different behavior in the high CLR tail.

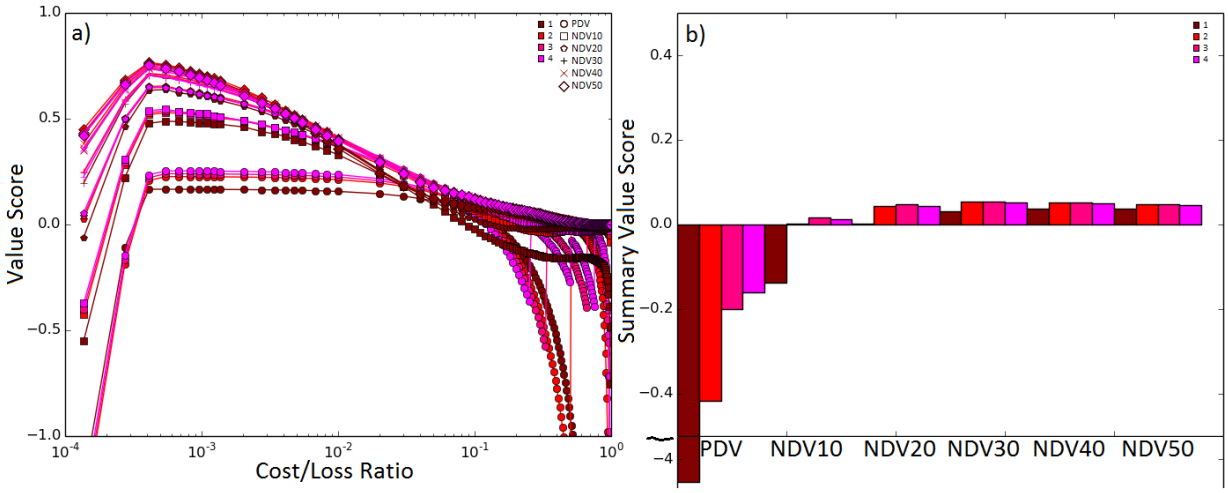

Figure 5.7: Value score comparisons for the DV algorithm ensemble configurations. Panel (a):Economic value diagrams with cost/loss ratio plotted on a logarithmic axis for clarity; panel (b) summary value scores for the configurations assuming a uniform distribution of end user CLRs.

Beyond DV approaches, results are still very much forthcoming. However, preliminary results are provided here for the interested reader. Fractions Skill Score comparisons for a variety ensemble and algorithmic configurations are presented in Figure 5.8. It can be seen by comparison of the maroon and red lines that PUR generally improves FSS when compared with PDV. The differences, however, are quite marginal except at the smallest evaluation radii, where approximately a 2% improvement is observed. A bigger difference is observed between UR and DV in a neighborhood context. FSS results for NDV are reproduced in Figure 5.8 on the magenta line; the fuchsia line, representing DNUR at the

same 20 grid box neighborhood radius as the magenta line, strictly dominates NDV at each evaluation radius. Similar findings hold at the 40 grid box radius comparing the yellow and lime green lines, though the results here are notably inferior to PDV on the 3-member ensemble, consisting with the findings presented in Figure 5.2. Some preliminary experiments with individual member weighting were also conducted. Weights were generated using the procedure discussed in Section 5.1.2 for a 3-member ensemble consisting of the NSSL-WRF and two GEFS/R members. The results of a subset of the weighting experiments appear in the blue lines in Figure 5.8. Weights for these experiments were obtained through cross-validation; as an example, the weights obtained for the NSSL-WRF in this ensemble when generated over the entire August 2009-August 2014 validation period are presented in Figure 5.9. Though, absent any smoothing as is shown here, there are some spurious anomalies, the general result is that the NSSL-WRF typically is afforded more than the *a priori* one-third weight that would be assigned to it, likely in association to the superior skill evidenced in Figure 5.3, and the fact that its forecasts differ from GEFS/R forecasts more than GEFS/R members vary between each other (not shown). The other striking feature is the generally higher weight assigned to the NSSL-WRF in the east, owing likely to its unique ability to resolve many of the small-scale convective extreme precipitation features occurring in this half of the nation when compared with the GEFS/R model. FSS being one of the two components determining the weights, the spatial gradients in the NSSL-WRF weights seen in Figure 5.9 line up well with the model skill score comparison shown in Figure 3.20 (see, for example, the Gulf Coast FSS differences). The weights in general produce mixed results at higher evaluation radii compared with not using any weights, but a several percent improvement is observed at the lowest, <50 km, evaluation radii. Shown here are results of both DNUR and ENUR at a 20 grid box neighborhood radius; both perform remarkably similarly, but DNUR does outperform ENUR at all evaluation radii. Lastly, a proof-of-concept logistic regression model was trained based on a three member ensemble consisting of the NSSL-WRF and two GEFS/R members. Each training example had 10

features: 3 predictors per model and a final feature. The predictors from each underlying ensemble member were: 1) the mean QPF/RPT ratio within a 20-grid box radius of the forecast point, 2) the QPF/RPT ratio standard deviation within a 20-grid box radius of the forecast point, and 3) the fraction of points within a 20-grid box radius of the forecast point with a QPF/RPT ratio greater than unity. The total standard deviation of all candidate forecast ratios for the entire ensemble represents the tenth feature[5]. The results, shown in the aqua line in Figure 5.8, indicate that this primitive logistic regression model did not improve over PDV for any evaluation radius examined.

Reliability characteristics for the same suite of configurations are presented in Figure 5.10. It is evident by inspection of Figure 5.10c and comparison with Figure 5.6 that UR methods are able to generate FPs in the smallest non-zero probability bins when DV, even with the use of neighborhoods, cannot. PUR exhibits some unusual characteristics in the reliability diagram in Figure 5.10a; discontinuities are observed in the ORF when FP crosses a threshold indicating exceedance by another ensemble member. This is likely attributable to an issue with the UR formulation; when *k* of *n* members forecast an event, and none of the *n-k* members not forecasting the event forecast a near-event, the event is intuitively less probable than if *k-1* members forecast an event, and the remaining *n-k-1* members all forecast a near-event. UR is, however, constrained to never create FPs above or below certain thresholds based on the proportion of ensemble members forecasting an event at the forecast location, resulting in this behavior as these FP thresholds are crossed. Aside from this, UR-based reliability diagrams tend to track quite closely with their corresponding DV reliability diagram lines. Weighting ensemble members tended to have relatively little effect at low FPs, but result in a stronger positive bias at the higher FPs, resulting in slightly inferior reliability overall as reflected in the RSS values depicted in Figure 5.10d. The ensemble configuration yielding the most unique reliability characteristics

[5] Updated ADVA implementations for the full suite of ML algorithms discussed in Chapter 2.6 is in progress, using an updated feature set and training individually on different geographic regions using those depicted in Figure 4.3. Results thus far appear promising, but are too preliminary to include here.

is the logistic regression model, which displays excellent calibration at low FPs up to about 0.2 before becoming increasingly positively biased at the higher FPs, approaching that of PDV at the highest FP bin. However, it has very low frequency of forecasting these FPs, with much- several orders of magnitude- higher frequency of forecasting in the FP region in which it is well calibrated, as evidenced by Figure 5.10c. As a result, despite inferior FSS results, the logistic regression forecasts are the best in all summary reliability statistics presented in Figure 5.10d, with the highest RSS and bias and confidence scores closest to zero.

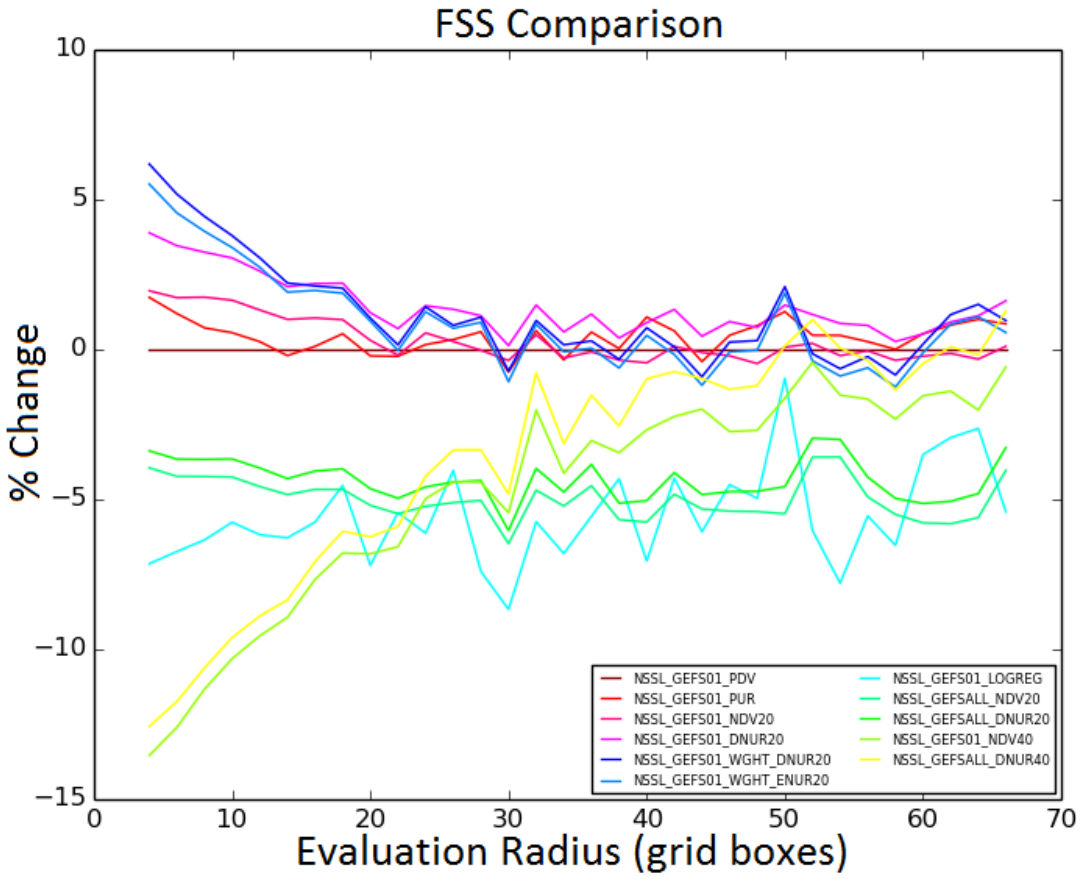

Figure 5.8: PFS FSS for a suite of various algorithmic configurations as a function of evaluation radius. Results are presented with respect to PDV on the three member ensemble consisting of the NSSL-WRF, GEFS/R control member, and GEFS/R perturbation 1.

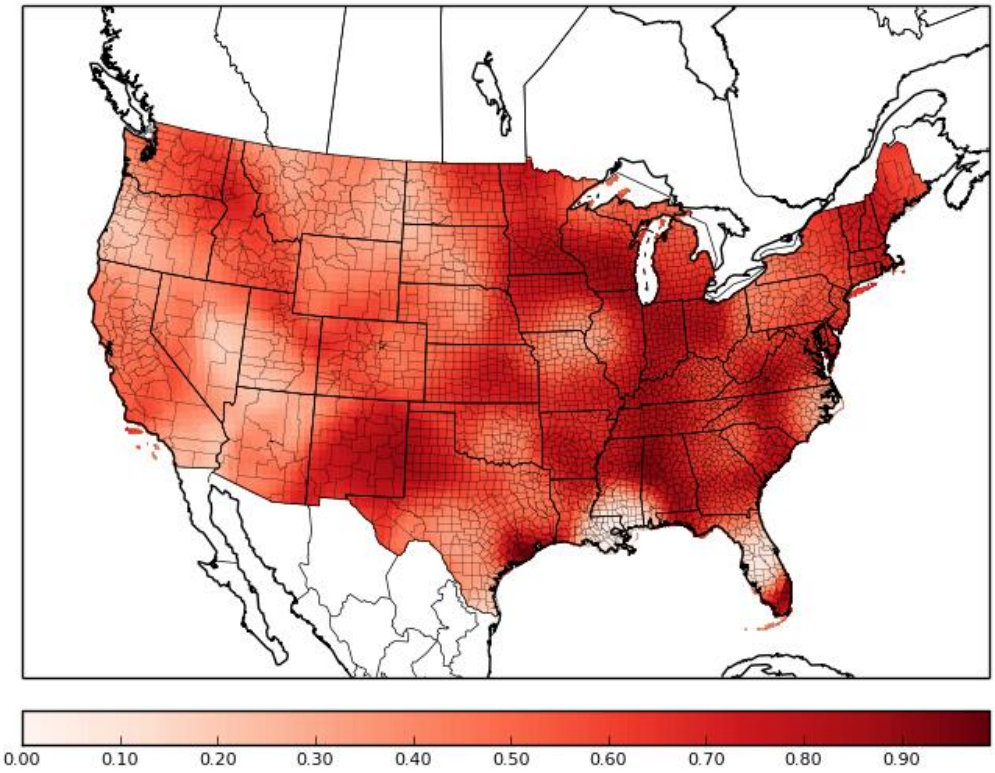

Figure 5.9: Preliminary weights for the NSSL-WRF model for those algorithms employing weights. Weights generated using the methodology described in the Section 5.1.2 of the text.

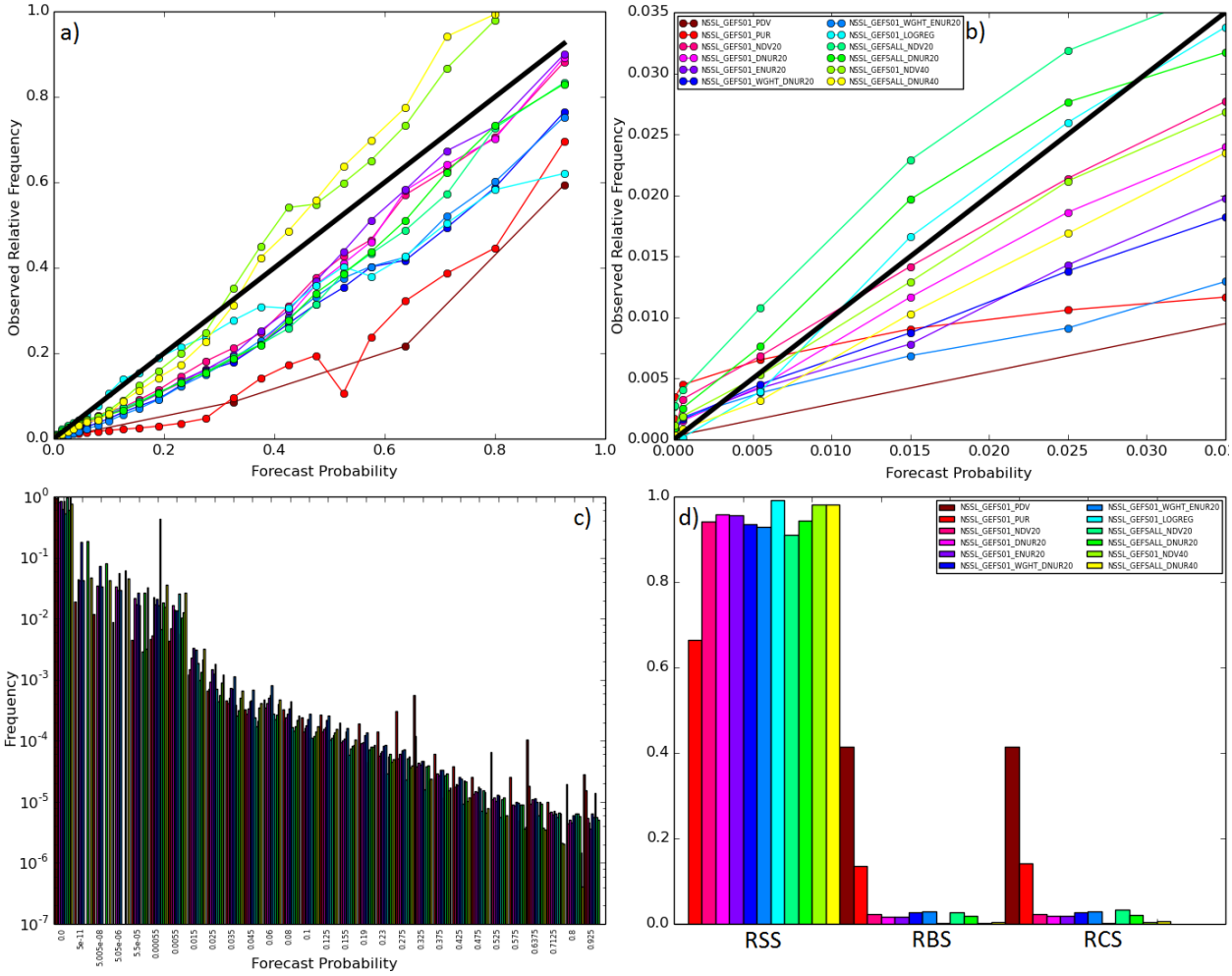

Figure 5.10: PFS reliability for a suite of various algorithmic configurations. Panel (a) shows the traditional reliability diagrams, (b) is zoomed for low FPs, panel (c) shows the corresponding sharpness diagram, and panel (d) shows the reliability summary statistics described in-text for the various algorithms. The reference for RSS in panel (d) is the 3-member ensemble PDV instead of the deterministic NSSL-WRF as used previously.

5.3 Discussion & Conclusions

Numerous findings of note may be gleaned from this preliminary work. An unsurprising but nonetheless critical aspect to note is the discrepancy in the assessment of PFS quality reached when comparing forecast skill versus reliability versus value. Inspection of just forecast skill in the DV PFSs assessed here would indicate that the best PFS uses a small neighborhood radius around 10 and includes only one of the individually inferior GEFS/R members. But evaluation of PFS reliability would reach a very different conclusion, namely that a much larger neighborhood radius of around 50 grid boxes will yield better PFS performance, as will the inclusion of a larger number of GEFS/R members (though even reliability analysis suggests that the inclusion of the full GEFS/R yields suboptimal performance). Finally, with respect to forecast value, a drastic improvement was seen in increasing from PDV to NDV with a 10 grid box radius. At low neighborhood radii, increasing the ensemble size also dramatically improves SVSs. SVS is rather insensitive to both neighborhood radius and ensemble composition at high radii, but the highest scores are obtained for a middling radius of approximately 30 grid boxes and an ensemble consisting of the NSSL-WRF and two GEFS/R members. In light of these contrasts, it is thus essential to clearly define system priorities when designing an EPS or PFS. Despite these discrepancies, there are still several overarching conclusions that may be reached. Switching from PDV to NDV with a small neighborhood radius and application of uniform ranks improve PFS performance in all statistics examined. The era of inspecting EPS information by means of traditional PDV probabilities has passed. Simple, inexpensive methods such as UR and neighborhoods can improve PFS information in every conceivable metric. There is no rational reason to apply strictly dominated probabilistic techniques when such simple and superior alternatives are so readily available.

Future work in this realm first and foremost includes three expansions. First, investigation of the NIVA and INTA methods which have been preliminarily explored herein must be completed. Results thus far, in addition to past literature (e.g. Eckel 2003, Theis et al. 2005), suggest that UR based

approaches can appreciably outperform their corresponding DV counterparts, both for point and neighborhood methods. Both intuition and preliminary findings also suggest ensemble member weighting can substantially improve PFS skill and reduce sensitivity to inferior or redundant forecast information, but insufficient research has been, as of yet, devoted to this subject. Second, a much more thorough exploration of the application of machine-learning based ADVAs towards this forecast problem is warranted. Research in this realm is currently in progress, and the results thus far, such as the first results from a highly oversimplified logistic regression model presented here, illustrate the capability of this class to do what no other algorithm classes can do. This is evidenced, for example, by the superior reliability statistics presented by the logistic regression model over any other approach of the same ensemble composition. The results from this model are considered to be approximately the baseline for ADVA-based PFS performance, and it is thought that further work in this realm will prove highly fruitful. Third, this research must be extended to other RPs. Here, only the 2-year RP has been examined, and it is certainly possible that observed performance characteristics will change for rarer events at higher RPs.

6 Forecast System in Action: A Case Study

6.1 Meteorological Overview

A confluence of factors came together to make late May of 2015 one of the wettest periods on record for the southern Great Plains, with widespread devastating flooding and flash flooding. As can be seen by inspection of Figure 6.1, even from the first three weeks of the month, much of the southern plains had been much wetter than normal, with the vast majority of the region having received at least 150% of normal precipitation and many areas, particularly northern Texas and southern Oklahoma, receiving upwards of 300% of normal precipitation. To the northwest, the central Rockies of Colorado and Wyoming were also running in the vicinity of 200% of normal precipitation over this period. These antecedent conditions led to anomalously saturated soils during late May 2015 in the south and central US, leading to increased susceptibility to flooding and flash flooding in association with heavy rainfall. As discussed in Section 2.5.2 and in Doswell et al. 1996, the necessary ingredients for strong, flash flood-inducing convective storms are: 1) moisture, 2) atmospheric instability, and 3) a lifting mechanism. By inspection of Figure 6.2d, one sees that precipitable water values, a column integration of the atmospheric moisture content, were very high over much of the southern plains, with widespread values over 35 mm and several areas with over 50 mm (2 in.) of precipitable water at 0000Z on 24 May 2015. This can also be seen in the Oklahoma City, OK and Corpus Christi, TX soundings from the same time (Figures 6.3b and 6.3c), where 44 and 46 mm of precipitable water was observed, respectively. This suggests that ample moisture was present- certainly sufficient to cause flooding concerns. Furthermore, low level southerly flow (Figure 6.2c) is helping to continue the advection of warm, moist air into the southern plains region. Up in the central Rockies, precipitable water readings were much lower- the Riverton, WY sounding (Figure 6.3a) at 0000Z on 24 May indicates only approximately 16 mm

of precipitable water- this is still much moister than climatology for this location. This is illustrated in Figure 6.4, where it is seen that the approximately 0.63" precipitable water reading was over the 90th percentile for that location and date. Returning to Figures 6.2a and 6.2b, it is evident that the upper level pattern is rather disturbed, with strong troughing over the western CONUS, and an amplified ridge to the east. Additionally, there is a pronounced shortwave embedded in the upper level pattern; this feature is most evident over south central Texas, with an associated 500 mb vorticity maximum seen in Figure 6.2a. As is most evident in Figure 6.2b, the southern plains region also falls broadly in the right entrance region of a reasonably straight upper level jet streak. The upper level divergence and positive vorticity advection associated with the right entrance region of this jet streak, in addition to that being advected by the incoming shortwave trough, support synoptic scale forcing for ascent in this region. Inspecting the OUN and CRP soundings in Figures 6.3b and 6.3c, one sees modest, but certainly sufficient, convective available potential energy (CAPE) values in the vicinity of 1500 J/kg. Thus, all the necessary synoptic-scale ingredients for strong, flood-producing storms are present over the southern plains at this time. The mechanisms are somewhat different in the central Rockies of WY, with the topography serving as an important source of lift in this region, but the ingredients are in place here as well.

These ingredients came together to produce intense rainfall and catastrophic flood impacts over these regions between 1200Z on 23 May 2015 and 1200Z the subsequent day (and numerous notable events followed in the same region later in the month). Stage IV Precipitation Analysis over this period is indicated in Figure 6.5. Three areas of particularly heavy rainfall were observed: 1) over south central Texas in the Austin/San Antonio vicinity, where over 200 mm (8"+) was over observed over a relatively large area; 2) over much of Oklahoma, with widespread accumulations over 80 mm and two small areas over the western half of the state receiving in access of 200 mm; and 3) in central Wyoming, where several storms produced in excess of 75 mm in this region- a highly unusual occurrence in this part of

the country. It is should also be noted that these extreme rainfall regions span all three observational return period datasets; Oklahoma thresholds are determined in Atlas 14, Texas thresholds are specified by the TP-40 data, and Wyoming thresholds are taken from Atlas 2. This may result in some discrepancies in precipitation frequency analysis, since, for example, Atlas 2 thresholds tend to be lower than neighboring Atlas 14 thresholds (see Figures 3.1, 3.2). An example of the dramatic impacts the Texas rainfall had on area rivers is depicted in Figure 6.6; a gauge on the Blanco River rose from 5 feet to over 40 feet in approximately four hours, and the associated impacts were devastating to the region, with numerous fatalities recorded. Similar flash flood impacts were witnessed in Oklahoma, and though less publicized- likely in part due to the smaller population in the region- major flooding impacts were observed in Wyoming as well, prompting the declaration of a federal disaster for state flooding beginning on this date.

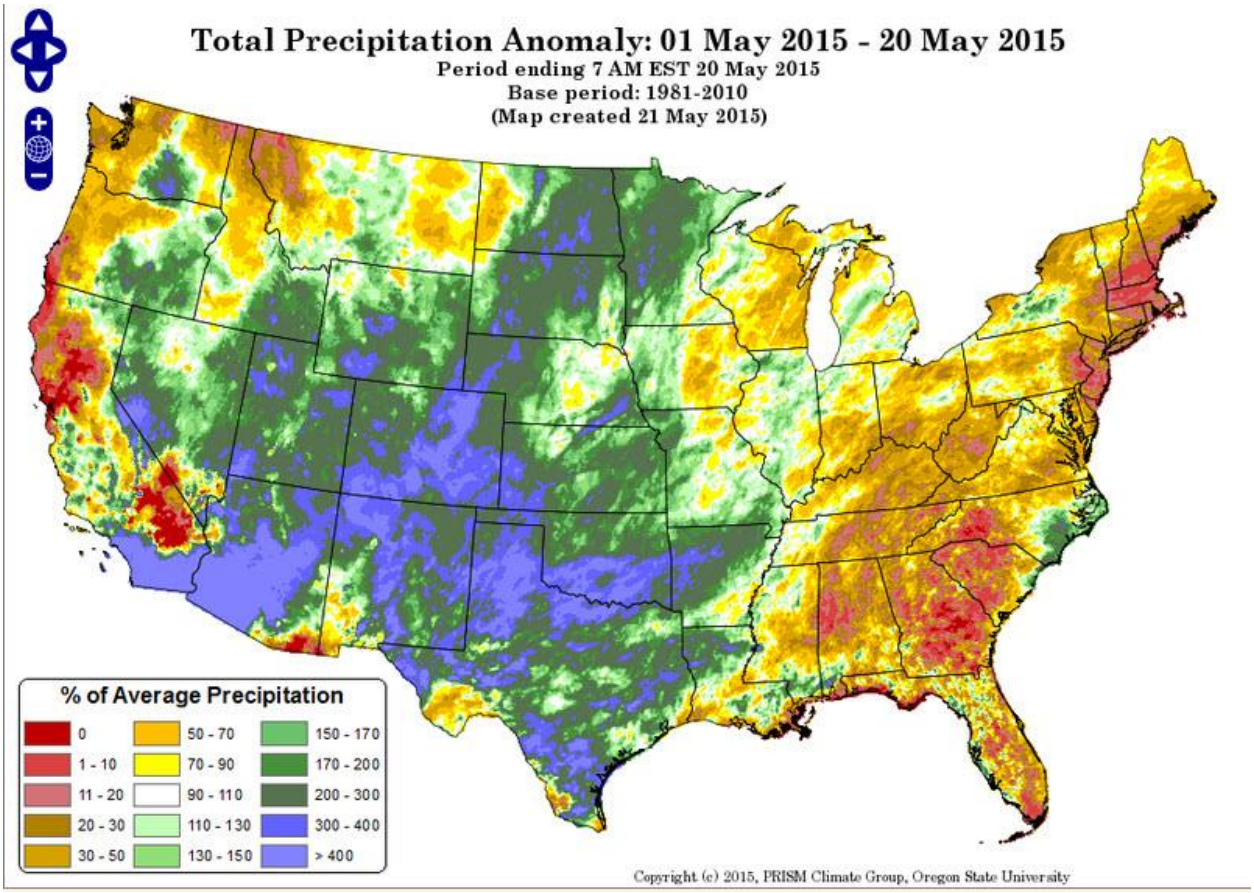

Figure 6.1: Antecedent precipitation expressed in proportion to local climatology from 01 May 2015-20 May 2015, approximately covering the three-week period prior to the 23-24 May 2015 extreme precipitation events. Taken from Beau Dodson's Weather Talk Blog: http://talk.weathertalk.com/may-23-2015-a-beautiful-saturday/

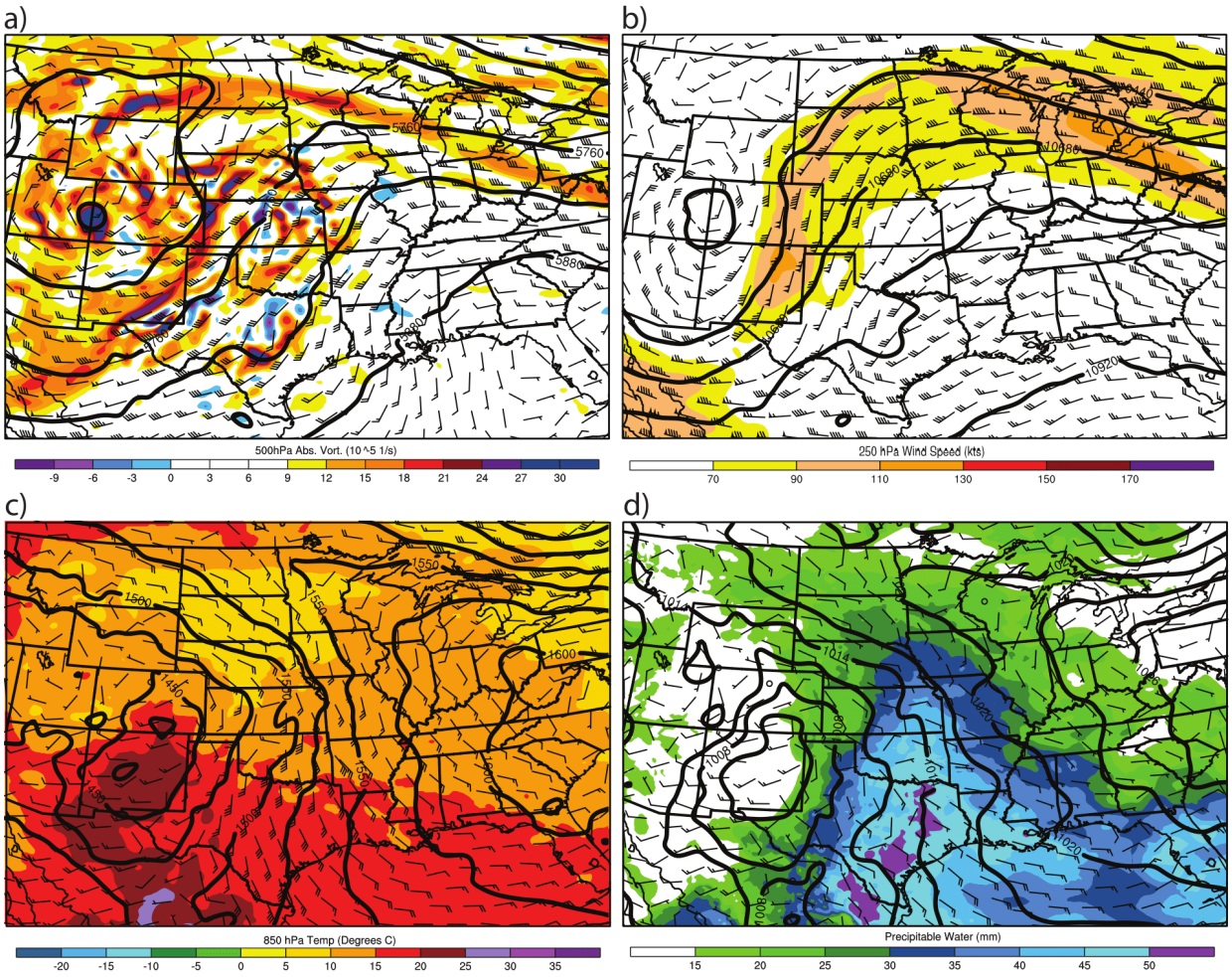

Figure 6.2: Synoptic-scale conditions over the central United States at 0000 UTC on 24 May 2015. Panel (a) contours 500 hPa heights with colored vorticity and plotted wind barbs at the same pressure level; panel (b) mirrors panel (a) for the 250 hPa level except that colors reflect 250 hPa isotachs; panel (c) reflects the 850 hPa conditions with colored isotherms; and panel (d) contours mean sea level pressure and colors precipitable water, with wind barbs reflecting 10-meter winds. All fields based on Rapid Refresh (RAP) analysis.

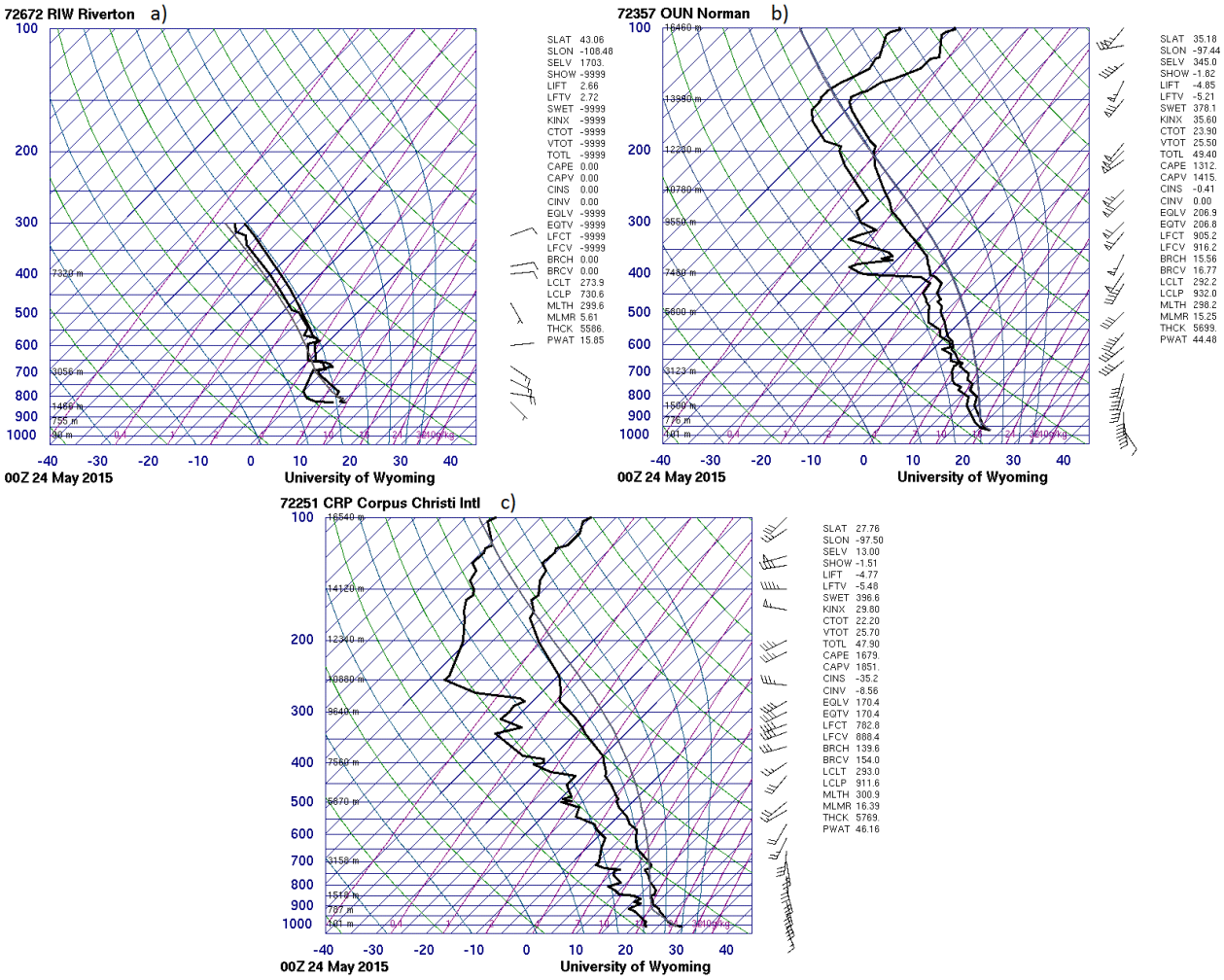

Figure 6.3: Observed soundings plotted on skew-T diagrams taken at several sites at 0000 UTC on 24 May 2015. Panel (a) reflects the Riverton, WY (RIW) sounding, panel (b) plots the Norman, OK (OUN) sounding, and panel (c) illustrates the Corpus Christi, TX (CRP) sounding. Images taken from the University of Wyoming sounding database: http://weather.uwyo.edu/upperair/sounding.html.

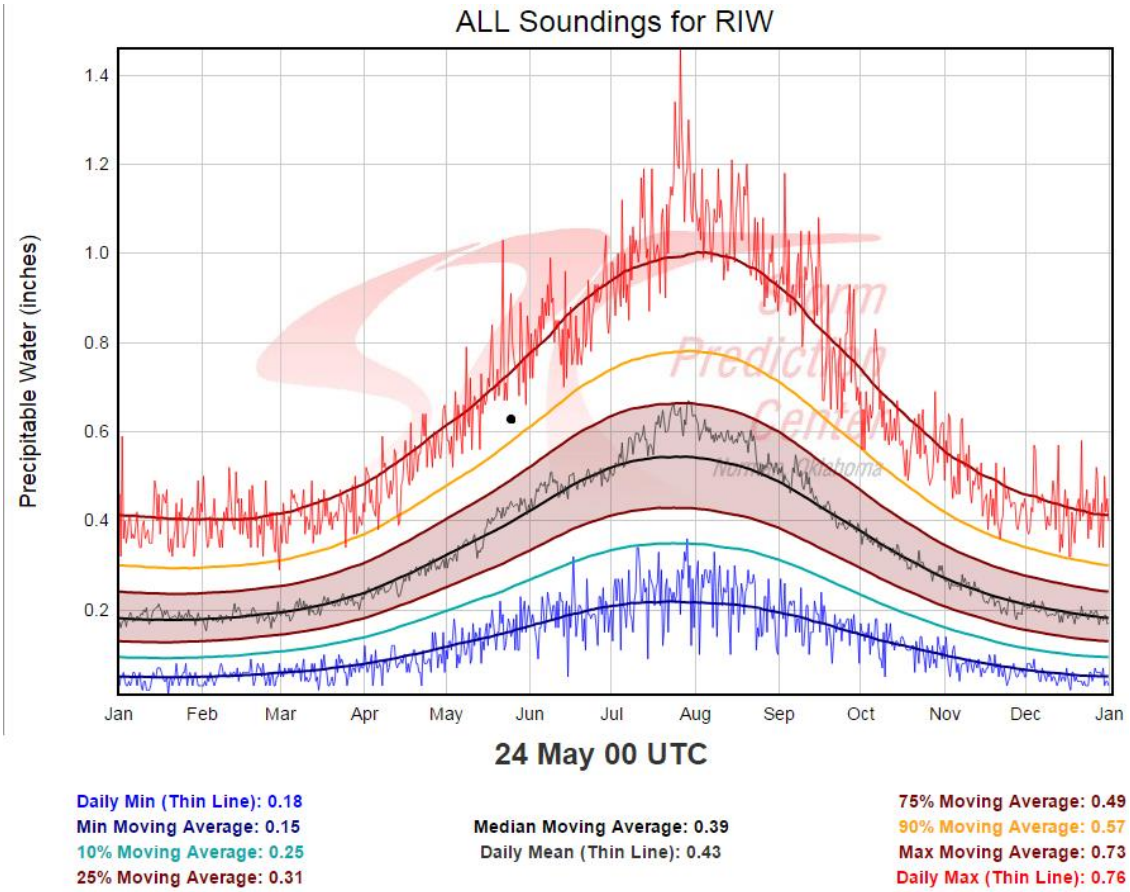

Figure 6.4: Climatological precipitable water values for Riverton, WY, shown here for contextual reference. Black dot indicates the approximate location on the diagram of the corresponding sounding in Figure 6.3a. Lines depict climatological percentiles in the color scheme noted at the bottom of the figure. Image taken from Storm Prediction Center's sounding climatology webpage: http://www.spc.noaa.gov/exper/soundingclimo/.

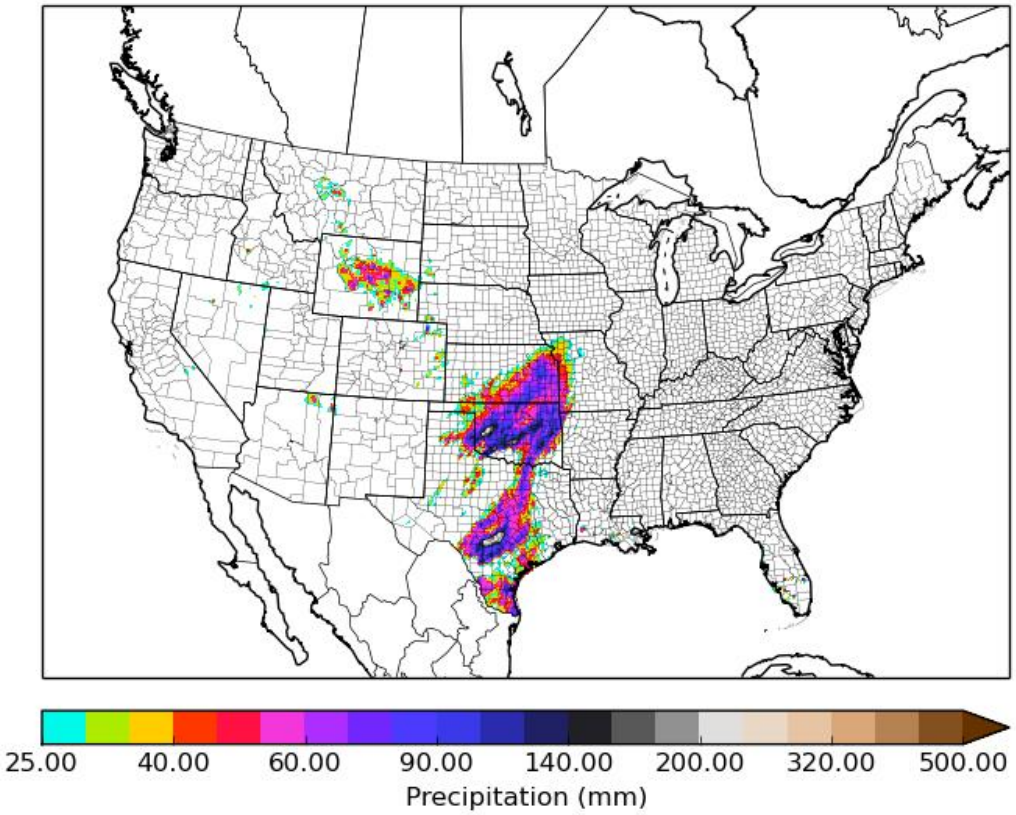

Figure 6.5: Stage IV precipitation analysis for the 24-hour period ending at 1200 UTC 24 May 2015.

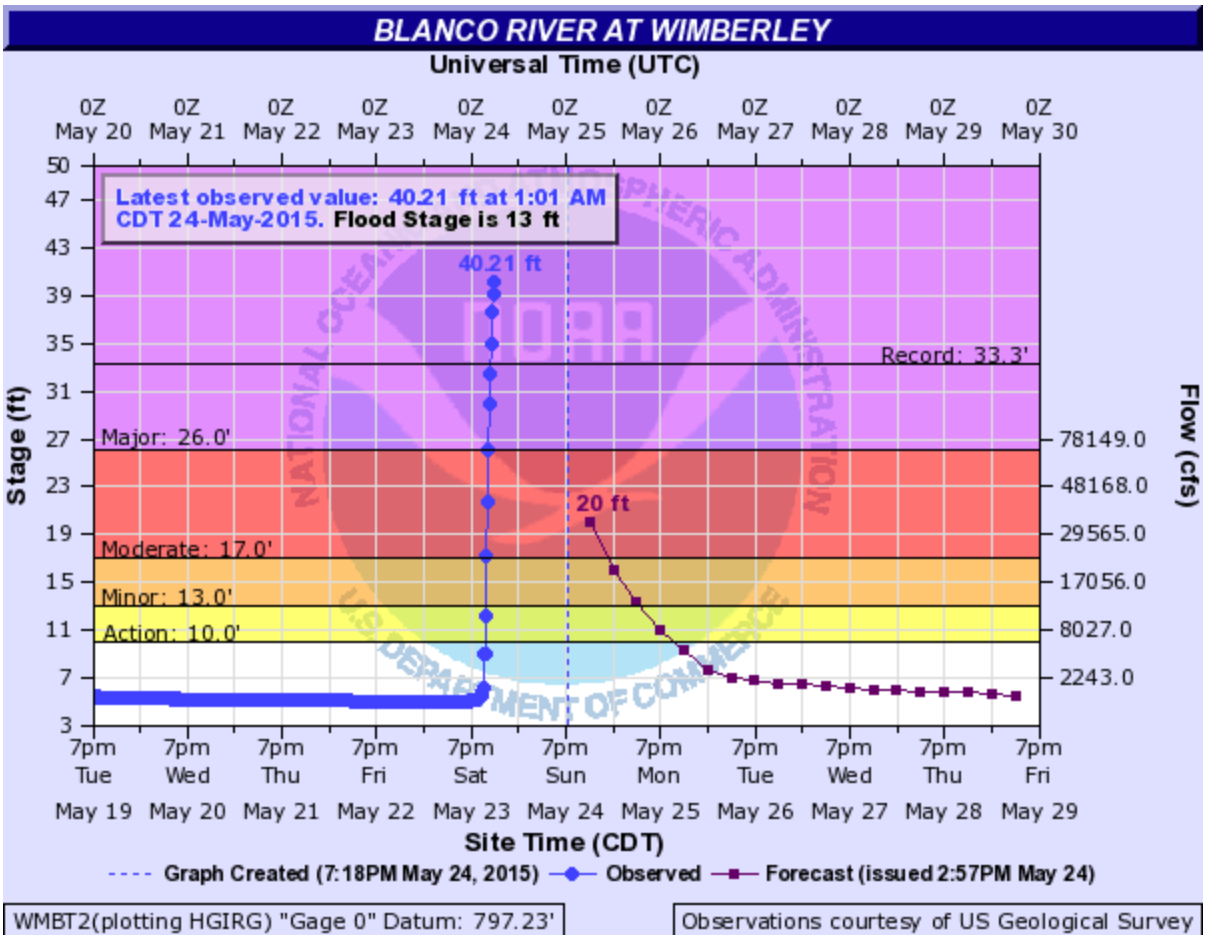

Figure 6.6: Hydrograph of the Blanco River at Wimberly, TX from 19 May 2015 to 29 May 2015, indicating the peak river level near 0500 UTC (0000 CDT) on 24 May 2015. Plotted values after this time reflect forecast values at that time, and not observed values. Image taken from: http://www.srh.noaa.gov/ewx/?n=memorial_weekend_floods_2015.

6.2 Member Forecasts

As discussed in section 6.1, the synoptic-scale ingredients were certainly in place for locally extreme rainfall in the places that observed it on this day. This suggests high predictability of the general notion of heavy precipitation in the region. However, several mesoscale features helped pinpoint the exact locations of heaviest rainfall, which had an enormous effect on the local impacts. Given this, how did the numerical models perform with their forecasts for extreme precipitation on this day? This question will be explored in this section.

Figure 6.7 depicts the 1200-1200Z 23-24 May 2015 precipitation accumulation forecast for the eleven members of GEFS/R based on the 0000Z 23 May initialization. The eleven members are shown in panels (a)-(k), and the Stage IV verification is repeated in Figure 6.7l for convenience. Perhaps due to the coarseness of the model (see section 2.2), in addition to the well-documented tendency for underdispersion, the GEFS/R members depict rather similar solutions, all differing from each other much less than they differ from the verifying solution. All of the runs do an acceptable job at forecasting relatively heavy precipitation in the general vicinity of where it was observed. A couple of members, and in particular member 2, depict locally heavy precipitation over central Wyoming. However, most members put the heavy precipitation in that region in SW South Dakota, to the northeast of where it was actually observed. Though GEFS/R provides some indication of a threat for heavy precipitation in Wyoming, a forecaster looking at only this guidance would rightly place more emphasis and attention to South Dakota. With regards to the precipitation over the southern plains, all the members correctly show an axis of heavy precipitation extending into SE Kansas. Most members do have two regions of enhanced precipitation embedded along this axis, one generally in northern Texas (to S OK) and the other somewhere in Oklahoma (to SE KS). These are displaced well to the north, and somewhat to the east of the observed maxima, though the eastern displacement does not appear to be associated with displacement error in the precipitation axis. The axis depicted in GEFS/R is apparent in the observations, but is quite weak with minimal surrounding precipitation in northern Texas and far southern Oklahoma. This distinction is seen to an extent in members 3 and 4 (Panels d and e), but is not nearly as apparent as in the observations. None of the members accurately predict the intensity of the precipitation west of the main axis in Oklahoma, which ended up being the location of heaviest precipitation in the state. Member 2 is the only one to really show this feature at all, but has the precipitation much weaker outside the primary N/S axis. Figure 6.8 presents the same information as in Figure 6.7, but framed in the context of return periods using the Atlas 14 thresholds. This presentation really highlights the

displacement errors with the Oklahoma and especially with the Texas precipitation maxima. Additionally, all the members are quantitatively much too weak, with members forecasting at most 5-year RP exceedances; most members only forecast 1- or 2-year events. This framework also highlights the importance of the Wyoming precipitation much more than the traditional QPF plots in Figure 6.7. Inspecting Figure 6.7 from a national perspective, one's attention is drawn much more to the precipitation over the southern plains than what is happening in the north. But the impacts of the precipitation were substantial there, and this is evident looking at the verification in Figure 6.8l, where the Wyoming precipitation and TX/OK events are given somewhat similar emphasis. This is also seen in the GEFS/R member forecasts, where this is perhaps even more emphasized than the southern plains precipitation, with 5-year events forecast in several members. Again the severity of the precipitation was underforecast in the GEFS/R members, but Figure 6.8 highlights the value of considering heavy precipitation from the RP framework.

Turning to CAMs, a sampling of high-resolution model guidance initialized at the same time as GEFS/R is shown in Figure 6.9. The NSSL-WRF, shown in Figure 9a, gives a fairly realistic looking broad depiction of the southern plains precipitation, with two areas of heavy precipitation, one over south-central Texas, and the other over Oklahoma, with scattered lighter precipitation in-between. The NSSL-WRF also shows some heavy rain-producing convective cells over the mountains of Wyoming, even showing some of the lesser cells that were observed in NE AZ and NE NV. The NSSL-WRF seems to do quite well with both location and magnitude with the Texas precipitation. In Oklahoma, like the GEFS/R, it fails to capture the very significant precipitation seen over the western part of the state, instead showing no significant precipitation at all in that area. The QPF magnitudes in the WY mountains look reasonable as well, accurately highlighting the threat in that area. The NAM-NEST (Figure 6.9b) exemplifies its typical behavior described in Chapter 3; very high precipitation maxima are seen in all the regions in which it is observed. Over SE WY, a large region of locally very heavy precipitation is

observed. The corresponding RP exceedance forecast plots are shown in Figure 6.10; the WY precipitation in the NAM-NEST is especially evident in Figure 6.10b, with a very large area of 50+ year RP exceedances. Unlike the NSSL-WRF, the NAM-NEST also forecasts, albeit to a lesser extent than in Wyoming, 50+ year exceedances in Idaho, Utah, and Nevada, none of which are seen to verify. Like the NSSL-WRF and especially GEFS/R, the NAM-NEST produces an axis of very heavy precipitation over the southern plains. In this case, the model accurately captures the intensity of the areas of heaviest rainfall, but the spatial structure of the precipitation is largely off, failing to forecast the heavy precipitation west of the main axis, having a westward displacement and clockwise rotation to the main axis relative to what is observed, and failing to capture the lack of precipitation in northern Texas, instead forecasting some high RP exceedance events in this area. Inspecting the HIRESW solutions (Figures 6.9c and 6.9d), the HIRES-ARW run does quite well in the north, doing an excellent job of pinpointing the locations of heaviest precipitation, and only underforecasting the precipitation by a bit. To its further credit, it correctly forecasts a strong cell in NE CO which produced rather heavy rainfall in the area- the only model to really forecast this event. In the south, however, it performs considerably worse. It does show heavy precipitation in Texas, but it is displaced too far to the south. It also predicts the local minimum in precipitation over northern Texas, but the precipitation over Oklahoma and vicinity is of the wrong character and far too weak. A forecaster looking at just this piece of guidance would not be likely to appropriately perceive the extreme rainfall threat in Oklahoma. They would at least have some notion though; Figure 6.10c shows some scattering of 10+ year events in OK. The main precipitation is also displaced too far to the NE, again along the main precipitation axis seen in all the model guidance. The HIRES-NMM (9d) performed very poorly overall. Some evidence of the Wyoming precipitation threat appears in the model solution, but it is underplayed both in magnitude and spatial extent. The precipitation over Oklahoma and Texas is so woefully underrepresented that it provides

almost no indication of the severity of the rainfall threat; this is especially apparent in the RP context shown in Figure 6.10d.

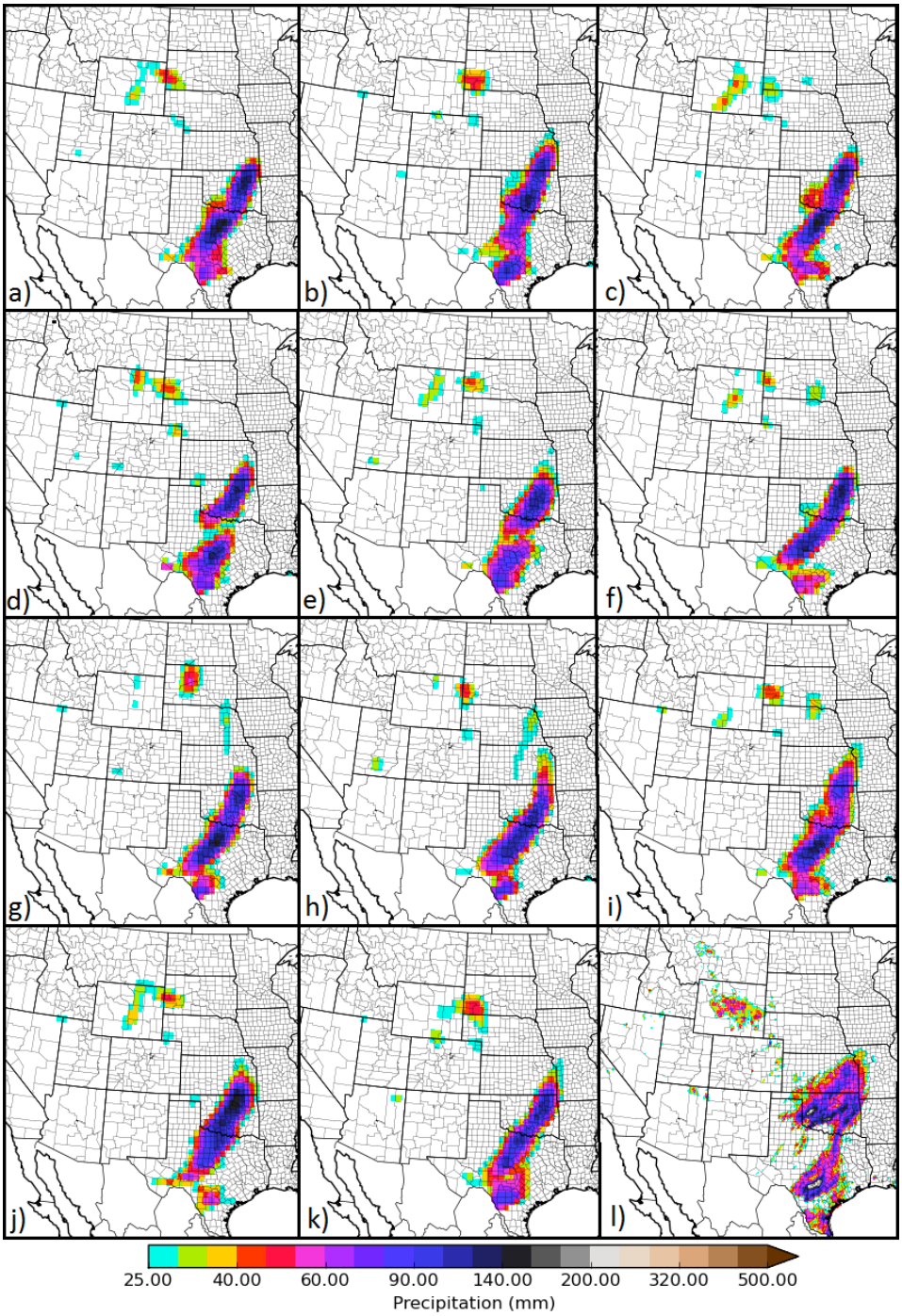

Figure 6.7: All GEFS/R precipitation accumulation forecast for the 24-hour period ending 1200 UTC on May 24 2015 based on the 0000 UTC 23 May 2015 initialization appear in panels (a)-(k) for the control and subsequently members 1 to 10, in sequence. Panel (l) reproduces the Stage IV precipitation analysis over the same period shown in Figure 6.5, and is reproduced here for convenience.

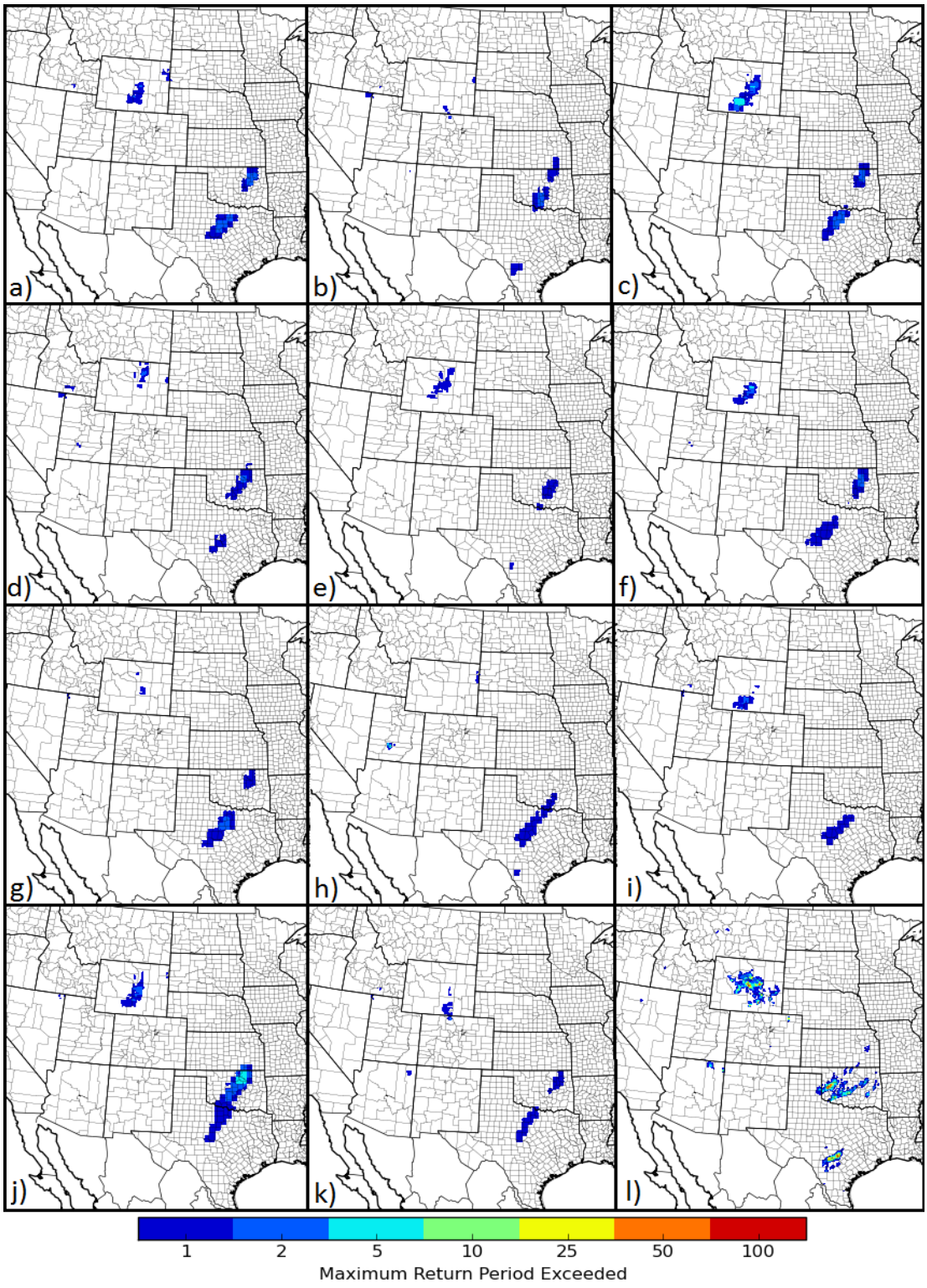

Figure 6.8: Same as Figure 6.7, but precipitation values plotted with respect to maximum return period (among 1-year, 2-year, 5-year, 10-year, 25-year, 50-year, 100-year) exceedance forecast or observed.

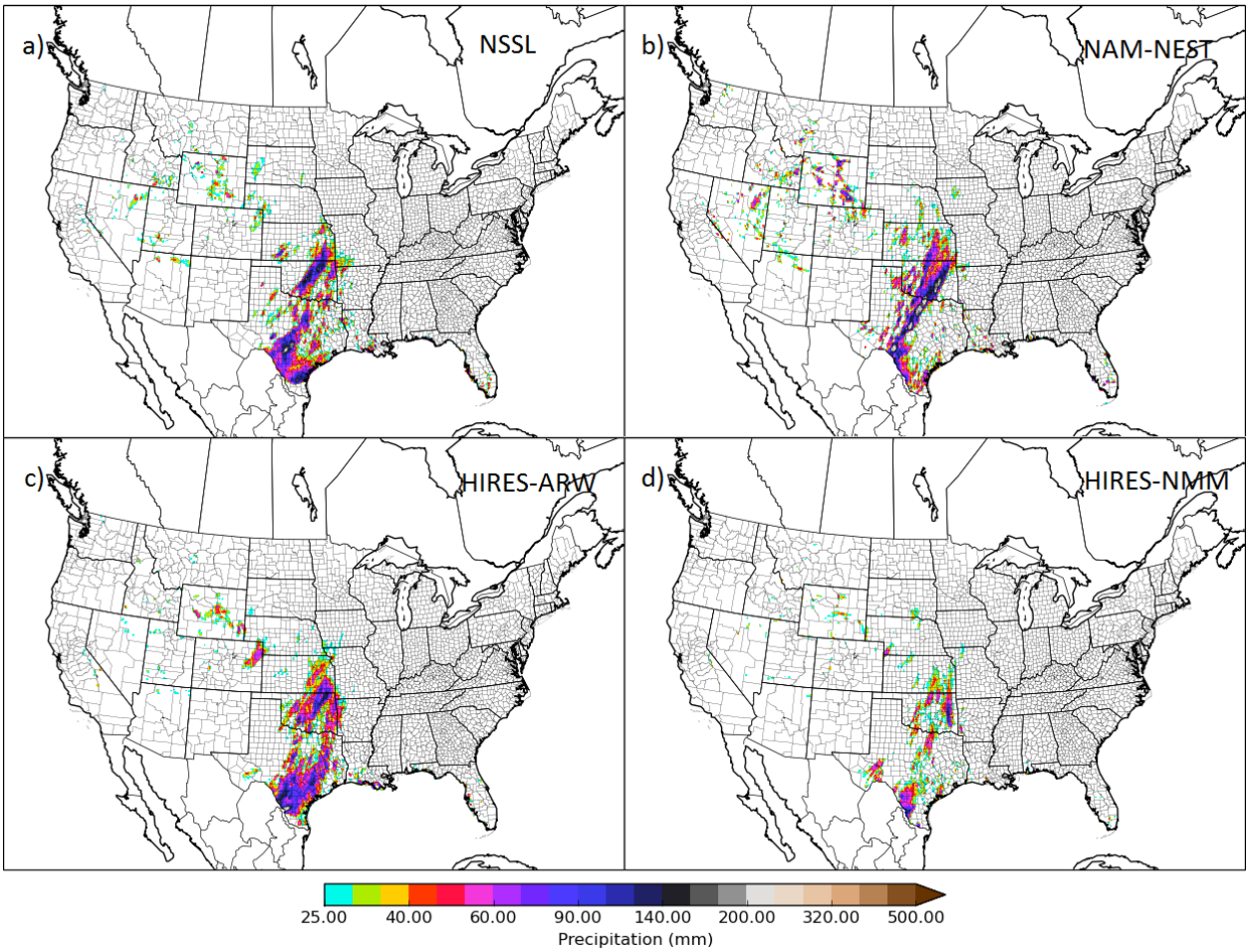

Figure 6.9: Precipitation forecasts for the 24-hour period ending 1200 UTC on 24 May 2015 based on each model's 000 UTC 23 May 2015 initialization. Panel (a) depicts the NSSL-WRF forecast, (b) corresponds to the NAM-NEST, (c) do the HIRESW-ARW, and (d) to the HIRESW-NMM.

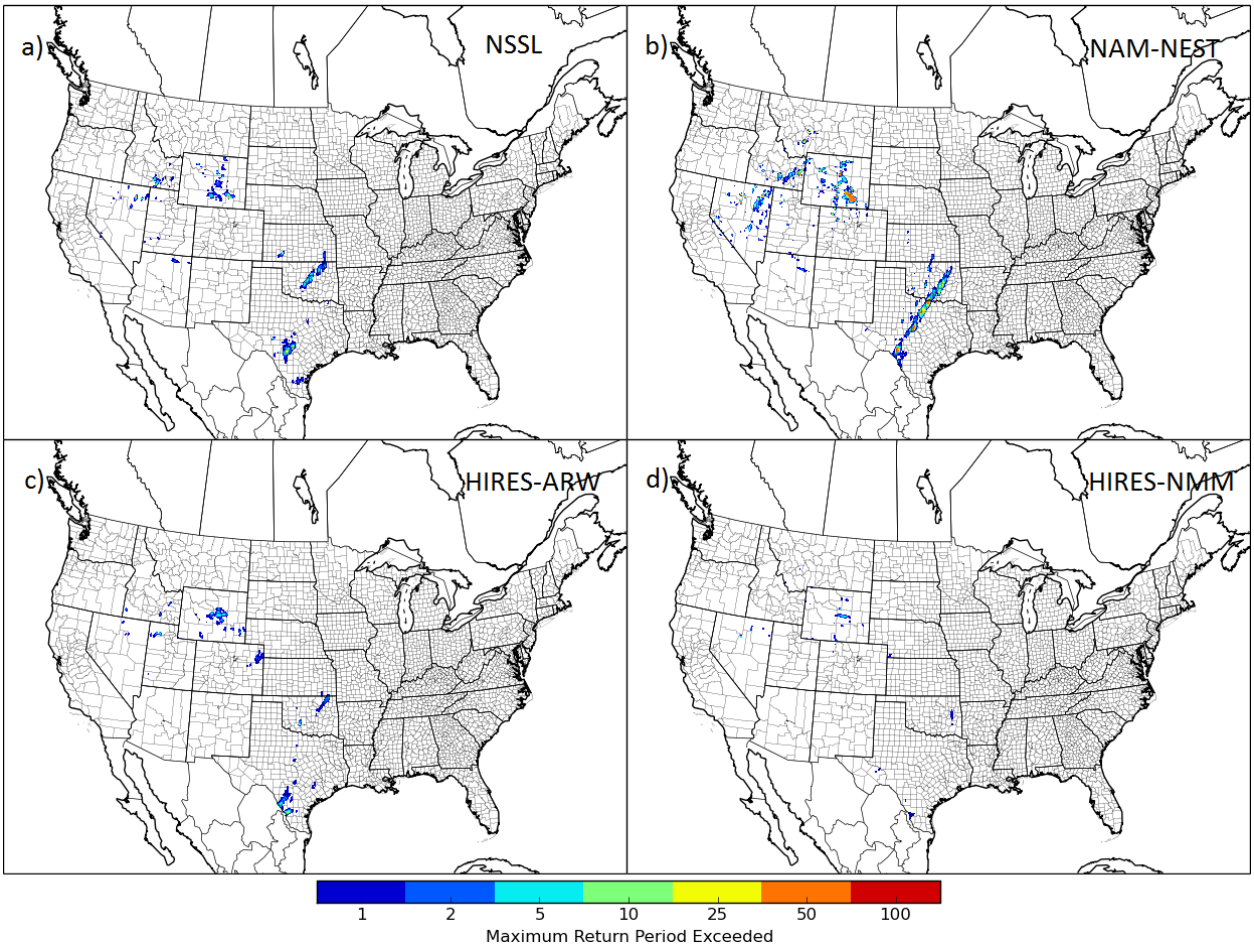

Figure 6.10: Same as Figure 6.9, except precipitation values plotted with respect to maximum return period (among 1-year, 2-year, 5-year, 10-year, 25-year, 50-year, 100-year) exceedance forecasted.

6.3 Using the Forecast System: Probabilistic Forecasts

Probabilistic forecasts for this case using some of the basic NIVA methods outlined in Chapter 5.1.1 are depicted in Figure 6.11. All FPs in that figure are derived from an ensemble consisting of all the member forecasts shown in Section 6.2: the full GEFS/R (11 members), the NSSL-WRF, the NAM-NEST, HIRES-ARW, and HIRES-NMM. PDV, as shown in Figure 6.11a, is about what one would expect from a simple synthesis of 6.8 and 6.10. Very noisy FPs with three areas of particularly high (> 0.5) probability of exceeding the 2-year 24-hour return period thresholds are seen: one in north-central Texas, one in northeast Oklahoma, and one in central Wyoming. Comparing with the verification in Figure 6.8l, we see that the aggregate of forecasts did quite well with the Wyoming system, even having a probability

bullseye in the approximate area that the most extreme precipitation occurred. FPs to the south were not nearly as skillful, with the emphasized region in Oklahoma being one of the areas of the state with the least coverage of 2-year RPT exceedances, and the region which had the highest coverage has very low PDV FPs. Furthermore, no 2-year RPT exceedances were observed in the high FP zone in north-central Texas, and where locally extreme rainfall was observed well to the south, FPs were very low. Admittedly there was some local enhancement in FPs to the immediate east of where extreme precipitation was observed in south-central Texas, with FPs as high as 0.25. Ultimately, the raw guidance, as synthesized with the PDV FPs, performed so-so at predicting extreme rainfall in this case, but there was certainly considerable room for improvement, particularly with the southern systems. PUR FPs (Figure 6.11c), at first glance, appears very similar to PDV, and rightly so; PUR can't change local FPs by more than $\frac{1}{n}$, where n is the ensemble size, from PDV FPs. But upon closer inspection (e.g. 6.11d), there are some minor differences. Principally, PUR acts to raise FPs slightly in the regions where precipitation is observed in some members, but no members' QPF exceeds the 2-year RPT. This is most evident in the mountainous west, and on the peripheries of model heavy precipitation features, where the model QPF is near, but just under the critical event threshold, thus leaving PDV to assign no weight, but PUR to assign a large fraction of the total possible FP contribution from the relevant ensemble member. One can see that PUR correctly enhanced FPs over southern Texas and over west-central Oklahoma, and decreased the magnitude of the FP bullseye over north-central Texas.

A variety of neighborhood based probability methods are compared for this case in Figure 6.12. As in Figure 6.11 going from PDV to NDV20, increasing the neighborhood radius further in NDV50 (6.12b) greatly smooths out the point FPs, with no FP exceeding 0.3 using NDV50, but probabilities in excess of 1% covering almost all of Oklahoma, Wyoming, and north and central Texas. It is evident by comparison of Figures 6.11a, 6.11e/6.12a, and 6.12b that, unsurprisingly, increasing neighborhood radius decreases forecast sharpness. It is less clear, just from inspection, what choice of neighborhood

radius maximizes forecast skill; that requires bulk analysis as assessed in Chapter 5. In this case, the 50 grid box neighborhood radius clearly results in a worse forecast, with the highest probabilities seen in the local observation minimum in north Texas, and lower probabilities over Oklahoma and south Texas. DNUR, with a neighborhood radius of 20, appears to operate, at least in this case, similarly to increasing the neighborhood radius: FPs are increased at the outskirts of regions of locally high FPs, while FPs are slightly decreased in the center. This makes sense- in DV, a precipitous drop in FP occurs crossing from a point where an ensemble member exceeds the critical threshold to one where it does not. In UR, there is no such discontinuity, and thus FPs are accordingly higher on the non-exceedance side and lower on the exceedance side. The same phenomenon is true for ENUR20, and the effect is actually amplified, likely due to the deceased number of points considered; in ENUR applied here, the spacing *s* was 10 grid boxes versus 5 for DNUR, and no post-smoothing was applied to either FP field. It should also be noted that, in areas of low probabilities, generally where $FP < \frac{1}{n}$, where n is the ensemble size, both DNUR and ENUR act to enhance FPs slightly. This is evident, for example, near the NV/ID/UT intersection.

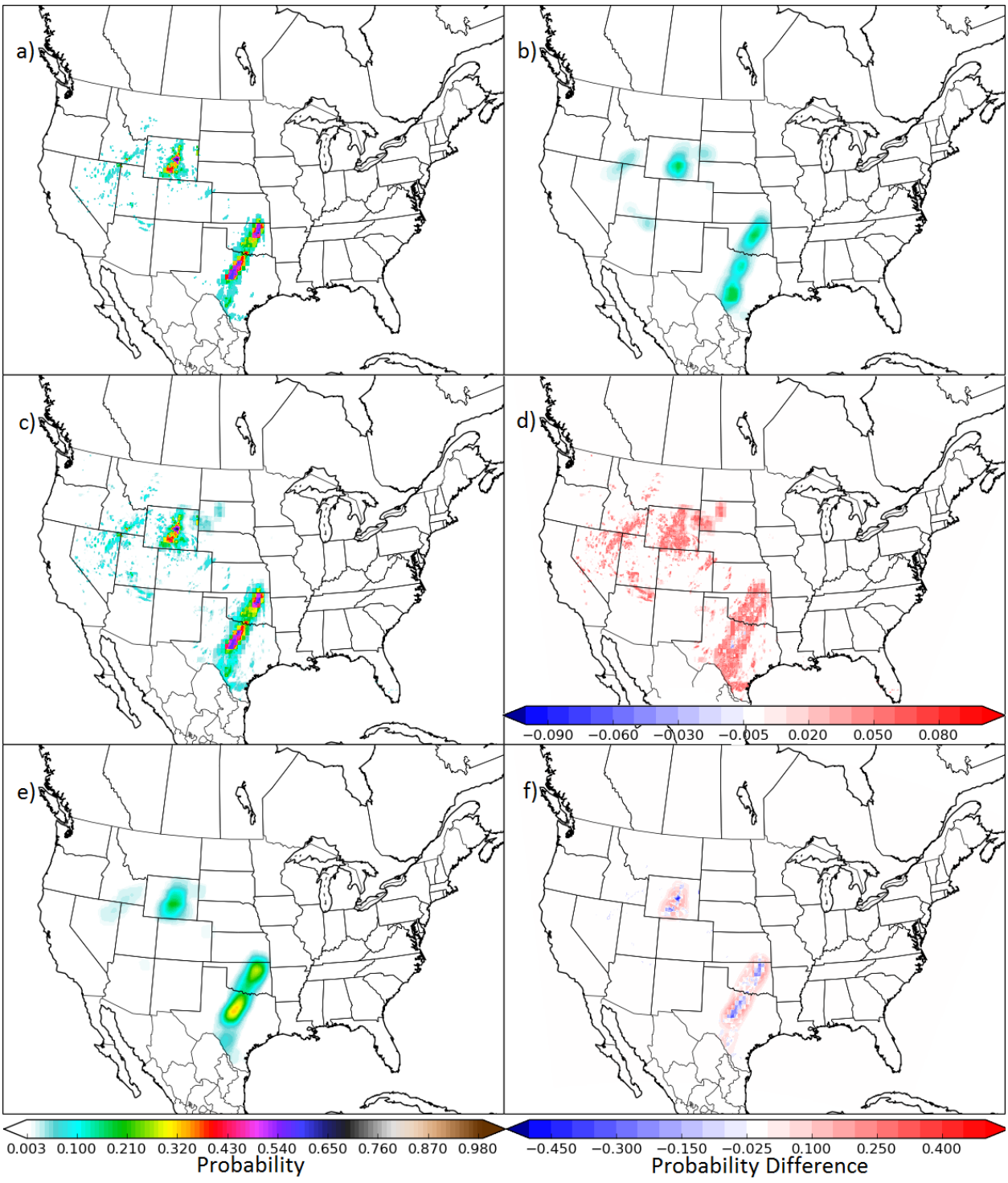

Figure 6.11: Comparison of several basic NIVA methods for generating point forecast probabilities. Probabilities correspond to the event of exceeding 2-year, 24-hour return period thresholds via the PDV method for the 12-36 hour forecasts of the 0000 UTC 23 May 2015 forecasts initialization. Panel (a) applies PDV; panel (b) applies logistic regression to a 3-member ensemble as discussed in Chapter 5; panels (c) and (d) plot PUR FPs and the difference from PDV, respectively; and panels (e) and (f) depict NDV FPs with a neighborhood radius of 20 grid boxes (NDV20), and its departure from PDV FPs. All methods, unless specified otherwise, are applied to an ensemble consisting of full GEFS/R, NSSL-WRF, NAM-NEST, and the HIRESWs, for a total of 15 ensemble members.

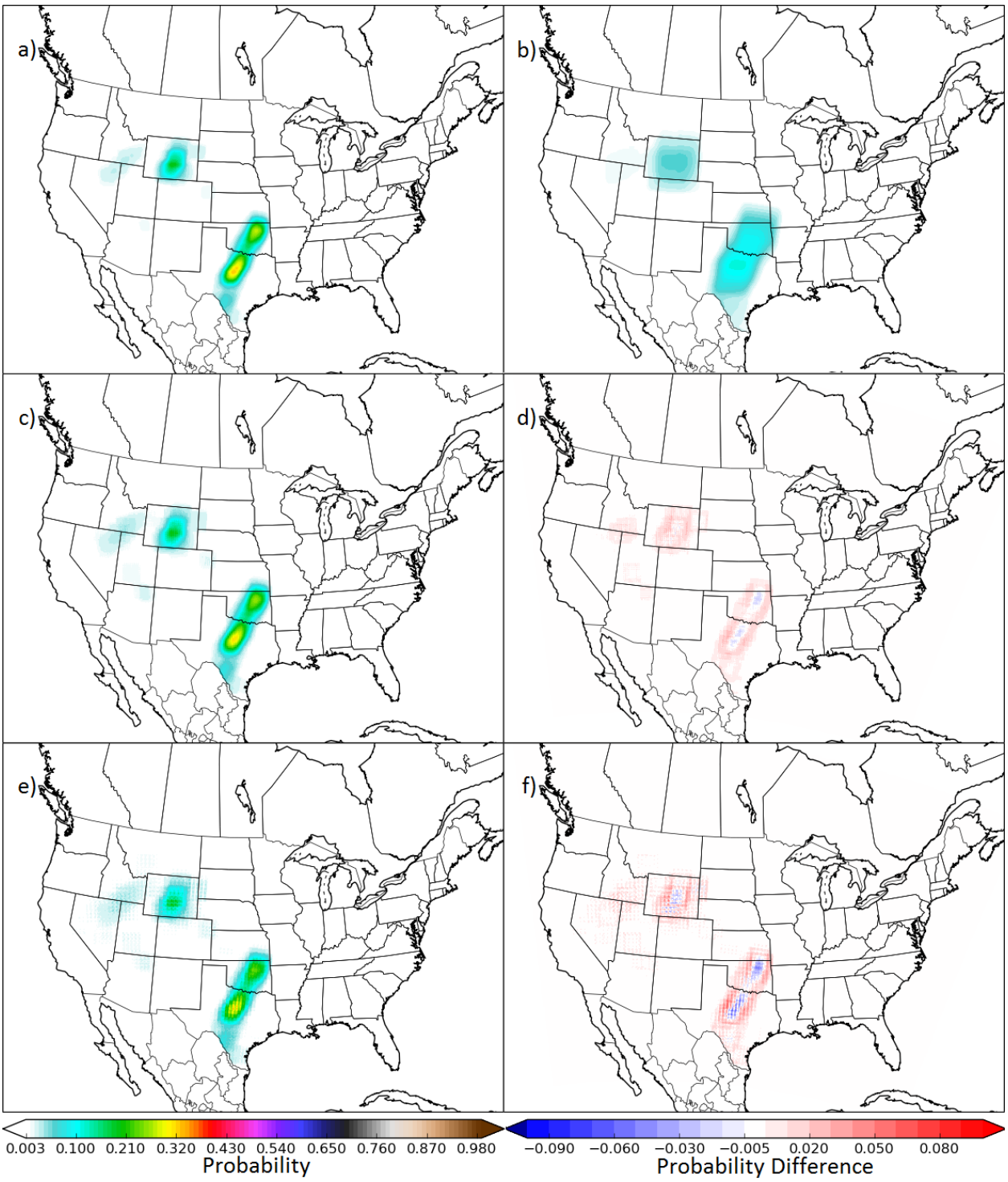

Figure 6.12: Comparison of several neighborhood methods for generating point forecast probabilities. Probabilities correspond to the event of exceeding 2-year, 24-hour return period thresholds via the PDV method for the 12-36 hour forecasts of the 0000 UTC 23 May 2015 forecasts initialization. Panel (a) applies NDV with a neighborhood radius of 20 grid boxes (NDV20), and panel (b) applies NDV with a neighborhood radius of 50 grid boxes (NDV50). Panel (c) plots FPs from DNUR with a 20-box radius (DNUR20), and (d) depicts the probability difference from NDV with the same radius. Panels (e) and (f) plot ENUR20 FPs and the corresponding probability difference from NDV20, respectively. All methods applied to ensemble consisting of full GEFS/R, NSSL-WRF, NAM-NEST, and the HIRESWs, for a total of 15 ensemble members.

7 Summary, Conclusions, and Future Work

A forecast system for applying NWP model guidance from a plethora of sources and on different scales towards the important forecast problem of locally extreme rainfall framed in the context of probabilistic return period exceedance forecasting has been presented. Numerous models, both convection-parameterizing and convection-allowing -have been individually evaluated to discern historical model skill in forecasting extreme rainfall, both in regional and bulk senses, as well as to determine model bias characteristics for extreme precipitation thresholds. Coarser models largely failed to adequately simulate small-scale convective events, while often adequately handling extra-tropical cyclones impacting the US west coast during the cool season, and tropical cyclone induced events impacting the eastern and southeastern US. High-resolution models produced a spectrum of skill and bias characteristics; most tended to overforecast extreme events relative to what was observed in Stage IV analysis, but some, notably the NAM-NEST, dramatically overforecasted events all across CONUS, while others were much more tempered. Overall, high resolution models tended to perform slightly better than the coarser models, though the skill differential was perhaps less than one might expect.

The two models with the longest data record- the GEFS/R and NSSL-WRF- had model precipitation climatologies fitted based on their data record, and these climatologies were used to create model-specific RP precipitation thresholds which correct for the model biases relative to the observationally derived thresholds. A number of right-skewed distributions were employed to assess the model bias characteristics with respect to extreme precipitation. GEFS/R distribution fits almost unequivocally produced RP thresholds much lower than those seen in observations as a result of failing to resolve many of the extreme precipitation events which have occurred. NSSL-WRF derived thresholds were highly sensitive to the choice of right-skewed distribution and exact methodology employed to generate

the fits, but were generally much more in line with the observationally-derived thresholds. A variety of techniques were then assessed in order to diagnose the deterministic RPT thresholds, among the available options, which maximized model predictive skill. Using a combination of observational and model-derived thresholds was found to improve model skill at predicting locally extreme rainfall at all return periods assessed, and the improvements were statistically significant for the medium RPs.

Lastly, preliminary work at using an ensemble of models to generate probabilistic locally extreme rainfall forecasts by means of point exceedance probabilities was presented. Raw methods using the quantitative precipitation values rather than a simple binary relationship between model QPFs and the event threshold, employing model neighborhoods to inform point probabilities, and use of ensemble weighting to alter forecast probabilities from the traditional fraction of members exceeding the event threshold at a point method were explored. Forecasts were assessed in terms of skill, reliability, and value. Though each class of metric resulted in different conclusions about the best ensemble configuration, all metrics agreed that substantial forecast improvement over the simple point democratic voting approach could be achieved easily and efficiently.

Extensive ongoing and future work is anticipated in association with this project. Initial methodology presented here will be refined. Many of the more involved methods discussed in Chapter 2 such as more involved machine learning applications for FP generation will be examined in much more depth than the very limited treatment given here. More sophisticated and theoretically sound methods for distribution fitting model data and choosing appropriate model-specific RP thresholds that more intelligently make use of known or derivable spatial relationships between precipitation distributions will be examined in further depth. This is especially important for operational applications since so many existing models have notably shorter data record lengths than the models examined in detail here. The initial inquiries examined here will also be considerably extended; it is desired that the complete

forecast system, incorporating the entire methodology framework discussed herein, be applied to: 1) both 6- and 24-hour AIs; 2) 1-, 2-, 5-, 10-, 25-, 50-, and 100-year RPs; 3) 0000 and 1200 UTC model initializations, perhaps incorporating multiple model cycles as ensemble members to yield a time-lagged ensemble; and 4) lead times beyond those examined here, perhaps out to forecast hour 60 for 6-hour accumulations (days 1 and 2) and 84 for 24-hour accumulations (days 1, 2, and 3). It is hoped to implement these models and apply them in a publicly accessible, real-time setting as a tool for forecasters, decision makers, and the interested public. This may also provide an avenue to explore forecast communication issues and nuances unique to extreme events, a highly important interface between atmospheric science, statistics, and social sciences which has historically been underexplored.

As discussed in Chapters 1 and 2, the performance gap between human and automated forecast skill has been decreasing over recent decades and years. It is frequently noted that the instances where forecasters are able to consistently improve over the automated guidance occurs during rare and extreme events. Forecasters are able to recognize rare event potential and unusual meteorological situations in general and appropriately adjust their forecasting framework and mindset to produce a better forecast. Most existing operational automated guidance cannot do this; most guidance is trained to perform well on most (typical) days, and with no training examples in the model residing near the potential verification in feature space, most models do not extrapolate well into extreme scenarios, resulting in poor predictions. The exacerbation of dynamical model biases in extreme scenarios makes dynamical guidance an unreliable source for automated predictions as well. It is my belief that automated guidance could be significantly enhanced by first quantifying the anticipated rarity of the meteorological situation, likely on a regional scale, and then using this to apply appropriate automated forecast methodology to generate a forecast. This could be used in the context of the forecast system presented here in making automated traditional QPFs. Some work towards the rarity quantification has already been done: it is done indirectly here, and ECMWF's extreme forecast index (EFI), among others,

attempts to do this. But there is not, to my knowledge, yet an existing connection between this and automated forecast methodology; this connection is where I believe immense forecast value may be added. Lastly, though an effort was made in the forecast system framework presented here to use a forecast predictand as impact-proportional as possible in an atmosphere-only modeling framework, the connection is still less direct than desirable. Extending this work towards probabilistic hydrologic modeling, either dynamically or possibly statistically, could go a long way towards furthering this goal, and may prove a highly productive avenue to pursue.

8 References

Applequist, S., Gahrs, G. E., Pfeffer, R. L., & Niu, X.-F. (2002). Comparison of Methodologies for Probabilistic Quantitative Precipitation Forecasting*. *Weather and Forecasting, 17*(4), 783-799. doi:10.1175/1520-0434(2002)017<0783:COMFPQ>2.0.CO;2

Benjamin, S. G., Dévényi, D., Weygandt, S. S., Brundage, K. J., Brown, J. M., Grell, G. A., . . . Smith, T. L. (2004). An hourly assimilation-forecast cycle: The RUC. *Monthly Weather Review, 132*(2), 495-518.

Bentzien, S., & Friederichs, P. (2012). Generating and Calibrating Probabilistic Quantitative Precipitation Forecasts from the High-Resolution NWP Model COSMO-DE. *Weather and Forecasting, 27*(4), 988-1002. doi:10.1175/WAF-D-11-00101.1

Bjørnar Bremnes, J. (2004). Probabilistic Forecasts of Precipitation in Terms of Quantiles Using NWP Model Output. *Monthly Weather Review, 132*(1), 338-347. doi:10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2

Bonnin, G., Martin, D., Lin, B., Parzybok, T., Yekta, M., & Riley, D. (2004). Point precipitation frequency estimates, precipitation-frequency atlas of the United States NOAA atlas 14, v. 2, version 3 NOAA, National Weather Service, Silver Spring, Maryland; accessed January 4, 2007.

Bonnin, G., Martin, D., Lin, B., Parzybok, T., Yekta, M., & Riley, D. (2006). Precipitation-frequency atlas of the United States: National Oceanic and Atmospheric Administration (NOAA) atlas 14, v. 1, version 4. *National Weather Service, Silver Spring, Md., available online at http://hdsc. nws. noaa. gov/hdsc/pfds*.

Breiman, L. (2001). Random forests. *Machine learning, 45*(1), 5-32.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78*(1), 1-3.

Buizza, R., Hollingsworth, A., Lalaurette, F., & Ghelli, A. (1999). Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Weather and Forecasting, 14*(2), 168-189.

Clark, A. J., Kain, J. S., Stensrud, D. J., Xue, M., Kong, F., Coniglio, M. C., . . . Gao, J. (2011). Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Monthly Weather Review, 139*(5), 1410-1418.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning, 20*(3), 273-297.

De Haan, L., & Ferreira, A. (2007). *Extreme value theory: an introduction*: Springer Science & Business Media.

Doswell III, C. A., Brooks, H. E., & Maddox, R. A. (1996). Flash flood forecasting: An ingredients-based methodology. *Weather and Forecasting, 11*(4), 560-581.

Eckel, F. A. (2003). *Effective mesoscale, short-range ensemble forecasting.* University of Washington.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology, 8*(6), 985-987.

Friederichs, P. (2010). Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes, 13*(2), 109-132.

Friederichs, P., & Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly Weather Review, 135*(6), 2365-2378.

Friederichs, P., & Hense, A. (2008). A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting, 23*(4), 659-673.

Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, 1189-1232.

Fritsch, J. M., & Carbone, R. E. (2004). Improving Quantitative Precipitation Forecasts in the Warm Season: A USWRP Research and Development Strategy. *Bulletin of the American Meteorological Society, 85*(7), 955-965. doi:10.1175/BAMS-85-7-955

Germann, U., Zawadzki, I., & Turner, B. (2006). Predictability of precipitation from continental radar images part IV: limits to prediction. *Journal of the Atmospheric Sciences*, 63(8), 2092-2108.

Glahn, H. R., & Lowry, D. A. (1972). The use of model output statistics (MOS) in objective weather forecasting. *Journal of Applied Meteorology, 11*(8), 1203-1211.

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 69*(2), 243-268.

Guttman, N. B., Hosking, J., & Wallis, J. R. (1993). Regional precipitation quantile values for the continental United States computed from L-moments. *Journal of Climate, 6*(12), 2326-2340.

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau Jr, T. J., . . . Lapenta, W. (2013). NOAA's second-generation global medium-range ensemble reforecast dataset. *Bulletin of the American Meteorological Society, 94*(10), 1553-1565.

Hamill, T. M., & Colucci, S. J. (1997). Verification of Eta-RSM short-range ensemble forecasts. *Monthly Weather Review, 125*(6), 1312-1327.

Hamill, T. M., & Whitaker, J. S. (2006). Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review, 134*(11), 3209-3229.

Hamill, T. M., Whitaker, J. S., & Mullen, S. L. (2006). Reforecasts: An important dataset for improving weather predictions. *Bulletin of the American Meteorological Society, 87*(1), 33-46.

Hamill, T. M., Whitaker, J. S., & Wei, X. (2004). Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly Weather Review, 132*(6), 1434-1447. doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029-1054.

Hershfield, D. M. (1961). Technical Paper No. 40: Rainfall Frequency Atlas of the United States, Department of Commerce. *Weather Bureau, Washington, DC*.

Hosking, J. (1992). Moments or L moments? An example comparing two measures of distributional shape. *The American Statistician, 46*(3), 186-189.

Hosking, J. (1997). FORTRAN routines for use with the method of L-moments, Version 3.04. *IBM Research*.

Hosking, J. (2006). On the characterization of distributions by their L-moments. *Journal of Statistical Planning and Inference, 136*(1), 193-198.

Hosking, J., & Wallis, J. (1993). Some statistics useful in regional frequency analysis. *Water Resources Research, 29*(2), 271-281.

Hosking, J. R., & Wallis, J. R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics, 29*(3), 339-349.

Hosking, J. R. M., & Wallis, J. R. (2005). *Regional frequency analysis: an approach based on L-moments*: Cambridge University Press.

Hsu, W.-r., & Murphy, A. H. (1986). The attributes diagram A geometrical framework for assessing the quality of probability forecasts. *International Journal of Forecasting, 2*(3), 285-293.

Kalnay, E. (2003). *Atmospheric modeling, data assimilation, and predictability*: Cambridge university press.

Leith, C. (1974). Theoretical skill of Monte Carlo forecasts. *Monthly Weather Review, 102*(6), 409-418.

Lin, Y., & Mitchell, K. E. (2005). *1.2 the NCEP stage II/IV hourly precipitation analyses: Development and applications.* Paper presented at the 19th Conf. Hydrology, American Meteorological Society, San Diego, CA, USA.

Marsh, P. T., Kain, J. S., Lakshmanan, V., Clark, A. J., Hitchens, N. M., & Hardy, J. (2012). A method for calibrating deterministic forecasts of rare events. *Weather and Forecasting, 27*(2), 531-538.

Marzban, C. (1998). Scalar measures of performance in rare-event situations. *Weather and Forecasting,* 13(3), 753-763.

Miller, J., Frederick, R., & Tracey, R. (1973). NOAA atlas 2. *Precipitation-frequency atlas of the western United States, 3*.

Mullen, S. L., & Buizza, R. (2002). The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF Ensemble Prediction System. *Weather and Forecasting, 17*(2), 173-191.

Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology, 12*(4), 595-600.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*: MIT press.

Mylne, K. R. (2002). Decision-making from probability forecasts based on forecast value. *Meteorological Applications, 9*(3), 307-315.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825-2830.

Perica, S., Dietz, S., Heim, S., Hiner, L., Maitaria, K., Martin, D., . . . Unruh, D. (2011). Noaa Atlas 14 Volume 6 Version 2.0, Precipitation-Frequency Atlas of the United States, California. *NOAA, National Weather Service, Silver Spring, MD*.

Perica, S., Martin, D., Pavlovic, S., Roy, I., St Laurent, M., Trypaluk, C., . . . Bonnin, G. (2013). NOAA Atlas 14 Volume 9 Version 2. Precipitation-Frequency Atlas of the United States, Southeastern States. *NOAA, National Weather Service: Silver Spring, MD, 171*.

Pilon, P. J., & Adamowski, K. (1992). The value of regional information to flood frequency analysis using the method of L-moments. *Canadian Journal of Civil Engineering, 19*(1), 137-147.

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review, 133*(5), 1155-1174. doi:10.1175/MWR2906.1

Ralph, F., & Dettinger, M. (2011). Storms, floods, and the science of atmospheric rivers. *Eos, Transactions American Geophysical Union, 92*(32), 265-266.

Roberts, N. M., & Lean, H. W. (2008). Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review, 136*(1), 78-97.

Roebber, P. J. (2013). Using evolutionary programming to generate skillful extreme value probabilistic forecasts. *Monthly Weather Review, 141*(9), 3170-3185.

Rogers, E., Black, T., Deaven, D., DiMego, G., Zhao, Q., Baldwin, M., . . . Lin, Y. *Coauthors, 2009: The NCEP North American Mesoscale modeling system: Recent changes and future plans.* Paper presented at the Preprints, 23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction, Omaha, NE, Amer. Meteor. Soc. A.

Roulin, E., & Vannitsem, S. (2011). Postprocessing of Ensemble Precipitation Predictions with Extended Logistic Regression Based on Hindcasts. *Monthly Weather Review, 140*(3), 874-888. doi:10.1175/MWR-D-11-00062.1

Schaffer, C. J., Gallus, W. A., & Segal, M. (2011). Improving Probabilistic Ensemble Forecasts of Convection through the Application of QPF–POP Relationships. *Weather and Forecasting, 26*(3), 319-336. doi:10.1175/2010WAF2222447.1

Scheuerer, M., & Hamill, T. M. (2015). Statistical Post-Processing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions. *Monthly Weather Review*. doi:10.1175/MWR-D-15-0061.1

Schumacher, R. S., & Johnson, R. H. (2006). Characteristics of US extreme rain events during 1999-2003. *Weather and Forecasting, 21*(1), 69-85.

Schwartz, C. S., Kain, J. S., Weiss, S. J., Xue, M., Bright, D. R., Kong, F., . . . Wandishin, M. S. (2010). Toward Improved Convection-Allowing Ensembles: Model Physics Sensitivities and Optimizing

Probabilistic Guidance with Small Ensemble Membership. *Weather and Forecasting, 25*(1), 263-280. doi:10.1175/2009WAF2222267.1

Skamarock, W. C., & Klemp, J. B. (2008). A time-split nonhydrostatic atmospheric model for weather research and forecasting applications. *Journal of Computational Physics, 227*(7), 3465-3485.

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., & Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Monthly Weather Review, 135*(9), 3209-3220.

Sobash, R. A., Kain, J. S., Bright, D. R., Dean, A. R., Coniglio, M. C., & Weiss, S. J. (2011). Probabilistic forecast guidance for severe thunderstorms based on the identification of extreme phenomena in convection-allowing model forecasts. *Weather and Forecasting, 26*(5), 714-728.

Coe, R., & Stern R.D. (1984). Fitting models to daily rainfall data. *Journal of Applied Meteorology, 21(7), 1024-1031.*

Stevenson, S. N., & Schumacher, R. S. (2014). A 10-year survey of extreme rainfall events in the central and eastern United States using gridded multisensor precipitation analyses. *Monthly Weather Review, 142*(9), 3147-3162.

Theis, S., Hense, A., & Damrath, U. (2005). Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorological Applications, 12*(3), 257-268.

Waxberg, G. (2015). Detail-Oriented: National Weather Service Introduces the HRRR. *Weatherwise, 68*(2), 28-33. doi:10.1080/00431672.2015.997565

Wilks, D. S. (2011). *Statistical methods in the atmospheric sciences* (Vol. 100): Academic press.

Wilks, D. S., & Hamill, T. M. (2007). Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review, 135*(6), 2379-2390.

Williams, R., Ferro, C., & Kwasniok, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society, 140*(680), 1112-1120.

Wilson, P.S. & Toumi R. (2005). A fundamental probability distribution for heavy rainfall. *Geophysical Research Letters, 32(14).*

Yussouf, N., & Stensrud, D. J. (2008). Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecasting system during the 2005/06 cool season. *Monthly Weather Review, 136*(6), 2157-2172.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., & Mylne, K. (2002). The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society, 83*(1), 73-83.