VERIFICATION OF RIVER STAGE FORECASTS

by

Edwin Welles

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF HYDROLOGY AND WATER RESOURCES

In Partial Fulfillment of the Requirements
For the Degree

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN HYDROLOGY

In the Graduate College

THE UNIVERSITY OF ARIZONA

2 0 0 5

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by Edwin Welles

entitled Verification of River Stage Forecasts

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

_____ Date: April 19, 2005
Soroosh Sorooshian

_____ Date: April 19, 2005
Hoshin Vijai Gupta

_____ Date: April 19, 2005
Donald R. Davis

_____ Date: April 19, 2005
William J. Shuttleworth

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

_____ Date: April 19, 2005
Dissertation Director: Soroosh Sorooshian

STATEMENT BY AUTHOR

This dissertation as been submitted in partial fulfillment of the requirements for an advanced degree at the The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department of the dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED:_____

Edwin Welles

ACKNOWLEDGEMENTS

Over the long course of this project many, many, many people helped me. First of all, the professors and my fellow students at the University of Arizona provided me with the background I needed to undertake this project. Then the professors and students at the University of Maryland helped out when I needed a couple more classes. The Department of Energy Global Change Fellowship helped to fund my initial scholastic endavours. The National Weather Service staffs at the Office of Hydrology, the Hydrologic Services Division and the River Forecast Centers provided me with invaluable insight into the hydrologic science and the *Hydrologic Forecast Science*. The NWS also supported me in my second round of classes and dissertation credits monetarily and with computer resources. I also thank my committee members who read and commented on my work. My advisor, Soroosh Sorooshian offered me numerous insightful comments that helped me to understand the work I had done. My family and friends supported me throughout with encouragment and patience.

Many thanks to everyone.

DEDICATION

Though it is an infinitesimal token of my gratitude for their help and encouragement, this dissertation is dedicated to my lovely wife and my two wonderful boys.

TABLE OF CONTENTS

TABLE OF CONTENTS – Continued

TABLE OF CONTENTS – Continued

TABLE OF CONTENTS – Continued

LIST OF FIGURES

LIST OF FIGURES – Continued

LIST OF TABLES

ABSTRACT

Little verification of hydrologic forecasts has been conducted to date, and therefore little is known about the skill of hydrologic forecasts. This dissertation presents a verification study of river stage forecasts with lead-times up to three days for sixteen locations in the United States for a period spanning the past decade. The verification metrics from this limited sample indicate that the below flood stage forecasts are skillful, and so are the day 1 above flood stage forecasts. However, by day 3, the longer lead-time, above flood stage forecasts appear to have little skill (when compared with simple persistence). Further, they have not improved during the period of record despite a number of forecast process improvements. A path to improving the forecasts is suggested, via a new approach to selecting enhancements to the hydrologic forecast process. In support of this method, two fundamental building blocks of a robust verification program are presented: a method to pinpoint sources of skill in forecasts, and a standardized process for verifying forecasts.

One element of a complete verification system is a process to determine why forecasts behave as they do. Forecasters need to be able to determine what causes a forecast to be good and what causes it to be bad. Therefore, an operationally implementable method for conducting this type of verification analysis is described and demonstrated. The method is used to evaluate the influence of model calibration, model initial conditions, and precipitation forecasts on the skill of single-valued (deterministic) river forecasts.

A second important element of any forecast process, is a well defined, standard verification methodology. This dissertation proposes a standard verification system for deterministic river forecasts as a foundation for future discussions and for development of a well accepted set of verification practices for hydrologic forecasts. The proposed standards account for the needs of users, forecasters, scientists and administrators and are designed to be easily implemented within the constraints of an operational system.

1 INTRODUCTION

Hydrologic forecasting supports the safety and economic well being of our nation. River forecasts warn of future floods to save lives and property, and water volume forecasts support the efficient management of our water supplies for drinking, agriculture and industry. Hydrologic forecasting is also very difficult. The forecasts depend upon imperfect, mathematical descriptions of the physical processes governing runoff generation and river routing. Hydrologic forecasts depend upon meteorological forecasts, and therefore, the hydrologic forecasts include all the uncertainty in the meteorological forecasts. The forecasts must also account for the human intervention in the river systems through the operation of reservoirs and locks. Like any other scientific problem, hydrologic forecasting requires rigorous objective analysis, including objective analysis of the forecast skill, objective analysis of the sources of error and skill in the forecasts, and objective analysis of the way enhancements made to the forecast process change the forecast skill. This objective, analytic framework is provided by verification, which refers to the use of objective measures to study forecasts and their corresponding observations.

Historically, the verification of hydrologic forecasts has received little attention from the operational and scientific communities. As a result, although hydrologic forecast products are issued routinely, decision makers have cause to wonder how much confidence to place in any specific forecast, with potentially serious consequences for

public safety, environmental conditions, and economic well-being. Further, decisions must be made about investments in hydrologic research and operational forecast system upgrades. With little forecast evaluation having been conducted to date, forecast users, researchers, forecasters, and administrators have little concrete, objective guidance for their decisions.

Almost a decade ago, when the National Research Council (NRC) reviewed US National Weather Service (NWS) hydro-meteorological operations, they identified the need to begin verifying hydrologic forecasts. The NRC report stated "the verification of hydrologic forecasts is inadequate," (NRC, 1996, pp 3) and one high priority recommendation was that the NWS implement a comprehensive verification program for hydrologic forecasts. Since then, the NWS has implemented modest evaluation programs nationally and regionally, but these programs are far from complete. The programs characterize a few aspects of the forecast skill at a few selected locations, but do not provide a comprehensive control on the forecast process. In addition, only a few forecast evaluation studies have been conducted by the larger hydrology community and none of these studies considered the skill of short term single-valued river stage forecasts (the most common type of hydrologic forecast). One indicator of the extent to which hydrologists have ignored forecast verification is the current edition of the World Meteorological 0rganization (WMO) Guide to Hydrological Practices (WMO, 1994); it

does not mention verification[1]. As Stephenson and Joliffe (2004) put it when referring to flood forecasts "Such forecasts are difficult to verify" (p. 196). Not only do hydrologists not verify their forecasts, there even appear to be some who think that they can not be verified.

In their work with quantitative precipitation forecasts, Krzysztofowicz and Sigrest (1999a) found that just the simple act of performing a verification of forecasts leads to improved forecast skill. They provided forecasters with verification statistics the month after they issued their forecasts, and although, Krzysztofowicz and Sigrest made no changes to the forecast process, the forecasts tangibly improved. They inferred from this observation, that the forecasters have an opportunity to "calibrate" themselves when they are confronted with objective verification measures to review. If these meteorological results are representative, then hydrologists have a lot to gain from verifying their forecasts.

To draw attention to this important, but under-studied, component of the hydrologic forecast process, the work reported here conducted an assessment of National Weather Service short term (< 3 day leadtime) river stage forecasts. A surprising outcome of this assessment (presented in Section 3 of this dissertation) is that contrary to expectations, the forecasts do not appear to be very skillful, nor have they improved with time. While

1 The new edition of the WMO Guide to Hydrological Practices, being written now, will include a section on forecast verification (Personal communication, C. Barrett, Jan 2004).

the sample size for the study is small (only 16 forecast locations) evidence from other areas in the country suggest that these results are likely to be representative of the skill of NWS river stage forecasts across the Nation.

In light of these results, a new approach to developing the forecast process is proposed. Because so little forecast verification has been conducted to-date, the direction of development of the forecast process has necessarily relied upon the expert opinions of trusted scientists. This method does not appear to have yielded the desired results. It is therefore, proposed that the development of the forecast process be driven by objective verification metrics instead of expert opinion. In other words, by beginning to verify forecasts, hydrologists can turn the process of hydrologic forecast development into an objective, scientific endeavor.

To arrive at the needed systematic and objective description of hydrologic forecast skill two things are needed; 1) a well defined methodology for determining sources of error in the forecasts, and 2) standards for verifying forecasts. The error analysis is needed to provide a means of identifying which elements of the forecast process require enhancement. Standardized verification procedures are needed to provide a consistent framework for verifying forecasts thereby improving communication between forecasting groups. Both of these tasks can benefit by borrowing from the procedures and methods being used by the meteorological community.

The scope of this dissertation is as follows. A methodology for the error analysis is presented in Section 4. The methodology is based upon simple comparisons of different forecast configurations, and has the advantage of being easily implemented into an operational setting. An illustrative analysis is presented in which the relative importance of runoff model calibration, initial state updating, and precipitation forecasts is evaluated for river stage forecasts with lead-times out to three days. The results demonstrate that simple comparisons can be used to pinpoint sources of forecast error and skill.

Next, a proposal for a standardized verification system for short term river forecasts is presented in Section 5. The standards include a description of the data to be archived, the metrics to be computed and the procedures for analyzing the forecast process. The proposed standards explicitly address the needs of administrators, forecasters, users and scientists. It is anticipated these standards will be dynamic and they will be updated when the forecast process changes and as improved methods for evaluating forecasts are discovered.

Finally, a concluding discussion summarizes the contributions of this work to hydrologic science and practice and proposes a new direction for future research for the scientific community to consider: the development of alternative forecast processes within the framework of an objective verification system.

2 LITERATURE REVIEW

A science of forecast verification has been developed by the meteorological forecast community. They have constructed verification measures; analyzed the characteristics of those measures; applied them to their forecasts; and adopted standard procedures for verifying forecasts. This literature review makes no attempt to provide an exhaustive description of the extensive meteorological literature on verification; it does, however, briefly introduce the meteorological literature, and then provide a detailed description of the hydrologic verification literature.

2.1 Brief Summary of Meteorological Literature

Murphy (1996) surveyed the development of meteorological forecast verification and credits the Finley (1884) paper on tornado forecasts with initiating the scientific discussion and the subsequent development of weather forecast verification techniques. Finley computed a Percent Correct of 96.6% on his tornado forecasts, and used this metric to declare his tornado forecasts skillful. However, his claim was energetically disputed by many including Gilbert (1884) who showed that Finley's Percent Correct would have been 98.2% had he always forecast “no tornado”. Since Finley's paper numerous alternative measures have been proposed, re-discovered, analyzed and applied.

Brier and Allen (1951) described the reasons for verifying forecasts and they sorted those reasons into three categories: 1) *administrative,* 2) *scientific* and 3) *economic*.

Administrative verification is descriptive, providing characterizations of the status of the forecast service. The status for one year may be compared to the status for the next year. A complete *administrative* verification program will include metrics to establish the efficacy of the entire forecast system: metrics for the timeliness of the forecast delivery, the number of forecasts issued and the like, in addition to an assessment of the overall forecast skill. The goal of *administrative* verification is to describe the overall skill and efficiency of the service so decisions can be made with respect to resource allocation, research directions and implementation strategies. *Scientific* verification is analytic, with a focus on understanding why forecasts perform well or poorly. The goal is to use objective methods to establish which elements of the forecast process control the skill and error of the forecasts under specific conditions. *Economic* verification refers to the evaluation of the economic benefit (or dis-benefit) accrued to the user through the improved decision making enabled by the forecasts. This class of verification is difficult to generalize because the economic impact of a set of forecasts for a user will be specific to each user. User oriented verification must be conducted in conjunction with the users themselves.

Whatever the purpose of verification, it is important to understand the characteristics of the metrics used to analyze the forecasts. Murphy (1996) identified the process of analyzing verification metrics themselves as “meta-verification”. The two most important “meta-verification” characteristics are that the scores must be *proper* and

equitable. A strictly *proper* score does not permit hedging. That is, a forecaster cannot improve his/her score by simply forecasting based upon the characteristics of the metric: for example, by using the most likely event. An *equitable* score (Gandin and Murphy 1992) is one for which all random or constant forecasts (i.e. no skill forecasts) result in the same expected value of the metric. More generally, Murphy (1993) proposed a set of characteristics for the ways a forecast can be good (or bad)m describing forecast goodness in terms of *consistency, quality, and value*. *Consistency* means the consistency between the forecasters' best estimate of the future event and the forecast. Q*uality* refers to the correspondence between the forecasts and the events they forecast, and v*alue* refers to the expense the users avoid or the benefits they accrue as a result of the forecasts. Most verification metrics characterize forecast *quality* with *equitable* and *proper* metrics thereby encouraging forecasters to be *consistent*.

Two substantial contributions to the verification science were made in the 1980's. First, Mason (1980 and 1982) introduced Signal Detection Theory and the Relative Operating Characteristics (ROC) to the verification literature. The ROC is most useful with probability forecasts as it allows users to compare the effects of different probability thresholds on the forecast skill and then summarize the overall skill of a set of forecasts. Single valued forecasts can be plotted on a ROC diagram to compare single- and multi-valued forecasts. The ROC also provides a means of assessing the meta-verification characteristics of categorical metrics.

Second, in the late 80's Murphy and Winkler (1987) proposed a general framework for verification based upon the joint distribution of forecasts and observations, *p(f,o).* They introduced the concepts of *Discrimination* and *Reliability* which are derived from factoring the joint distribution into conditional distributions given the observations, *p(f | o),* or given the forecasts, *p(o | f).* Each conditional distribution yields different information about the forecast-observed relation. Within the diagnostic framework, *Discrimination* refers to the ability of the forecasts to distinguish between future events. *Discrimination* is assessed by evaluating the conditional distributions of the forecasts given the observations, *p(f | o).* *Reliability* refers to the forecasts ability to say something correct; that is, if an event is forecast, did it occur. *Reliability* is evaluated with *p(o|f).*

In addition to developing the computational theory for verification, the meteorological community has verified forecasts in both a research and operational context. Evaluations of specific forecasts have been published (e.g. Livezey and Jamison, 2003; Wobus and Kalnay, 1995; Brooks et al, 1997). Hartmann et al. (2002) focused on user directed verification and demonstrated the importance of understanding the users perspective to determine the skill of the forecasts.

The meteorological community haa also developed verification standards to support forecast operations. For example, the NWS Operations Manual provides detailed

instructions and requirements to weather forecasters for verifying forecasts (NWS, 2001a). Environment Canada published a handbook of verification methods in 1989 which describes standard methods for verifying meteorological forecasts in order "bring consistency of thought ....to the design, execution and interpretation of weather forecast verification." (Stanski et al 1989, pp i) More recently, the WMO Commission for Basic Systems published a Standardized Verification System for Long Range Forecasts (WMO, 2002). The procedures can be applied to any long range meteorological forecast service. Standardized procedures provide a common language of verification for everyone who is interested in forecasting thereby enhancing the communication amongst forecast users, forecasters and forecast technique developers. An example of a comprehensive verification procedure is that of the Finnish Meteorological Institute (Nurmi et al, 2003). This extensive evaluation is published on a semi-annual basis and includes metrics for temperature, probability of precipitation, cloudiness, rain amount and marine weather. Categorical metrics, ROC diagrams, and the mean absolute error are computed and plotted for the various elements, and a summary of the results is provided.

The meteorological community has also produced several useful publications summarizing the current state of forecast verification science. Joliffe and Stephenson (2003) published a collection of chapters by verification experts surveying the breadth of the verification science in their book A Practitioner's Guide to Verification. While some consider their summary insufficiently broad (Glahn, 2004), Joliffe and Stephenson's book

offers anyone interested in verification an informative description of current verification practices. Wilks (1995) summarizes the commonly used metrics and provides useful examples to illustrate computational details. Stanski et al (1989, from their abstract) ”describe commonly used verification methods for weather elements and numerical weather prediction model forecasts ... in terms of a general verification model.” They identify and discuss “the advantages and disadvantages of each verification method” and include “numerous examples using meteorological data.” (Stanski et al, 1989, abstract) Several scientists have collaborated to post a web site sponsored jointly by the World Meteorological Organization (WMO), World Weather Research Program and the Working Group on Numerical Experimentation. (WMO, 2004) The web site summarizes reasons to do verification, how it is commonly done and links to other sites where scientists have posted verification research.

2.2 Hydrologic Verification

Just over one hundred years after Finley's paper, Morris (1988) made an early contribution to the discussion of hydrologic verification. Morris focused on flood events and used the categories of Minor, Moderate and Major flood to categorize the forecasts. He proposed several measures derived from the standard meteorological measures Probability of Detection (POD) and False Alarm Ratio (FAR), modifying them slightly to account for the differences in importance to the forecast user of the rising and falling limbs of the hydrograph. He proposed a new measure, the Average Categorical Error

(ACE). It is computed only for those forecasts which fall outside of the observed category, and it is the Mean Absolute Error (MAE) between the forecast value and the nearest boundary of the observed category. Morris also proposed measures to characterize the timing error in forecasts using a complex series of lead time calculations for each category.

No other authors have proposed methodologies for hydrologic verification; they have all focused upon applying meteorological techniques to hydrologic forecasts. Schwein (1996) evaluated the effect of precipitation forecasts on the resulting river forecasts by comparing river stage forecasts computed with and without precipitation forecasts. She used the Mean Error, Mean Absolute Error, and inspections of hydrographs to assess the differences between the forecast scenarios, concluding that precipitation forecasts do not harm the river forecast skill.

Franz et al (2003) conducted a hindcast study of NWS water supply forecasts. They converted these ensemble forecasts into deterministic forecasts and then used three metrics to evaluate them: 1) the Mean Absolute Error (MAE), 2) the Percent Bias and 3) the Correlation Coefficient. To facilitate comparisons between forecast locations, they added a Relative MAE which is the MAE divided by the standard deviation of the observations at the forecast location. They also conducted probabilistic verification of the ensembles.

In a similar vein, Pagano et al., (2004) evaluated 80 years of water supply forecasts issued by the Natural Resources Conservation Service (NRCS). They summarize previous evaluations of water supply forecasts and describe the change in the techniques for verification from simple differences between the forecasts and the observations to the current probabilistic verification techniques. For his evaluation of the forecasts they use the Nash-Sutcliffe efficiency to track changes in skill across the western U.S. over the 80 year period of record.

2.3 Operational Hydrologic Verification Schemes

Operational procedures for hydrologic forecast verification are only slightly more common than published schemes. The NWS, the largest supplier of hydrologic forecasts in the United States (and the agency mandated to issue Flood Watches and Warnings) has verified Flash Flood Warnings since 1986 using the FAR, the POD, the Critical Success Index (CSI) and the Lead Time. Recently, at the request of the Office of Management and Budget, the NWS added the Percent of Warnings with a positive Lead Time. The NWS is currently developing and implementing a verification program for River Flood Warnings; it will follow the same procedures and use the same measures as the Flash Flood Warning verification. Both of these programs evaluate the classic binary categorical event, in this case Flood/no-Flood. They do not evaluate the accuracy of the forecast river stages.

NWS River Flood Watches and Warnings are based upon stage time series forecasts issued by the NWS River Forecast Centers (RFCs). They are equivalent to model runs computed by meteorological forecast centers from which weather forecasts are constructed. When the National Research Council (NRC) reviewed the NWS hydrology program in 1996 (NRC, 1996), one of their high priority recommendations was for the NWS to implement a verification program for RFC time series forecasts. To that end, in April of 2001, the NWS initiated a program to collect the river stage time series issued by the RFCs at 173 of the 4000 NWS river forecast locations. The forecasts for each basin are paired with an observation and then sorted into one day, two day, and three day lead times with each pair sorted into above and below flood stage categories based upon the observations. To compute statistics, the selected 173 forecast locations are sorted into three river sizes based upon the average basin response time categories (listed in Table 1). For each subset of the forecast and observation pairs, the Mean Error (ME), the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) are computed. Flash Flood Watches and Warnings are issued from different datasets, and at this time there is no verification program for the computations which underly the Flash Flood Watches and Warnings.

In addition to the verification program implemented Nationally, the NWS Southern Region also implemented a verification program for RFC time series in 2001 (Olsen,

Category	River Response Time
Fast	< 24 hours
Medium	24 < and < 60 hours
Slow	> 60 hours

Table 1 River response categories

2001). The Southern Region program is based upon the work of Morris described above and the existing verification programs for meteorological forecasts. For the Southern Region program the FAR, the POD, the CSI, and the ACE are computed, as well as a simplified version of Morris' lead time calculations, an Average Categorical Lead Time (ACLT). The ACLT is computed as the lead time of the forecasts which fall in the same category as the observed. Because, public dissemination of verification metrics is a key to any successful verification program, the Arkansas Red River Basin River Forecast Center, has published the verification metrics for their forecast locations on their web site. However, the rest of the NWS Hydrology program has not followed this example, and neither the National metrics nor the metrics for the other RFCs have been published.

The longest standing NWS river stage forecast verification program is the NWS Central Region Mississippi and Missouri mainstem verification program. The Central Region has archived Mississippi and Missouri mainstem forecasts and observations since 1983 and they compute the RMSE at monthly time-steps from these data. Other federal, state and local agencies and university research groups also generate hydrologic forecasts and publicly visible verification programs are as sparse for these other forecast systems, as they are at the NWS.

One important limitation of the existing verification programs is that none has had the benefit of rigorous peer review. Through the work of the meteorological community, the

science of forecast verification is supported by a well explored theoretical foundation. An important step toward developing a robust tradition of verification within the hydrologic community will be to fold verification into the mainstream of hydrologic research with the review of operational verification strategies and the publication of verification results. In this way, operational procedures can be developed with the benefit of input from the scientific community, and by reviewing published verification results, the research community will be able to understand what research will help improve the forecast skill.

3 ADMINISTRATIVE VERIFICATION OF DETERMINISTIC RIVER STAGE FORECASTS

This section presents a verification study of U.S. National Weather Service (NWS) deterministic river stage forecasts at 16 locations in two regional datasets. One dataset is twenty years long, and the other ten years long. The purpose of this study is to describe the general characteristics of the forecast skill for each dataset and to address the following three specific questions.

- How does the performance of the actual forecasts compare to the performance of the persistence forecasts?
- How does the forecast performance change with lead time?
- How does the forecast skill change with time?

3.1 Method for Conducting the Administrative Verification

To conduct this verification study, a set of metrics were selected from the large set of possible metrics found in the meteorological literature, and these metrics were applied to data from two distinct regions. The forecasts and observations were paired together and then sorted into informative subsets with the selected metrics being computed on each subset.

3.1.1 Statistics to be Computed

From the multitude of verification statistics, a set of commonly used metrics were selected for this study. Because the NWS uses the FAR and the POD extensively in all

their verification programs, these are included here. The ME and RMSE are computed to address the accuracy of the forecasts. These are augmented by two popular dimensionless measures, the correlation coefficient (CC) and a RMSE Persistence Skill Score (SS). The sample size is reported as a companion for all statistics to provide the reader with an indication of the significance of the calculations. All these metrics are well established meteorological verification measures which have been reviewed by many people, and their characteristics are well known. Deque (2003), Livezey (2003), Mason (2003) provide excellent, recnet introductory summaries of these metrics.

3.1.2 Pairing and Sorting the Forecasts and Observations

To apply these metrics to the data, the forecasts and observations are first paired. They are paired by matching each forecast with the observation closest in time to the forecast valid time within a window of plus or minus one hour of the valid time. If no observation is found within this window the forecast is not included in the verification. These pairs are then sorted into subsets, and metrics are computed to characterize the subsets.

There are almost as many ways of sub-setting the forecast-observation pairs as there are useful metrics. Hydrologic errors vary strongly with stage; consequently, all the existing hydrologic verification schemes include sorting by the stage height. The pairs can be sorted by either the forecast or the observed value in the pair. However, in this study, only sorting by the observed value is used. Sorting by the forecast value is not suitable

for use as an *administrative* metric because the sample set is variable and this variability may result in statistics degrading while the forecasts have in fact improved. Sub-setting by the forecast value is more suitable for in depth analysis than for general characterizations of the forecast skill. In addition to sorting by the stage height, the forecast-observation pairs are sorted by lead-time and by year.

3.1.2.1 Selecting a Stage Threshold

To sort the forecast-observation pairs by stage height, the dynamic range of the rivers must be divided into categories. There are numerous possible methods for making this division: stage intervals, flow intervals, probability intervals and critical forecast thresholds. For the purposes of this study the most relevant division is at the Flood Stage for the forecast point because the most common use of these stage forecasts is issuing Flood Watches and Warnings. The other methods produce categories which are not related to this important forecast result and therefore are not as relevant to the forecast use.

The Flood Stage, as defined by the NWS, is the "gage height at which a watercourse overtops its banks and begins to cause damage to any portion of the defined reach." (NWS, 2001b) While the Flood Stage is well defined, it is not an objectively determined value. The Flood Stage for a location is determined by the NWS staff at the local Weather Forecast Office in cooperation with local authorities The Flood Stage may be

changed because the characteristics of the forecast location change or it may be changed to accommodate some other forecast need. For example, a forecast office may raise the flood stage in order to reduce the frequency with which they issue Flood Warnings. However, as the primary purpose of these forecasts is issuing Flood Watches and Warnings, the Flood Stage will be used.

3.1.3 Persistence as a Control Forecast

A control forecast is essential to understanding the magnitudes of the computed metrics. Without a baseline to provide a perspective, it is not possible to determine if the magnitudes of the metrics indicate skill or not. Nor is it possible to determine if trends are the result of changes in the forecast process or changes in the seasonal weather patterns. Persistence is used as the baseline in this study. A persistence forecast is constructed for each of the issued forecasts in the data set with a persistence forecast defined as the observation at the time the forecast is issued. Although, lagged persistence is considered to be a more informative baseline than ordinary persistence for climate forecasts (WMO, 2002), the comparison of lagged persistence (also known as a one-step auto-regressive model (AR-1)) with ordinary persistence is outside the scope of this review and is left for a later study.

3.2 Data to be Used for the Administrative Verification

The metrics listed above are applied to two data sets consisting of time series forecasts issued by NWS River Forecast Centers (RFCs). One dataset is from forecast locations in Arkansas and Oklahoma the other is a dataset from forecast points along the mainstem of the Missouri River. The Arkansas/Oklahoma (A/O) data set consists of 10 years of forecasts and observations starting on April 1, 1993 and ending on November 30, 2002 for 4 locations in Oklahoma and 1 location in Arkansas. Annual precipitation varies from 30 to 45 inches on these basins with the majority falling in the spring. Frontal storms dominate the winter weather while convective thunderstorms dominate the summer. The Missouri mainstem (MM) data set consists of twenty years of forecasts and observations starting on January 1, 1983 and ending on November 30, 2002 for 11 locations along the mainstem of the Missouri river. Table 2 lists the total drainage area, the number of modeled upstream basins, the Flood Stage and the record flood stage for each forecast location.

3.2.1 Forecast Modeling Environment

The forecast modeling environment consists of semi-distributed, watershed based (non-gridded), locally calibrated model collections. For each basin, a collection of models is linked together to represent the physical processes present in the basin. The models will vary from basin to basin as the physical processes in the hydrologic system vary from basin to basin. To account for variations in the physical processes across a single

Location	Drainage Area (Sq Mi)	Modeled Basins	Flood Stage (ft)	Record Flood (ft)
Spring River at Quapaw, OK	2510	4	20.00	46.60
Illinois River at Watts, OK	635	4	13.00	25.96
Glover River at Glover, OK	315	1	16.00	29.72
Chickaskia River at Blackwell, OK	1859	2	29.00	34.38
Arkansas River at Morrilton, AR	155,479	369	30.00	42.00
Location	Drainage Area (Sq Mi)	Upstream Basins	Flood Stage (ft)	Record Flood (ft)
Missouri River at South Sioux City, NE	318,559	424	30.00	30.77
Missouri River at Omaha, NE	326,759	457	29.00	40.20
Missouri River at Nebraska City, NE	413,959	660	18.00	27.70
Missouri River at Rulo, NE	418,859	673	17.00	25.60
Missouri River at St. Joseph, MO	420,000	684	17.00	32.07
Missouri River at Waverly, MO	485,900	929	20.00	31.15
Missouri River at Glasgow, MO	498,900	997	25.00	39.50
Missouri River at Boonville, MO	500,700	1006	21.00	37.10
Missouri River at Jefferson City, MO	501,000	1014	23.00	38.30
Missouri River at Hermann, MO	522,500	1093	21.00	37.00
Missouri River at St. Charles, MO	524,000	1095	25.00	40.00

Table 2 Forecast point characteristics

watershed, a basin may be divided into sub-areas and a unique collection of models used on each sub-area. Most commonly, basins are divided to accommodate differences in snow accumulation and ablation due to large elevation differences between the top and the bottom of a basin in mountainous terrain. Water from upstream basins is routed through downstream basins with the local runoff added to the routed flow. All the models are calibrated prior to being used to issue forecasts.

On downstream basins the forecasts are an accumulation of numerous model calculations. Table 2 lists the number of basins upstream of the forecast locations and for example, it can be seen in this table that the Rulo, NE location on the Missouri mainstem has 673 basins upstream of the forecast location. Therefore, there are at least 673 uniquely parameterized model collections used to forecast the Rulo flows. (There will be more model collections than the 673 upstream basins because multiple elevation zones are used in the mountains.)

The RFC forecasters have over 30 models to choose from when setting up the models for a basin (NWS, 2003a). Those used for the majority of the modeling in these two datasets are described here. The Sacramento Soil Moisture Accounting model (Burnash, 1973) is used to compute the runoff from precipitation and a unit hydrograph (Linsley et al, 1975) is used to route the water from across the basin to the basin outflow. For snow covered basins, the Anderson snow model (Anderson, 1973) is used to model snow accumulation

and ablation. As noted above, basins with large elevation differences between the top and bottom of the basin are broken into elevation zones with unique snow models, Sacramento models, input precipitation and temperature time series for each zone. The runoff from the multiple Sacramento models is aggregated via an areally weighted average and a single unit hydrograph is used to route the aggregated runoff to the basin outflow location. For non-headwater locations, water from upstream basins is routed through the downstream basin with the Lag/K method (Linsley et al, 1975) calibrated to the local river reach. The runoff generated on the local basin is then added to the routed flow. The forecast outflow of a basin is statistically post-processed using a simple linear difference scheme to account for biases between the last observed and last simulated values (NWS, 2002a). When water is routed downstream from upstream basins, it is the post-processed time series which are routed downstream. Observed stage data is converted to flow (and the forecast flows are converted to stages) via rating curves which are updated as regularly as the gauging agency makes updates available. The models used for computing the forecasts are manually calibrated according to standard NWS procedures (NWS, 2002b) with a single set of parameters per model area.

There are numerous reservoirs, locks and diversions above the MM forecast locations and above the Morrilton, Arkansas forecast location in the A/O dataset. The reservoir/lock/diversion modeling strategy varies considerably from forecast location to forecast location. The reservoir operator may provide a forecast of their intentions

(usually as reservoir outflows) to the NWS forecasters. In which case, these forecast outflows are used as the reservoir outflow and they are routed downstream. In cases where the operating organization cannot or will not provide a forecast, the reservoir operating rules may be modeled or the diversion may be estimated based upon expected consumptive use. Where the reservoir or diversion is small, or the operations very unpredictable, the reservoir or diversion may be ignored. Reservoirs and diversions pose a considerable modeling problem to skillful river forecasting.

Precipitation and temperature time series to drive the models are computed as areal averages for each model area. The observed precipitation input to the models is computed from NEXRAD RADAR rainfall estimates and gauge measurements using the P1 methodology (Young et al., 2000) for the A/O data set and the method of Seo (Seo and Breidenbach 2002, Breidenbach et al., 1999, Seo 1998)) for the MM dataset. During the winter months in the MM basins, gauge only estimates of the precipitation may be used when the forecasters determine the RADAR precipitation estimates are likely to be unreliable due to winter weather conditions. Twenty-four hours of Quantitative Precipitation Forecasts (QPF) are used as input to the model, with zero used for the QPF after twenty-four hours. Where temperatures are used, they are computed as basin averages from stations using distance and elevation weighting parameters defined by the forecasters prior to forecast time. Daily station Max/Min temperature forecasts are converted to six hour areal average time series using the same distance and elevation

weighting schemes used for the observed temperatures. The interpolated Max/Min forecasts are disaggregated to six hour time steps using a fixed diurnal cycle. Potential evaporation (PE) is used in the A/O basins and it is computed from air temperature, dew point, wind speed, and solar radiation (NWS, 2003b). Future PE values are computed as a blend from the last observed value to the monthly climatic average.

3.2.2 Daily Forecast Process

The daily forecast process begins with quality controlling the input data and input forecasts, followed by an assessment of the model simulations and it finishes with quality control of the output forecasts. All input forecasts and observations are reviewed prior to using them for the model simulations. The methods of quality control depend upon the data and the forecast office; they consist of visual inspection of geographic trends, visual inspection of temporal trends, statistical range checking and nearest neighbor checks. Once the data has been reviewed and corrected as needed, the models are run and the simulations are assessed by the forecasters using visual techniques. The observed and simulated stage time series are compared in the observed period; the simulated model states are compared to the forecasters' expectation for the model states given the known conditions in the basins. Adjustments to the model inputs, the model states or to the output of the models are made to cause the models to perform according to the forecasters expert knowledge of the hydrologic system and the forecasters knowledge of the abstraction of the hydrologic system into the models. Finally, the forecasts themselves

are quality controlled by comparing the forecasters' expectation of the river response given the known hydrologic conditions to the forecast river response.

The forecasts are issued every morning with updates in the afternoons and evenings issued on an as needed basis. Observed data arrives at the forecast offices continuously and it is quality controlled as it arrives. However, the offices receive a large quantity of data shortly after 12 Greenwich Mean Time (GMT) as most 24 hour stations report at 12 GMT, and all this data must be quality controlled before it is used in the forecasts. The RFCs receive QPF guidance from the NWS Hydrometeorological Prediction Center (HPC) every six hours from which the RFC forecasters generate the QPF they pass to the hydrologic models. The local QPF is generally not considered very important as only a very short period of precipitation forecasts are used. Considerable energy, however, is spent upon the observed precipitation. Once the observations and the input forecasts have been prepared, the RFC hydrologic forecasters can begin forecasting the rivers. Where appropriate they contact cooperating agencies to coordinate their forecasts with the forecasts issued by the other agencies. The time requirements for issuing the forecasts vary by office and the needs of their users, but generally all the forecasts must be issued within several hours of the 12 GMT and the 00 GMT synoptic times. Most basins are quality controlled every day, however, some forecast locations are “Flood Only” locations. Forecasts for these locations are not issued unless the forecasters expect the river to rise above flood stage. Forecast updates are made throughout the day as needed depending upon user requests, the stages in the rivers, and the meteorological forecasts.

It is worth noting that the inputs and methods for hydrologic forecasting have a different structure than those used for meteorological forecasting. Meteorological circulation models run forwards into the future based upon internal model dynamics and solar forcing. They solve the equations of motion and the uncertainties associated with the model output are internal to the models. With the hydrologic forecasts on the other hand, the models are driven into the future with forecast precipitation and temperature and therefore much of the uncertainty in the forecasts is exogenous to the hydrologic modeling process. One important task in the development of robust hydrologic forecast verification will be to analyze how the internal and external sources of error interact and limit the predictability of the hydrologic systems.

3.2.3 Forecast Process Updates

Throughout the ten and twenty year periods of record for the forecasts studied here, the NWS has made many updates to the forecast modeling system and other elements of the forecast process with the intention of improving the forecasts. The updates vary from enhanced computing power, to new models, to new data displays, and Table 3 provides a time line of the major enhancements made to these two datasets. The rainfall-runoff models were initially Antecedent Precipitation Index (API) (NWS, 2003c) models and

Arkansas/Oklahoma Data Set Forecast Process Updates	
1993	Began using multi-sensor precipitation estimates to drive hydrologic models and Oklahoma Mesonet comes on line.
1995	Automated data quality control and graphical forecast monitoring software developed and implemented
1996	Switched multi-sensor precipitation processing to P1 algorithm
1999	Sacramento, Unit Hydrograph, Anderson snow model re-calibration using gauge climate precipitation and temperature
2000	Sacramento model re-calibration using operational multi-sensor precipitation estimates
Missouri Mainstem Data Set Forecast Process Updates	
Year	Process Update
1994	Switched from API to SAC-SMA runoff model in some mountainous areas
1996	Evening forecast shifts started
1996	All forecasts include QPF
1996	Multi-sensor (RADAR and gauge) precipitation used as primary precipitation input time series from late spring to early fall.
1997	Switched from API to SAC-SMA runoff model in the rest of the Missouri Basin using regionalized parameters
1997	Results of multi year calibration project for SAC-SMA parameters implemented basin wide including recalibration of mountains
1998	All forecast groups have 6-hour time-step
2000,	Completed modeling and implementation of areas below Canyon Ferry Dam in the Musselshell basin a part of the Upper Missouri basin, areas above Canyon Ferry Dam in the Upper Missouri and large portion of the sand hills of Nebraska. Fully defined modeling for Ft. Peck Reservoir.
2001	Improvements made to Lower Dakota Tributaries and additions to the Upper Missouri from below Canyon Ferry Dam down stream to Great Falls, MT.

Table 3 Forecast process updates

they have been replaced by the Sacramento model. The Sacramento model has been re-calibrated using longer historical data sets and improved tools and techniques as these become available. For the MM dataset, the Sacramento model was initially implemented using regional calibration methods: one set of parameters for large geographically and geologically similar regions. Currently, through the Advanced Hydrologic Prediction Services (AHPS) program, parameters for individual basins are being developed. The model state updating process was enhanced with improved displays of observed and simulated variables and user interfaces to improve the forecasters' adjustments to the model states. The algorithm for computing the multi-sensor estimates of observed precipitation is also under continuous development and it has transitioned from the original NEXRAD process (Ahnert et al. 1983, Ahnert et al 1986, Hudlow et al. 1988) to the current process (Seo and Breidenbach 2002, Breidenbach et al. 1999, Seo 1998). The number of precipitation gauges and the frequency of the observations has also generally increased with the implementation of the Automated Surface Observing System (ASOS), Data Collection Platforms (DCPs), Local Flood Warning Systems and Mesonets. The observed data quality control procedures have been enhanced through improved displays of range checking and spatial anomaly information as well as improved displays of the data itself. The precipitation forecast process has been updated because it was found the QPF had a patchwork characteristic when the forecasts from multiple Weather Forecast Offices (WFOs) were aggregated into a single composite (NWS, 1999). The QPF is now generated at the RFC rather than WFOs. The effect of these forecast process

improvements is not readily visible in the evaluation results presented here, indicating the need for more extensive verification efforts within the hydrologic community.

3.3 Results of the Administrative Verification

The most instructive results of this verification analysis are summarized in the following five bullets. A summary discussion of each bullet follows in subsequent sections along with graphs of the results.

- First, the below Flood Stage forecasts are accurate (in the sense of absolute values of the statistics) and skillful (in the sense of skill over persistence) for the one, two and three day lead-times considered here.
- Second, the above Flood Stage forecasts are also accurate and skillful for day one, though the skill converges toward the persistence skill by day 3.
- Fourth, the Missouri Mainstem (MM) forecasts are much more accurate than the Arkansas/Oklahoma (A/O) forecasts, but the two datasets show comparable skill over persistence.
- Fifth, few of the metrics show trends of improvement over the ten and twenty year periods of record.

3.3.1 Below Flood Stage Forecast Skill

The below flood stage metrics for the two data sets are plotted in Figures 1-4. The categorical metrics (POD and FAR) and the Correlation Coefficient are not presented for

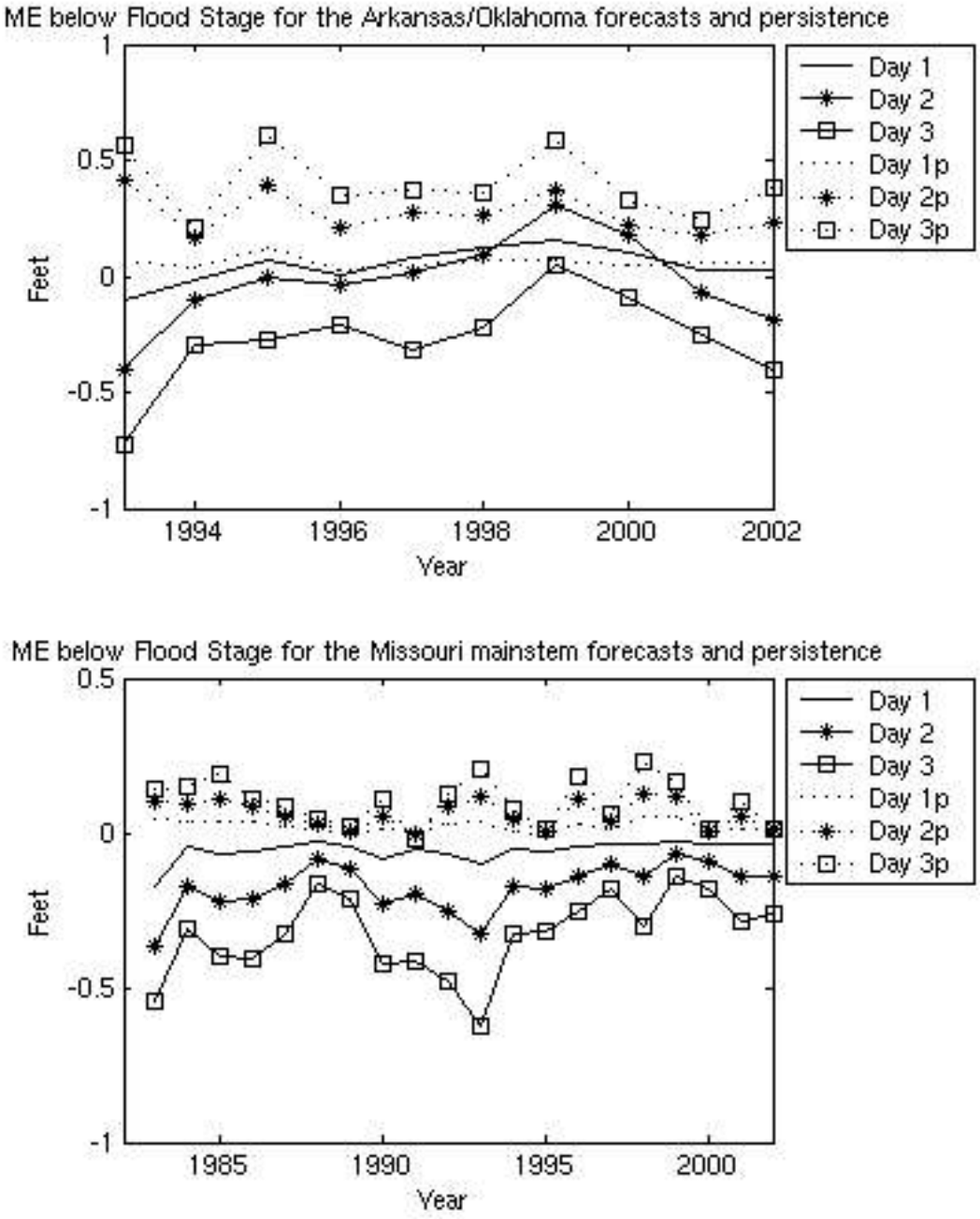

Figure 1: Annual ME for observations below Flood Stage on days 1, 2, 3.

the below Flood Stage category as they all vary only slightly near their optimal values and are therefore not very informative. The Mean Error (Figure 1) for both the A/O and the MM datasets is low (less than ∓0.5 feet). The issued forecasts under-forecast (negative ME) while the persistence over-forecasts (positive ME). The variation in the ME from year to year for the issued forecasts tends to follow the pattern of the persistence metrics although in the case of the MM dataset the sign is reversed. The RMSE (Figure 2) stays below two feet for the A/O dataset and except for the first year, below 1.5 feet for the MM dataset. For the A/O dataset, the day 3 issued forecasts show almost 1 foot less error than the day 2 persistence. For the MM dataset, the RMSE for the issued forecasts follows the persistence RMSE for the following day. The SS-RMSE (Figure 3) summarizes the relation of the persistence to the issued forecasts. The A/O dataset holds steady at 0.4 for all three days. The day 1 SS-RMSE for the MM dataset also varies around 0.4 with day 2 slightly less and day 3 a little lower. The large Sample Sizes (Figure 4) indicate small sampling errors, therefore, with positive skill scores and small ME and RMSE it seems reasonable to characterize these low stage forecasts as accurate and skillful.

3.3.2 Day One, Above Flood Stage Forecast Skill

For the above Flood Stage forecasts all six metrics are presented in Figures 5 to 11. The day 1 ME (Figure 5) is less than 1 foot for the A/O dataset and less than 0.5 feet for the MM dataset. For both datasets, the ME for the issued forecasts is less than the

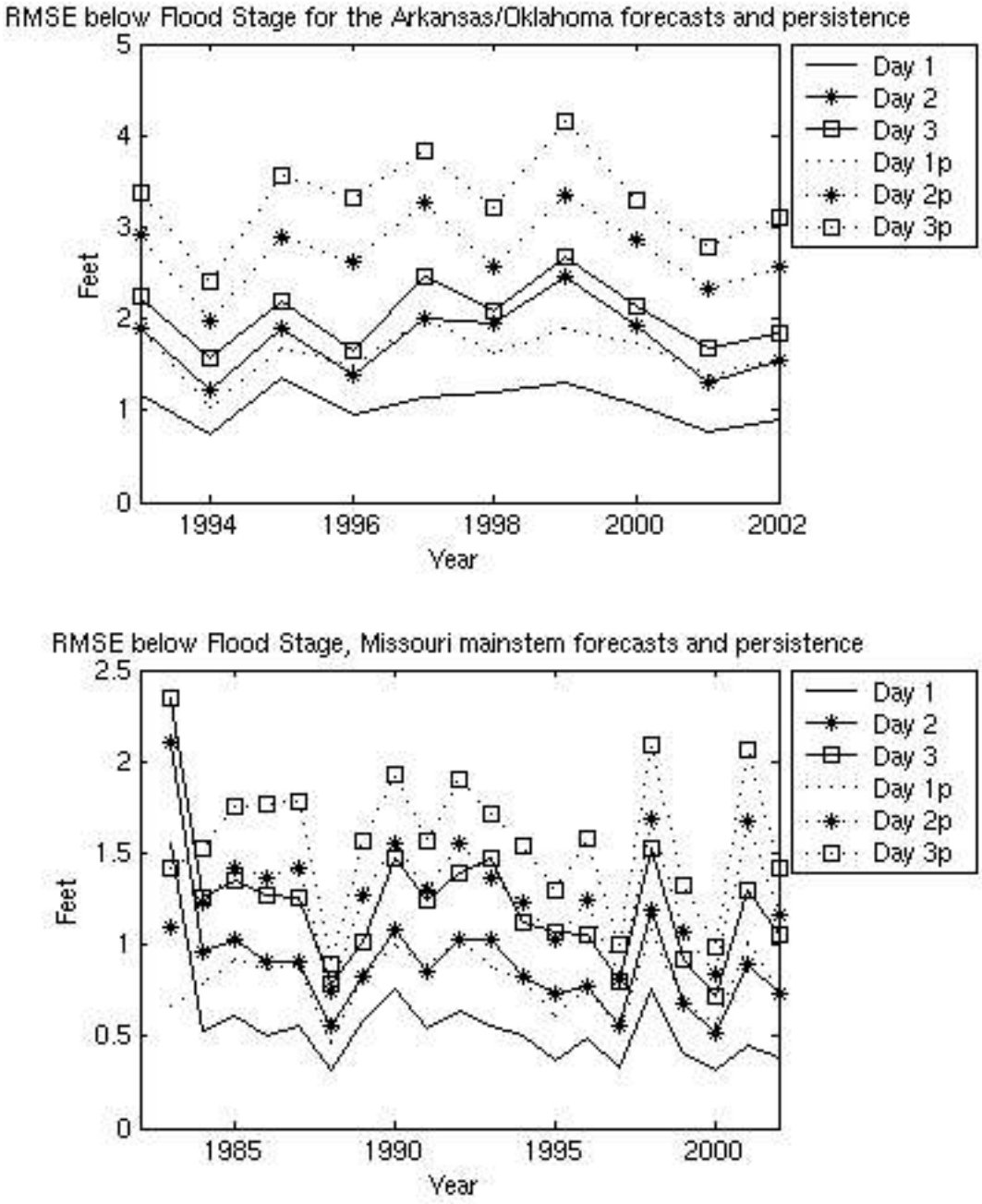

Figure 2: Annual RMSE for observations below Flood Stage on days 1, 2, 3.

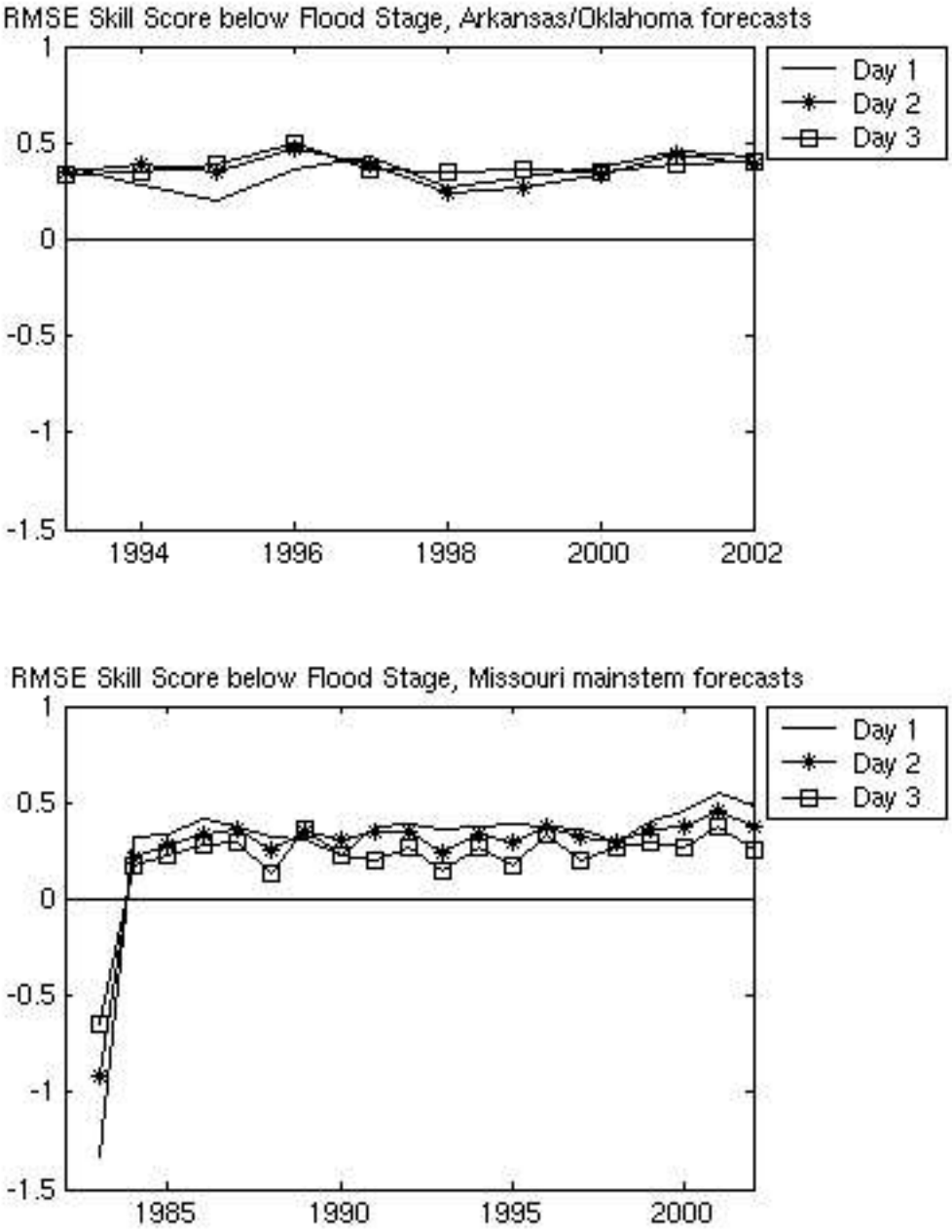

Figure 3: Annual RMSE Skill Score for observations below the Flood Stage on days 1, 2, 3.

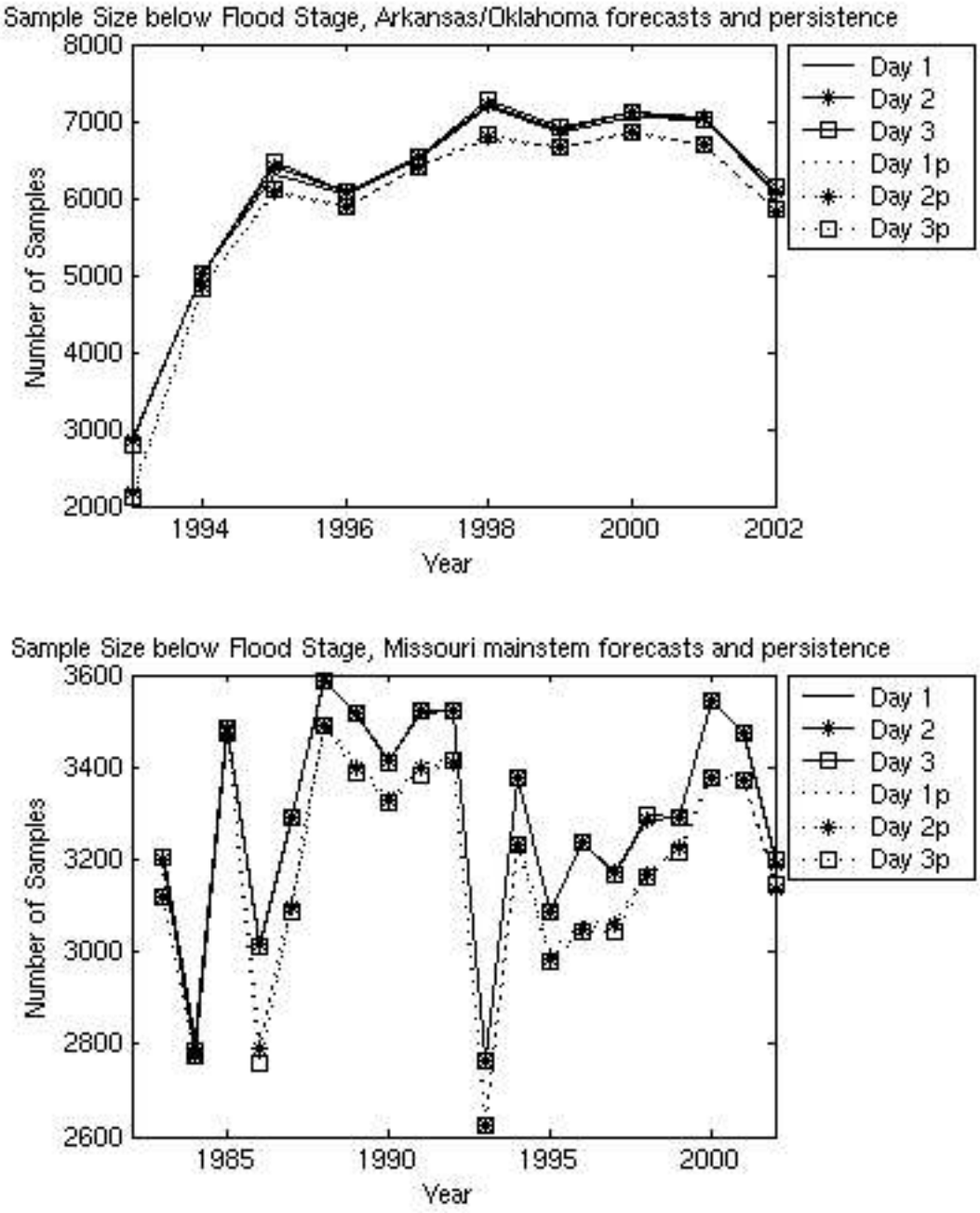

Figure 4: Annual Sample Size for observations below Flood Stage on days 1, 2, 3.

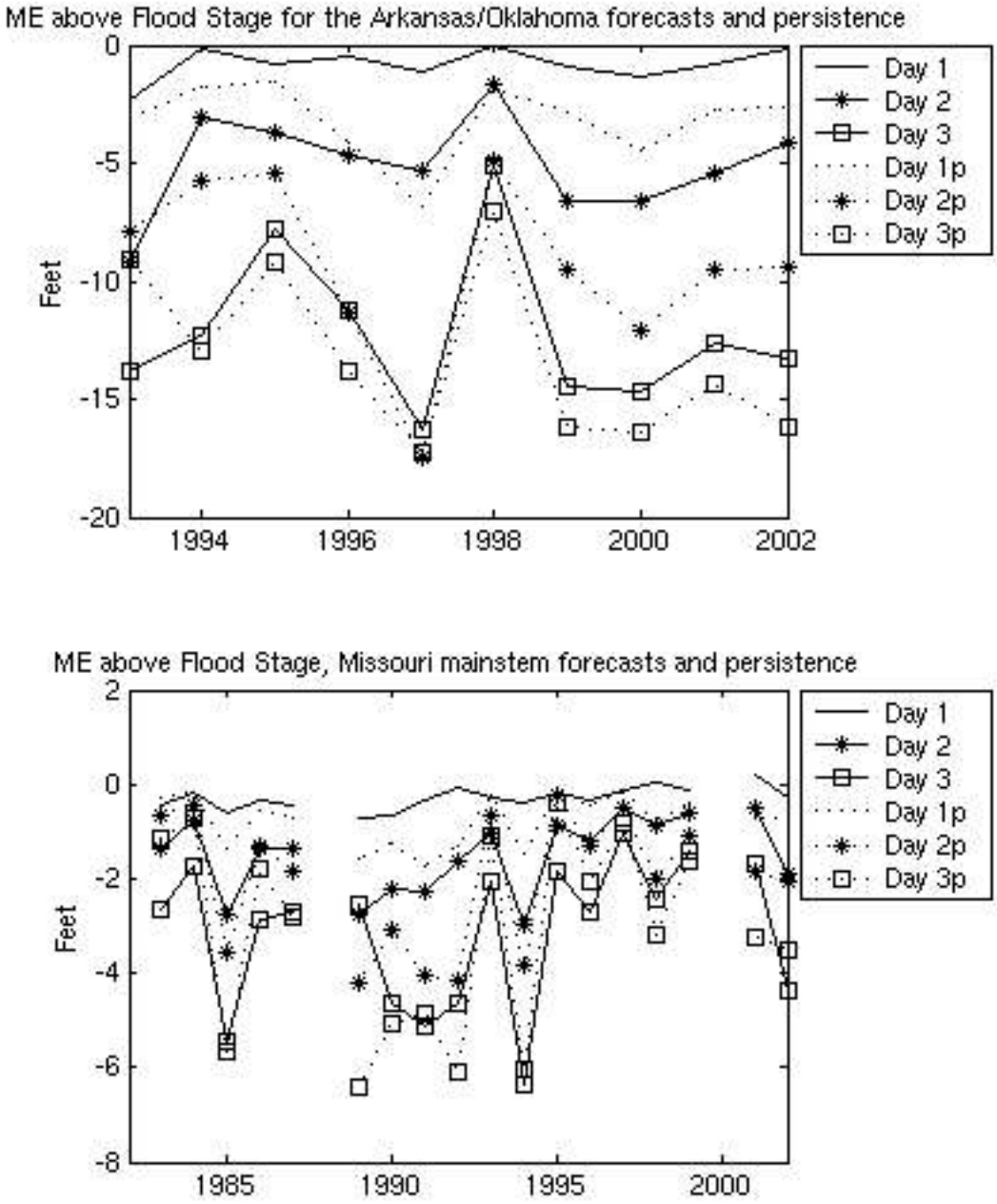

Figure 5: Annual ME for observations above Flood Stage on days 1, 2, 3.

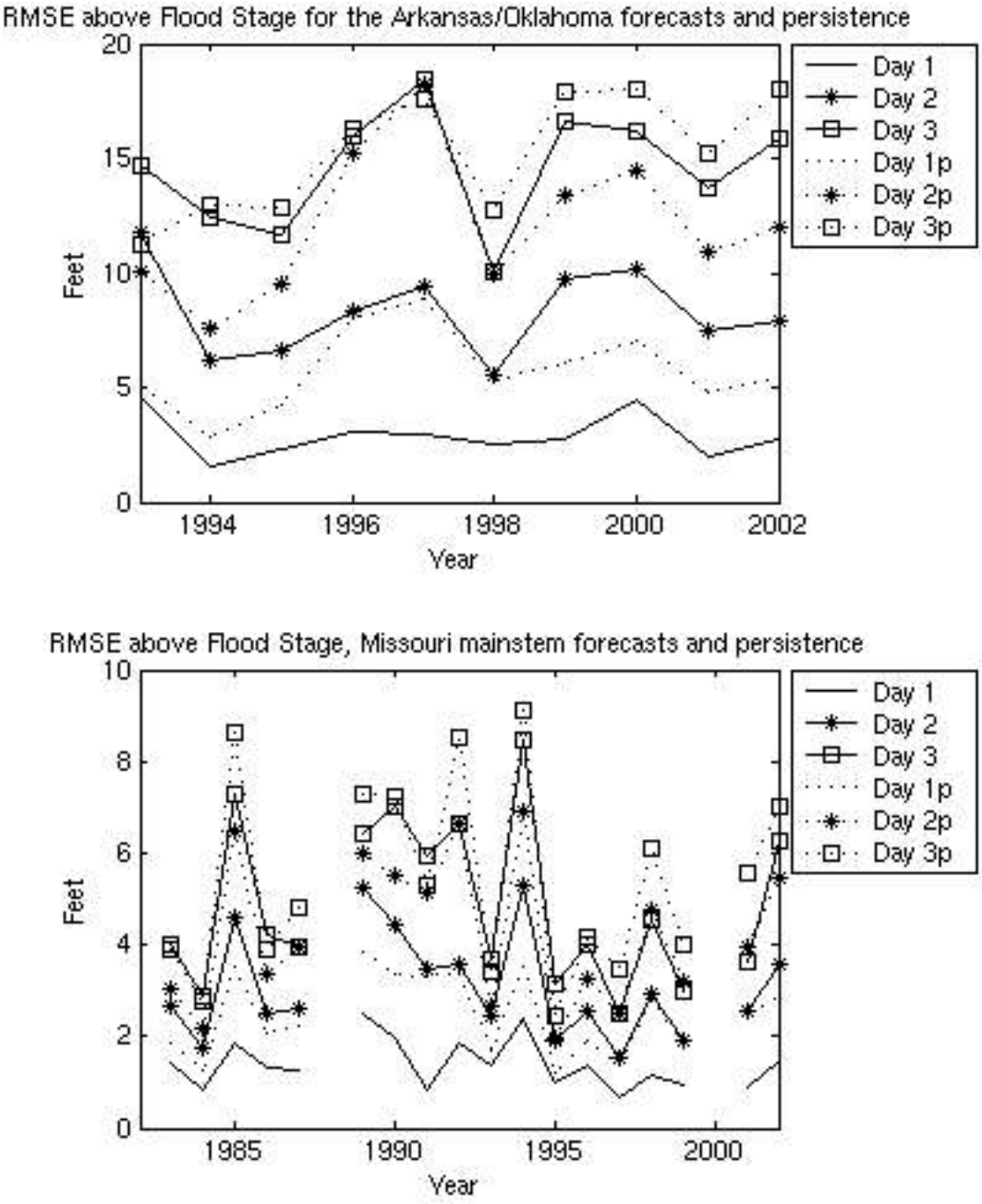

Figure 6: Annual RMSE for observations above Flood Stage on days 1, 2, 3.

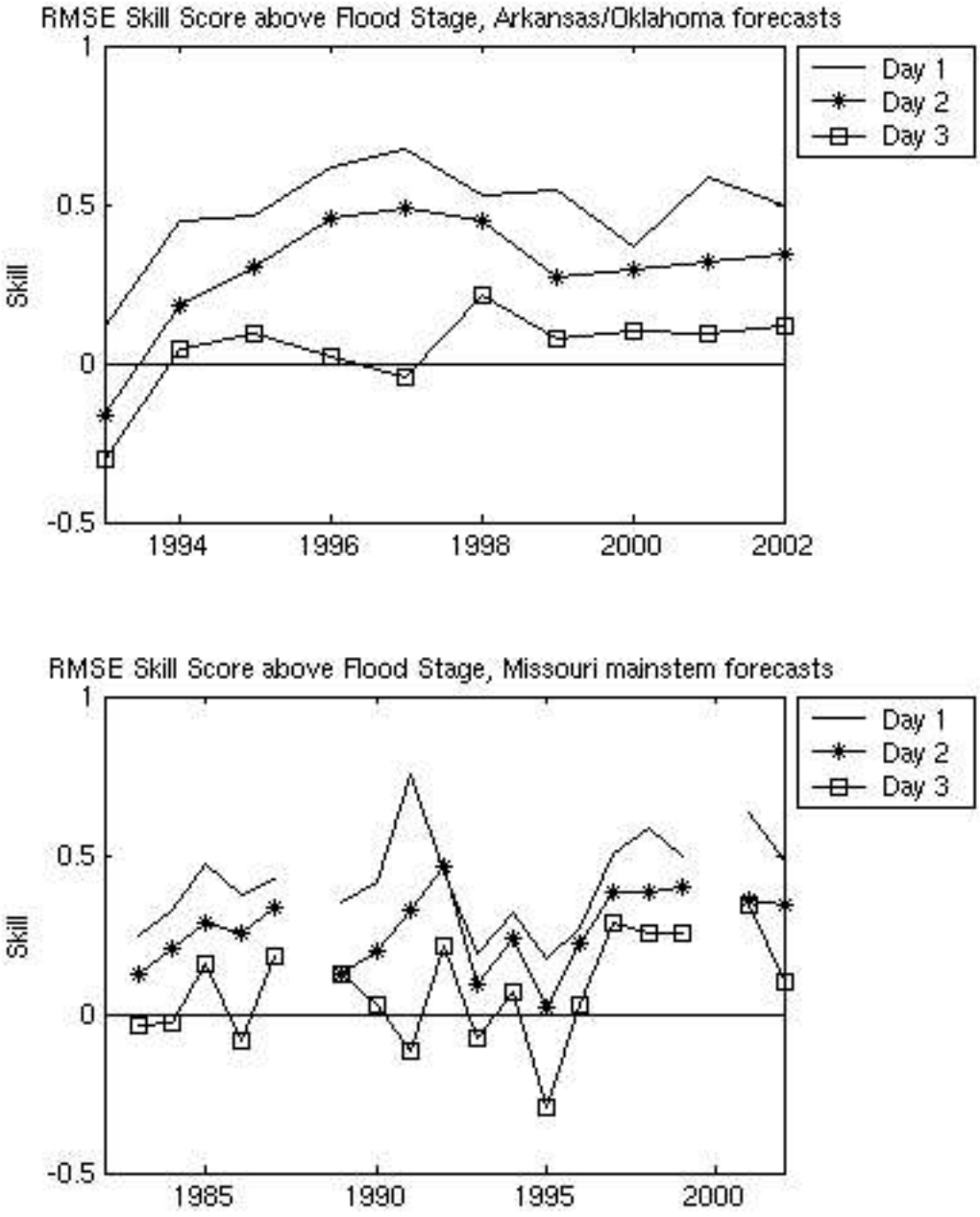

Figure 7: Annual RMSE Skill Score for observations above Flood Stage on days 1, 2, 3.

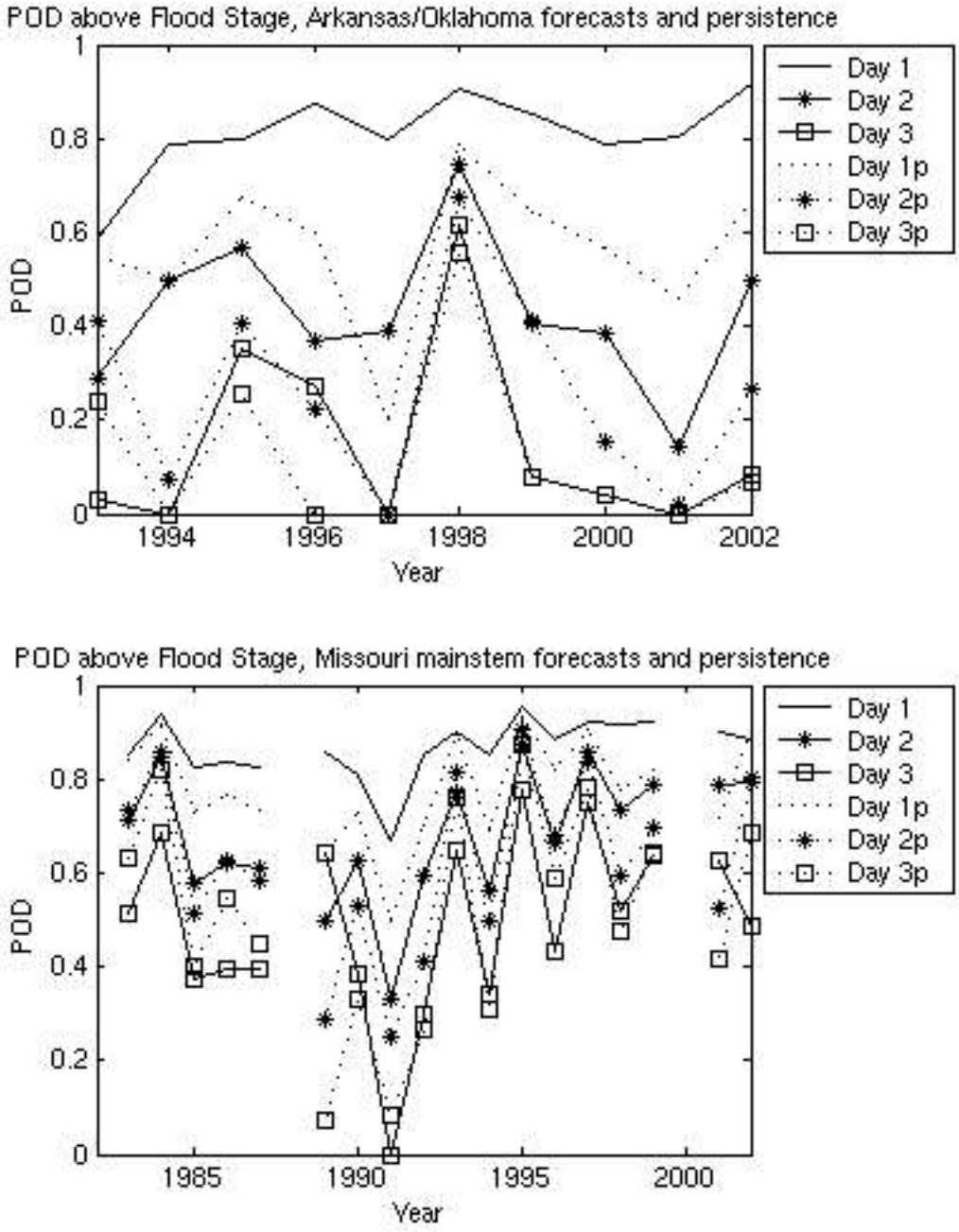

Figure 8: Annual POD for above Flood Stage category on days 1, 2, 3.

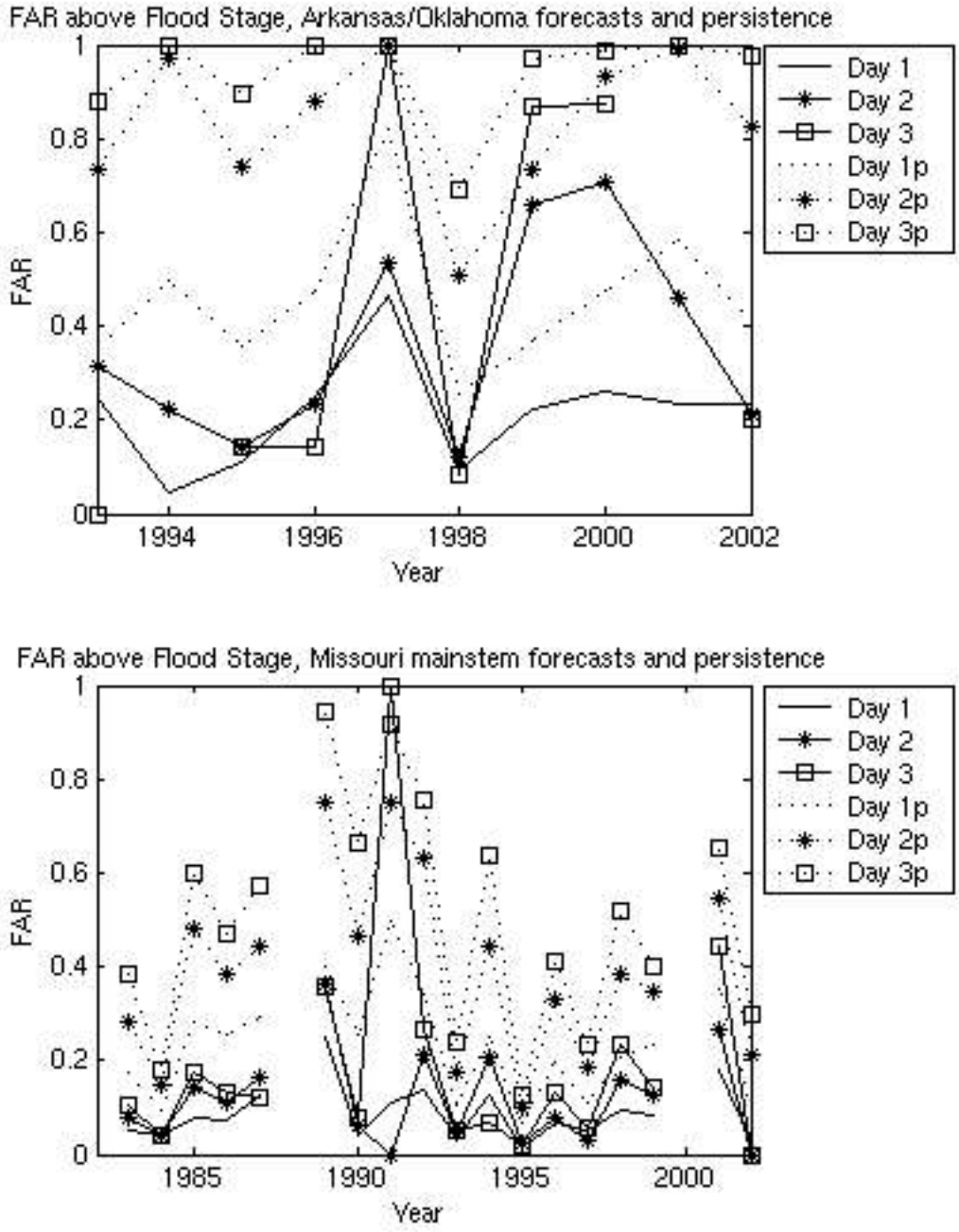

Figure 9: Annual FAR for above Flood Stage category on days 1, 2, 3.

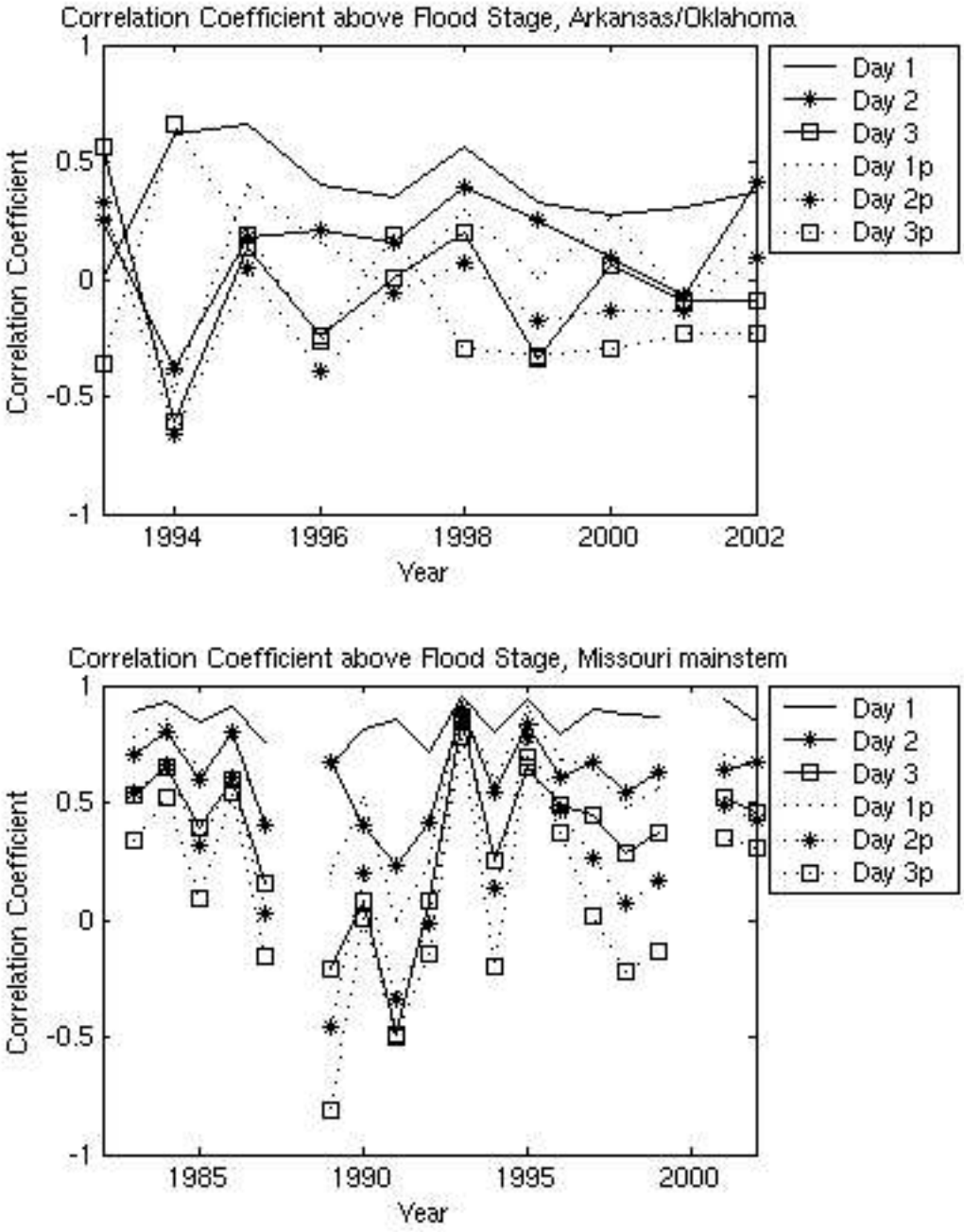

Figure 10: Annual Correlation Coefficient above Flood Stage on days 1, 2, 3.

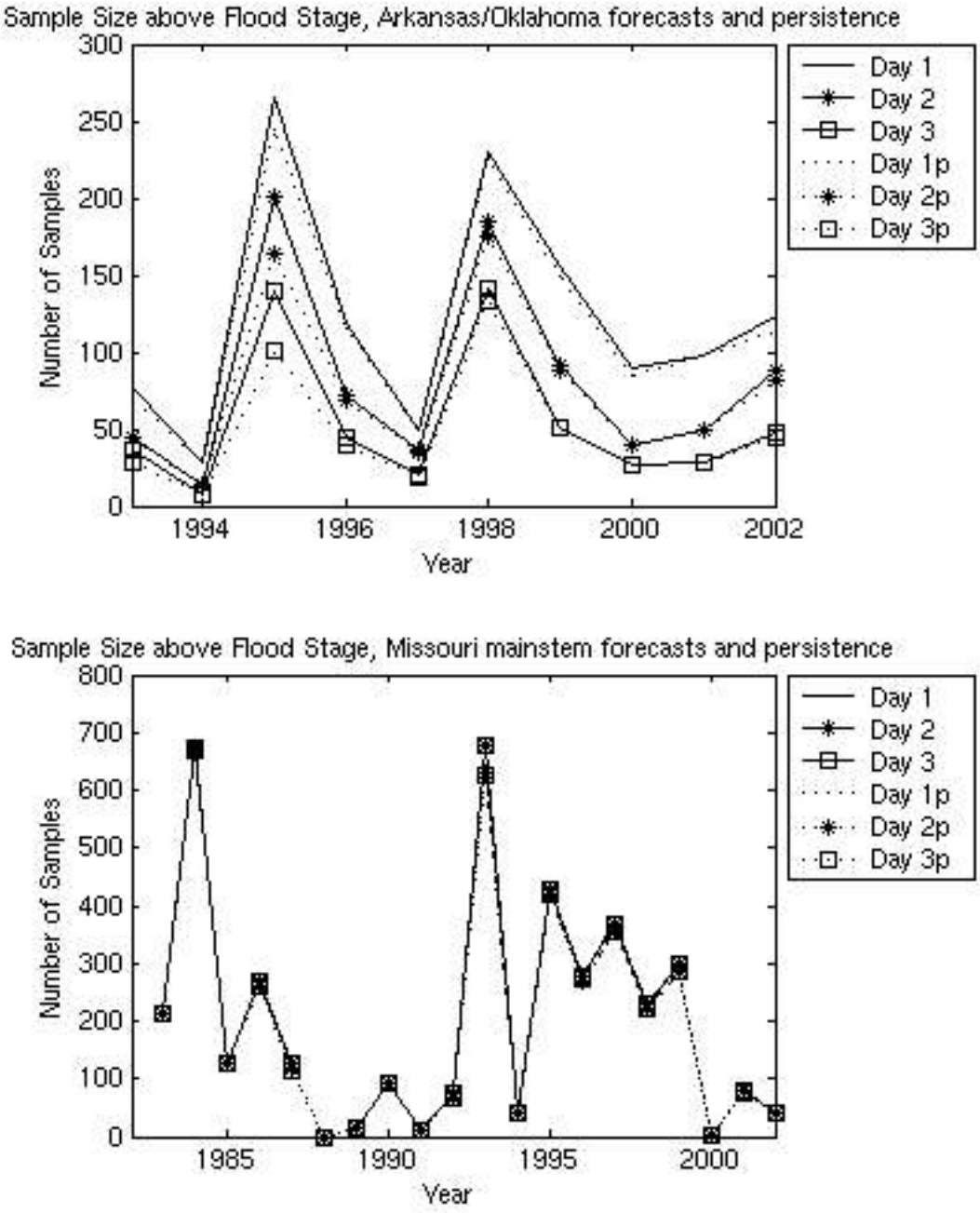

Figure 11: Annual Sample Size for observations above Flood Stage on days 1, 2, 3.

persistence ME. The breaks in the MM metrics in 1988 and 2000 indicate there were no observations above Flood Stage in those years. The RMSE (Figure 6) is similarly low for both datasets on day 1; it varies around 2 feet with two years jumping up to 5 feet for the A/O dataset and it is less than 1 foot for the MM dataset. The SS-RMSE (Figure 7) for the A/O dataset hovers around 0.5 while the MM dataset SS-RMSE rises as high as 0.7 but then falls as low as 0.2 for the day 1 forecasts. The POD (Figure 8) is consistently greater than 0.8 with the FAR (Figure 9) generally below 0.2 for both datasets. Both datasets are always at least as good as the persistence. In 1997 the A/O dataset FAR follows the pattern in the persistence and there is an increase to 0.4 in the FAR for the issued forecasts, but it drops back down in the following years. The CC (Figure 10) varies around 0.5 for the A/O dataset and 0.8 for the MM dataset. The CC for both datasets is consistently better than the persistence, although in 1993 and 1995 the MM dataset persistence CC rises up to 0.8. The Sample Sizes (Figure 11) for the above Flood Stage metrics are very small for some years. Therefore, the separation between the persistence and the actual forecasts is uncertain in those years. Nonetheless, it does not seem unreasonable to characterize these above Flood Stage, day 1 forecasts as accurate and skillful given the small absolute errors, the small number of false alarms, the high rate of detection and the positive skill scores.

3.3.3 Day Two and Three Above Flood Stage Forecast Skill

The day 2 and day 3 metrics are also plotted in Figures 5-11 alongside the metrics for the day 1 forecasts. As one would expect, the errors grow with each day, and by day 3 they have become similar in magnitude to those of the persistence baseline. The day 2 ME for the A/O dataset (Figure 5) ranges around -5 feet with the day 3 ME varying about -10 feet. The day 2 ME is well inside the day 2 persistence ME, but the day 3 ME is only slightly better than the persistence ME. The MM dataset, day 2 ME varies around -2 feet and the day 3 ME varies around -3 feet, but the issued forecasts show a larger ME than the persistence on some occasions even for the day 2 forecasts. The day 2 A/O RMSE (Figure 6) for the issued forecasts varies around 7 feet but the day 3 RMSE is on the order of 15 feet. The MM dataset, day 2 RMSE varies around 3 feet and for day 3 it covers a range between 9 and 3 feet. As with the ME, the RMSE for the issued forecasts are usually better than the persistence, though this is not always the case. The difference between the persistence and actual forecasts RMSE is well summarized by the SS-RMSE (Figure 7) which rises above 0.3 for day 2 of the A/O dataset but remains only slightly above 0.0 for day 3 of the A/O dataset. Again there is much more variability in the MM dataset than the A/O dataset, and the day 2 MM dataset SS-RMSE rises up to 0.5 and falls to zero, but it remains positive for the period of record. The day 3 SS-RMSE for the MM dataset however, switches between positive and negative skill score values. The day 2 and day 3 POD (Figure 8) follow the pattern in the persistence POD for both datasets. For the A/O dataset the issued forecasts have a better POD than persistence in day 2 in all

years except one, but by day 3 the A/O POD is less than or equal to the persistence for half the years. The MM dataset shows a similar pattern. The day 2 and day 3 FAR (Figure 9) are highly variable with breaks in the FAR record because no above Flood Stage forecasts were issued. The persistence FAR is large in day 2 and 3 and always larger than the FAR for the the issued forecasts. The day 2 and 3 CC (Figure10) is highly variable for both datasets; it falls below zero and climbs up above 0.5. According to the metrics reported here, the day 2 issued forecasts retain some of the accuracy and skill of the day 1 forecasts, but the day 3 forecasts provide little of the accuracy or skill seen in the day 1 forecasts.

Krzysztofowicz and Maranzano (2004) also found that above flood stage forecasts were not skillful by day 3 on a different set of basins using an entirely different method. In parameterizing their Bayesian Forecast System on four basins in the Monongahela River basin, they found the forecasts from the hydrologic models were “conditionally uninformative” for day 3 given a 24 hour precipitation forecast. They conclude the uncertainty is sufficiently great to limit the predictability of the river stages.

3.3.4 Comparison of the Two Datasets

The metrics computed for the issued MM forecasts and the issued A/O forecasts indicate the MM forecasts are better than the A/O forecasts. In the average, differences between the forecasts and the observations (see the RMSE, Figure 6) are less for the MM forecasts

with less consistent bias (see the ME, Figure 5) for the above Flood Stage forecasts, a stronger correlation (see the CC, Figure 10) between the forecasts and observations, and a greater likelihood of being warned of a flood (see the POD, Figure 8) with fewer false alarms (see the FAR, Figure 9). However, the same can also be said of the baseline persistence; the MM persistence is better than the A/O persistence. A comparison between the two data sets of the improvement over the persistence baseline (SS-RMSE, Figure7) shows a different picture; the A/O forecasts show at least as much skill over persistence as the MM forecasts. Therefore, although the metrics indicate the MM forecasts are more accurate than the A/O forecasts, they also indicate the NWS forecast process integrates more skill into the A/O forecasts than into the MM forecasts.

3.3.5 Changes in Skill Over Time

One expectation of these forecasts is the skill will have improved over the past ten and twenty years as a result of the enhancements made to the forecast process. However, this expectation does not appear to be met in many of these metrics. Through visual inspection a trend may be seen in the day 1 POD for the A/O dataset above Flood Stage category (Figure 8) and in the SS-RMSE for the day 1 MM below flood stage forecasts (Figure 3). The other metrics do not appear to have a trend in either direction. This is a surprising result, given the number of updates made to the forecast system.

3.4 Discussion of the Administrative Verification

Hydrologists have been developing the forecast process under the assumption that integrating improved science and computational methods will lead to better forecasts. Decisions regarding forecast process improvements have been based upon the experience of hydrologic science and forecasting experts. From the data for the limited set of forecast locations presented here, it appears this approach has not worked and a new approach is required.

In considering alternatives, the approach used by the meteorological forecast community serves as a good example. For instance, the development of Numerical Weather Prediction models is conducted by numerous, un-affiliated groups following different approaches with the results compared through objective measures of forecast performance. In other words, the forecasts are verified, and the research is driven, not by ad-hoc opinions postulated by subject matter experts, but by the actual performance of the forecasts as determined with objective measures. This is the approach hydrologists need to adopt; hydrologic forecasts need to be embedded in a comprehensive verification program.

In the remainder of this section, three high priority research activities to initiate the development of a comprehensive hydrologic forecast verification program are proposed.

They are:

- Define verification standards for hydrologic forecasts;
- Develop a comprehensive baseline description of hydrologic forecast skill;
- Describe the sources and sinks of forecast skill, and the interaction between those sources and sinks in the hydrologic forecast process.

3.4.1 Verification standards

Standards are the foundation for comprehensive verification. Carefully chosen standards can be a means of provoking a structured discussion to define important forecast characteristics and the best techniques to evaluate those characteristics, without being a limitation to "creativity". Well constructed standards provide a number of benefits to those working to enhance forecast performance. They support the second and third research goals by describing methods for characterizing forecast skill. They enhance the communication amongst those working to improve the forecast skill by defining a common language of verification. They help forecast agencies implement operational, verification procedures by providing a template for those procedures. Standards are only effective if they are widely accepted, and therefore, they must be proposed through publication. Once hydrologists have determined how to evaluate forecasts, the task of evaluating them can begin.

3.4.2 A baseline description of forecast skill

Developing a comprehensive baseline description of hydrologic forecast skill is the first step toward improving the forecasts. We must know where "we are," to determine where "we should go". The primary obstacle to such studies is data collection. There are two archive sources for past forecasts. First, archives at local NWS forecast offices and at the offices of other agencies and groups which issue forecasts. The NWS archives extend back to at least April 1, 2001 for the National verification program locations and possibly longer and for more locations depending upon the practices at the local NWS offices. A second source for NWS forecasts, is the Service Records Retention System (SRRS) at the National Climatic Data Center. It holds records for the NWS issued forecasts, including the RFC time series. Observations are also required for verification. If the forecasts are pulled from the NWS RFC archives, the observations will be available there as well. If the forecasts are pulled from the SRRS, observations may have to be collected from the agency which manages the gauge location or from the SRRS as well. Once the forecasts and observations have been collected, an in depth review of the forecast accuracy and skill can be conducted and published.

3.4.3 Identifying sources and sinks of forecast skill

The third priority item, describing the sources and sinks of forecast skill, and the interaction between those sources and sinks in the hydrologic forecast process is the key to understanding how to improve the forecasts. If the interaction between the forecast

process elements, including the forecasts input to the hydrologic models, is not well understood, then it will not be possible to identify how changes in one element in the forecast process will change the skill of the final forecasts. Monitoring the forecast process with a well structured verification process which includes control forecasts, as well as verification of all input forecasts can provide some information to analyze the sources of skill in the hydrologic forecasts. However, most of the required analyses will have to be done using hindcast methods. Krzysztofowicz has begun this work in his development of a Bayesian Forecast System (Krzysztofowicz, 1999b, Krzysztofowicz and Herr, 2001, Krzysztofowicz and Maranzano, 2004), and the next section of this thesis makes an additional contribution in the analysis of the sources of skill on precipitation driven headwater basins. Additional studies on snow covered basins, on non-headwater, downstream basins, on basins with reservoirs, on basins with dry versus wet climates and on basins with a variety of soil types also need to be conducted.

The following two sections initiate the formal development of verification methods for hydrologic forecasts by demonstrating a method for analysing the forecasts to determine sources of error and then proposing verification standards for short term deterministic hydrologic forecasts.

4 SCIENTIFIC VERIFICATION OF DETERMINISTIC RIVER STAGE FORECASTS

This section describes a method for determining the sources of skill and error in a set of forecasts. Many people have studied elements of the forecast process: for example, model calibration, model state updating, and precipitation forecasting, but the forecast process itself, with the various elements linked together, has not been studied extensively. This section presents a hindcasting experiment used to analyze NWS, short-term (lead-times less than 3 days), single-valued, river stage forecasts on precipitation driven, headwater basins. The purpose of this experiment is to illustrate a method for using the distributions oriented verification of Murphy and Winkler (1987) to identify sources and sinks of forecast skill. This analysis leads to several recommendations for operational verification systems.

Verification metrics and methods (taken from the meteorological literature) are applied to the hindcasts to address the following set of questions.

- What is the primary source of skill in the hindcasts at each lead time?
- What is the role of the calibration, the initial conditions, and the QPF in the hindcast skill?
- How does the quality of the calibration and the initial conditions affect the total hindcast error given the uncertainty in the QPF?
- What are the key requirements for an operational verification system if it is to provide insight into sources of error and skill in the forecasts?

Numerous similar studies on downstream forecast locations, snow covered forecast locations, reservoir outflow locations and the like will be required to build a robust understanding of hydrologic forecast skill and uncertainty.

4.1 Error and skill in hydrologic forecasts

Traditionally, the error in hydrologic forecasts has been divided into two categories: meteorological error and hydrologic error. Meteorological error is the error in the hydrologic forecasts caused by the error in the meteorological forecasts used to drive the hydrologic models into the future. This study focuses on the meteorological error resulting from Quantitative Precipitation Forecasts (QPFs). QPFs are single-valued precipitation forecasts, reported as depth of rain expected to fall over a basin in a given time. While the QPFs have improved over the past decades, they remain highly uncertain when evaluated at the short modeling time steps and the fine spatial scales used for hydrologic models, even if those models run at six hour time steps over lumped basins, hundreds of square miles in area. Temperature forecasts can be critical to short-term forecasts on basins where the precipitation type, rain or snow, determines if a flood event will or will not occur. However, the basins to be studied here are never snow covered; consequently, QPFs are the only meteorological forecasts considered in this analysis.

Hydrologic error consists of the errors caused by the hydrologic modeling. Within this broad category there are many contributing sources of uncertainty: model parameters,

model initial conditions, upstream flows routed into a basin, reservoir operations, rating curves, and the structure of the models. This study focuses on the hydrologic error for a single headwater basin and therefore does not include the error in the upstream routed flows or reservoir operations. All of these errors are interrelated, and errors of one type may exaggerate or mask errors of another type. As will be seen in the hindcasts, the interaction between the types of error changes with lead-time and is an important element in understanding the sources of error and skill in the hindcasts.

4.2 Hindcast experiments

With the growth of inexpensive computing power and disc space, hindcasting is becoming a more usable tool for analyzing forecasts. The experimenter sets up a system to recompute forecasts and based upon the prior observations and input forecasts, makes controlled runs to re-forecast a set of events. The forecast model is run with observed precipitation up to a date marked as the "present;" the initial conditions for the model system are stored and the model is then restarted with forecast precipitation. The forecast is computed and stored, and the model is run forwards with observed precipitation to a new date marked as the "present." In order for the hindcasts to be valid, it is critical no information (like the observed stage) is used in the calculations during the "forecast" period. The computational process, the input observations, and the input forecasts can be manipulated to evaluate alternate forecast procedures, or the probable effects of improved inputs upon the forecasts. Comparisons of alternate scenarios are facilitated because the

same climatic period is used for all computations, thereby eliminating the differences in forecast skill due to annual variability in the local climate.

A few previous authors have used hindcasts to analyze hydrologic forecasts. Krzysztofowicz has used hindcasts to parameterize the Bayesian Forecast System (BFS) (Krzysztofowicz and Herr, 2001; Krzysztofowicz and Maranzano, 2004) which integrates the hydrologic and the meteorological uncertainty into a single probability forecast. They used hindcasts computed at the NWS Ohio River Forecast Center with perfect QPF (observations) for the first 24 hours and with zero for the following 48 hours. Franz et al (2003) used hindcasts to evaluate the skill of long-range ensemble water supply forecasts. They recomputed initial conditions for past years and then generated hindcasts with the NWS Ensemble Streamflow Prediction system from these re-constructed initial conditions. Recently, Vivoni et al (2003) used hindcasts to evaluate the skill of their QPF Nowcaster for very short lead times for one storm event. They generated precipitation forecasts with their Nowcaster, and then ran these generated precipitation forecasts through a hydrologic model. They compared their results to those using a persistence precipitation forecast. Werner et al (2004) used hindcasts to evaluate several methods of computing temperature ensembles for use in mid- to long-range hydrologic forecasts. Like the Franz et al study, they have reconstructed initial conditions for past years and they are comparing seasonal volume hindcasts using their

various temperature ensembles. While the process of hindcasting has not been used extensively in hydrology yet, hindcasting can be an effective analytic tool.

4.3 Diagnostic Verification

The verification method demonstrated here follows the diagnostic approach of Murphy and Winkler (1987). When this diagnostic approach is applied, the forecasts are sorted into discrete subsets and then each subset is evaluated. For example, when sorting stage forecasts into two categories as is done in this analysis, the distributions to be evaluated when assessing *Discrimination* skill are $p((f, o) \mid o < T)$ for the low stage category and $p((f, o) \mid o \geq T)$ for the high stage category, where T is a stage threshold (e.g. Flood Stage). To assess the *Reliability* of the forecasts, the forecast-observation pairs are sub-setted based upon the forecast value. The distributions to be assessed are then $p((f, o) \mid f < T)$ for the low stage category and $p((f, o) \mid f \geq T)$ for the high stage category.

The terms *Discrimination* and *Reliability* are also used to describe probability forecasts, with *Discrimination* diagrams used to assess the Resolution of the forecasts and *Reliability* referring to the quality of the probability statements. In addition, the term *Discrimination* is associated with the measure proposed by Murphy, Brown and Chen (1989) labeled DIS. In this description, *Discrimination* refers to the skill of the forecasts when measured for subsets sorted by the observations, and *Reliability* refers to the skill of the forecasts when measured for subsets sorted by the forecasts. Although the

calculations are not identical, the underlying concepts are the same amongst all the references associated with *Discrimination* and *Reliability.*

4.4 Method for the hindcast experiment

The process for computing these hindcasts follows the NWS forecast process to the extent possible. The algorithms used are the NWS standards, and the input data was collected from NWS offices. The places where the hindcast process deviates from the NWS process are noted.

4.4.1 Algorithms used to compute the hindcasts

The forecast process to be analyzed here is the NWS short term deterministic river stage forecast process. This process was described in detail in the previous section, and is reviewed here for convenience. For precipitation driven headwater basins, the NWS uses a calibrated Sacramento model (Burnash et al., 1973) at six hour time-steps to compute runoff from rainfall, a unit hydrograph to route runoff to the basin outlet (Linsley et al, 1975), and then manual state updating to assimilate observed stages into the simulations. Precipitation forecasts are used for all lead times, although modeled precipitation is only used in the first 24 hours and zero is used after 24 hours. The hydrologic model output is post processed using a simple linear difference scheme (NWS, 2002a) to remove current model biases.[2] With this post processing scheme, the simulation is forced to pass through

2 At time step *i*, the adjusted forecast, $f_i^{adj} = f_i + (O_0 - f_0)\,(N\text{-}i)/N$, where f_i is the un-adjusted simulation at time *i*, O_0 and f_0 are the observation and un-adjusted simulation at time 0, and N is a manually selected number of time steps over which the adjustment is computed.

the observations thereby insuring consistency between the forecasts and recent observations. The post-processed time series is then converted to a stage time series with a rating curve and the stage time series is issued as the forecast. The components of the forecast process to be analyzed here are the calibration of the Sacramento model, the model state updating as it is reflected in the model initial conditions, and the QPF.

The forecast process cannot be reproduced exactly in the hindcast process. Most obviously, the manual state updating cannot be recreated because it would be too expensive and non-objective. The variational assimilation method (VAR) proposed by Seo et al. (2003) is used for the hindcasts. In general, the forecasters are able to integrate more information through the manual state updating process than can be done automatically and this ability can be important for basins with complex hydrology: for example, basins which include snow, routing upstream flows through the basin, or reservoir operations, as well as runoff computations. However, as was demonstrated by Seo et al (2003), on the precipitation driven headwaters studied here, the automated state updating can be effective. A second difference between the operational forecast process and the hindcast process is the simulation post-processing is not used in the hindcasts because it obscures the differences between the hindcast scenarios. The post-processing algorithm forces the simulations to run through the last observed value, therefore, if the post-processing was included, all the hindcasts would start at the same value and the only differences between them would be those discernible at the longer lead times. The third

difference between the actual forecast operations and the hindcasts is the forecast issuance time. The actual forecasts are issued once daily, at 12 GMT, unless flooding is imminent, in which case the forecasts are issued on an as needed basis. The hindcasts are "issued" twice daily, at 00 GMT and 12 GMT, and the schedule is not changed even if there is flooding.

4.4.2 Description of the data

One big obstacle to effective hindcasting is data archiving. Without a proper archive of the input to the original forecasts, they cannot be recreated. (Inexpensive computing resources are facilitating this aspect of the hindcast process as well.) Three basins for which there was a suitable archive of the input data were found: two basins in Oklahoma, the Illinois River at Watts, OK, and the Blue River at Blue, OK and one in Missouri, the Elk River at Tiff City, MO. These basins have been used in the Distributed Model Inter-comparison Project (Smith et al., 2004) and in testing the VAR (Seo et al., 2003). A detailed description of the basin geo-hydrology can be found in Smith et al. (2004). The observed precipitation for the hindcasts was taken from the NWS Stage III grids computed using the P1 process (Young, 2000) at the Arkansas-Red Basin River Forecast Center (ABRFC). The QPF was also provided by the ABRFC from their archive of operational QPFs. The river stage data is the operational stage data collected by the USGS and archived by the ABRFC. There was sufficient data for these basins to run in a hindcast mode for four years from 1997 to 2000.

4.4.3 Description of the hindcast scenarios

Three forecast process elements were studied here: calibration, state updating and QPF. For each forecast process element a “skilled” and an “un-skilled” implementation was developed. For the calibration, a calibrated and an un-calibrated set of parameters were used for the “skilled” and the “un-skilled” implementation. The calibrated parameters were derived by NWS experts within the NWS Hydrology Laboratory using manual calibration methods described in the NWS calibration handbook (NWS 2002b). The un-calibrated model parameters are derived from the pedological equations of Koren (Koren et al, 2003) with no additional manual calibration performed on the pedological results. Parameters derived from these pedological equations are commonly used as an initial parameter set to begin the manual calibration process. These pedological parameters are referred to as the un-calibrated or a-priori parameters.

The “skilled” and “un-skilled” state-updating was computed by running the hindcasts with the VAR turned on for the “skilled” implementation and off for the “un-skilled” implementation. Three QPF implementations were used: a “skilled”, an “un-skilled” and a “perfect” implementation. For the “un-skilled” implementation, the QPF is set to zero for the entire forecast period; this is called the Zero QPF scenario. For the “skilled” implementation, the operational meteorological modeled QPF is used for the first 24 hours then the QPF is set to zero for the remaining two days of the hindcast period; this is

called the Real QPF scenario. For the “perfect” implementation, the observed precipitation is used as the QPF; this is called the Perfect QPF scenario. The first two QPF implementations are commonly used in the NWS operational forecast process, and the third is commonly used for hydrologic model development, model uncertainty and calibration studies. The calibration, the state updating and the QPF types are matched for a total of twelve hindcast scenarios on each basin. Table 4 lists each hindcast scenario. Persistence hindcasts were also generated and are used to provide a perspective on the skill of the hindcasts. Persistence is defined as the observation at the basis-time of the forecast. That is, the observation at the time the forecast is issued is persisted into the future as the forecast for all lead-times.

4.4.4 Hindcast analysis process

The forecasts and observations are sorted into two subsets, high and low stages. It is possible to sort into finer categories, and when this is done the characterization of the hindcast–observed relation is similar to that for the two category sorting. In addition to sorting by stage height, the hindcasts are sorted into lead-times at six hour time-steps (the time-step of the hindcasts). Statistics for each subset are computed on the forecast observation pairs collected from all three basins, and then these statistics are compared to isolate the changes in skill provided by each forecast process element.

Scenario	Abbreviation
Perfect QPF with VAR and calibrated parameters	P-V-C
Perfect QPF without VAR and calibrated parameters	P-NV-C
Real QPF with VAR and calibrated parameters	R-V-C
Real QPF without VAR and calibrated parameters	R-NV-C
Zero QPF with VAR and calibrated parameters	Z-V-C
Zero QPF without VAR and calibrated parameters	Z-NV-C
Perfect QPF with VAR and un-calibrated parameters	P-V-U
Perfect QPF without VAR and un-calibrated parameters	P-NV-U
Real QPF with VAR and un-calibrated parameters	R-V-U
Real QPF without VAR and un-calibrated parameters	R-NV-U
Zero QPF with VAR and un-calibrated parameters	Z-V-U
Zero QPF without VAR and un-calibrated parameters	Z-NV-U

Table 4. The names of the hindcast scenarios.

To compare the different subsets of the hindcasts, the subsets must be characterized using summary statistics. Several commonly used verification statistics were investigated: the Mean Absolute Error, The Root Mean Square Error, the Mean Error, the False Alarm Ratio, the Probability of Detection, the Critical Success Index, the Area under the Relative Operating Characteristics (ROC) curve, a ROC Discrimination distance, and the Pearson Correlation Coefficient. It was found that the measures themselves were not the key to understanding the error in the hindcasts, rather the comparisons amongst the hindcasts and subsets made the verification meaningful. Therefore, the Root Mean Square Error (RMSE) is used in the presentation of the hindcast comparisons. The Sample Sizes are reported to provide an indication of the uncertainty in the computed metrics.

For each hindcast scenario, including the persistence hindcasts, for each lead-time, the RMSE is computed for the high stage and low stage *Reliability* and *Discrimination* subsets across all three locations. For each forecast process element, the scenarios which are similar except for the forecast process element of interest are compared. For example, to evaluate the contribution to the hindcast skill from the calibration, the hindcasts with the “skilled” and the “un-skilled” calibration but the same QPF and updating treatments are compared. The same is done to isolate the contribution of the initial conditions to the hindcast skill, the hindcasts with “skilled” and “un-skilled” updating but the same QPF and calibration treatments are compared. For the QPF the same procedure is followed:

the state updating and the calibration are held constant, and the different QPF scenarios are compared. A list of the comparisons are provided in Tables 5, 6 and 7.

4.5 Results of the Hindcast Experiment

The model calibrations, the QPF skill, and the general performance of the hindcasts in comparison to the persistence are described first as background information to explain the hindcast comparisons. The comparisons are then described, first the calibration, then the initial conditions and finally the QPF.

4.5.1 Description and comparison of the two calibrations

In order to provide some background to the hindcast behavior, the two calibrations are described and compared. It is customary within the NWS to use the ME to evaluate a calibration; it is the only recommended statistic in the NWS handbook for calibration (NWS, 2002b). Therefore, the ME is reported in addition to the RMSE. The Correlation Coefficient (R) is also reported as it is another commonly used calibration metric. The Perfect QPF hindcast with no state updating (P-NV) is the same as a standard calibration simulation as there is no state updating and observed precipitation is used to drive the models. The statistics computed from this hindcast scenario for the calibrated and un-calibrated model are summarized in Table 8.

Scenarios Compared	Abbreviation
Calibrated vs. A-priori parameters for Perfect QPF with VAR	P-V
Calibrated vs. A-priori parameters for Perfect QPF, no VAR	P-NV
Calibrated vs. A-priori parameters for Real QPF with VAR	R-V
Calibrated vs. A-priori parameters for Real QPF, no VAR	R-NV
Calibrated vs. A-priori parameters for Zero QPF with VAR	Z-V
Calibrated vs. A-priori parameters for Zero QPF, no VAR	Z-NV

Table 5. Scenarios for the calibration comparisons.

Scenarios Compared	Abbreviation
With and without VAR for Perfect QPF with calibrated parameters	P-C
With and without VAR for Perfect QPF, un-calibrated parameters	P-U
With and without VAR for Real QPF with calibrated parameters	R-C
With and without VAR for Real QPF, un-calibrated parameters	R-U
With and without VAR for Zero QPF with calibrated parameters	Z-C
With and without VAR for Zero QPF, un-calibrated parameters	Z-U

Table 6. Scenarios for the state updating comparisons

Scenarios compared	Abbreviation
Real QPF vs. Perfect QPF, with VAR, Calibrated Parameters	rvc-pvc
Real QPF vs. Perfect QPF, no VAR, Calibrated Parameters	rnvc-pnvc
ZeroQPF vs. Perfect QPF, with VAR, Calibrated Parameters	zvc-pvc
Zero QPF vs. Perfect QPF, no VAR, Calibrated Parameters	znvc-pnvc
Zero QPF vs. Real QPF, with VAR, Calibrated Parameters	zvc-rvc
Zero QPF vs. Real QPF, no VAR, Calibrated Parameters	znvc-rnvc
Real QPF vs. Perfect QPF, with VAR, Un-calibrated Parameters	rvu-pvu
Real QPF vs. Perfect QPF, no VAR, Un-Calibrated Parameters	rnvu-pnvu
Zero QPF vs. Perfect QPF, with VAR, Un-calibrated Parameters	zvu-pvu
Zero QPF vs. Perfect QPF, no VAR, Un-Calibrated Parameters	znvu-pnvu
Zero QPF vs. Real QPF, with VAR, Un-calibrated Parameters	zvu-rvu
Zero QPF vs. Real QPF, no VAR, Un-Calibrated Parameters	znvu-rnvu

Table 7. Scenarios for the QPF comparisons.

	Discrimination			*Reliability*		
	RMSE	ME	R	RMSE	ME	R
Low: Calibrated	1.3	0.2	0.85	1.3	0.2	0.85
Low: Uncalibrated	3.5	-0.1	0.51	2.6	-0.6	0.50
High: Calibrated	2.8	0.0	0.65	3.0	0.0	0.75
High: Uncalibrated	5.8	3.2	0.55	11.5	10.5	0.35

Table 8. Summary statistics to compare the model calibrations.

For the low stage *Discrimination* and *Reliability*, both the calibrated and the un-calibrated parameters have almost no ME. The un-calibrated low stage *Discrimination* RMSE (3.5 ft.) however, is more than twice the calibrated RMSE (1.3 ft.) and the un-calibrated Correlation (0.51) is only 60% of the calibrated Correlation (0.85). The low stage *Reliability* metrics show similar differences. For the high stages, for both *Discrimination* and *Reliability*, the expert calibration has almost no ME, a high Correlation (0.65 for *Discrimination*, and 0.75 for *Reliability*) and a modest RMSE (2.8 ft. for *Discrimination*, and 3.0 ft. for *Reliability*). The un-calibrated model on the other hand, tends to over-forecast the observed high stages (*Discrimination* ME of 3.2 ft.), and it tends to forecast too many high stages (*Reliability* ME of 10.5 ft.). In addition, the high stage *Discrimination* and *Reliability* Correlations for the un-calibrated model are low (0.55 for *Discrimination*, and 0.35 for *Reliability*). It is possible to make extensive comparisons of model calibrations, but from this brief summary, it can be seen the expert calibration provides a considerable improvement to the simulations compared to the a-priori parameters.

4.5.2 Description of the QPF skill

A short summary of the QPF skill is presented in order to help explain the behavior of the hindcasts. The same procedures used to analyze the stage hindcasts are used for the QPF except the Mean Error (ME) is reported in addition to the RMSE. The ME is reported because the bias characteristics of the QPF are important to understanding the effect of

the QPF on the hindcasts. The threshold (25 mm) was selected so the number of observations in the high precipitation category was near that for the high stage category. A zero QPF forecast was used as a baseline rather than persistence, because zero QPF is a common alternative to modeled QPF while persistence is not. There is little variation in the QPF metrics across the four lead-times so the 6 hour forecasts were pooled into a single sample set. That is, the forecast precipitation amounts were not added together to produce a single 24 hour qpf, the forecasts were collected into a single sample set for the 24 hour period. The metrics reported in Table 9 are for the 24 hour period.

For both *Discrimination* and *Reliability*, the NWS issued forecasts have lower RMSE and ME than the Zero QPF in both categories, demonstrating the NWS QPF forecast process adds skill over a zero QPF. However, there is still considerable uncertainty in the QPF. For example, the *Discrimination* RMSE (25.7 mm) and ME (-22.8 mm) for the issued forecasts are almost equal to the mean of the high precipitation observations (33.2 mm). For the lower category, the *Discrimination* ME for the issued QPF is small (0.2 mm), but the accumulated depth of incorrectly forecast rain for this category is 6136 mm. On the other hand, in those critical times when there were large rain events (obs > 25 mm), the forecasts are too low. The accumulated depth of rain under-forecast for the high *Discrimination* category is -3329 mm. These characteristics of the QPFs, not enough rain when there should be rain and too much rain when there should not be any, are seen later in the description of the hindcasts.

In order to provide some perspective on the quality of these QPFs in relation to the QPF across the United States (and therefore the relevance of these results to other places in the U.S.), the statistics from the NWS National Precipitation Verification Unit (NPVU) (McDonald et al, 2000 and NPVU, 2004) are provided in Table 10. The POD and the FAR are included because they are commonly used for meteorological verification. As can be seen from the table, the differences between these national statistics and the local statistics are small. The uncertainty seen in the QPFs on the hindcast basins may be considered representative of the the uncertainty in the QPFs across the country, and the error in the hdyrologic simulations caused by the QPF in the hindcasts, representative of the QPF driven error elsewhere in the U.S.

4.5.3 The hindcasts in relation to persistence

The persistence provides an interesting baseline perspective for the hindcast skill. For low stage *Discrimination* and *Reliability*, the only hindcasts which perform better than the persistence are the well calibrated scenarios with Perfect QPF at lead-times greater than 18 hours (P-V and P-NV in Figure 12). The un-calibrated parameters (P-V and P-NV in Figure 13) for both the low stage *Discrimination* and *Reliability* never perform better than persistence, even for the Perfect QPF scenarios.

In the case of the high stages however, the value of the NWS forecast process is more evident, as the hindcasts generally perform better than the persistence. For the high stage

		ME (by obs)	RMSE (by obs)	ME (by fcst)	RMSE (by fcst)	Samples (by obs)	Samples (by fcst)
Actual QPF	<=25 mm	0.2	2.5	0.1	3.0	31800	31920
Zero QPF	<=25 mm	-0.7	2.6	-0.7	3.5	31800	31946
Actual QPF	>25 mm	-22.8	25.7	15.5	20.3	146	26
Zero QPF	>25 mm	-33.2	34.7	NA	NA	146	NA

Table 9. The actual QPF compared to the zero QPF for the three hindcast basins.

		ME (by obs)	RMSE (by obs)	ME (by fcst)	RMSE (by fcst)	FAR	POD
National	<=25 mm	0.1	2.4	0.1	2.4	na	na
Hindcast basins	<=25 mm	0.2	2.5	0.1	3.0	na	na
National	>25 mm	-24.7	29.1	16	23	0.76	0.10
Hindcast basins	>25 mm	-22.8	25.7	15.5	20.3	0.77	0.04

Table 10. National Precipitation Verification Unit QPF statistics and the QPF statistics for the 3 hindcast basins.

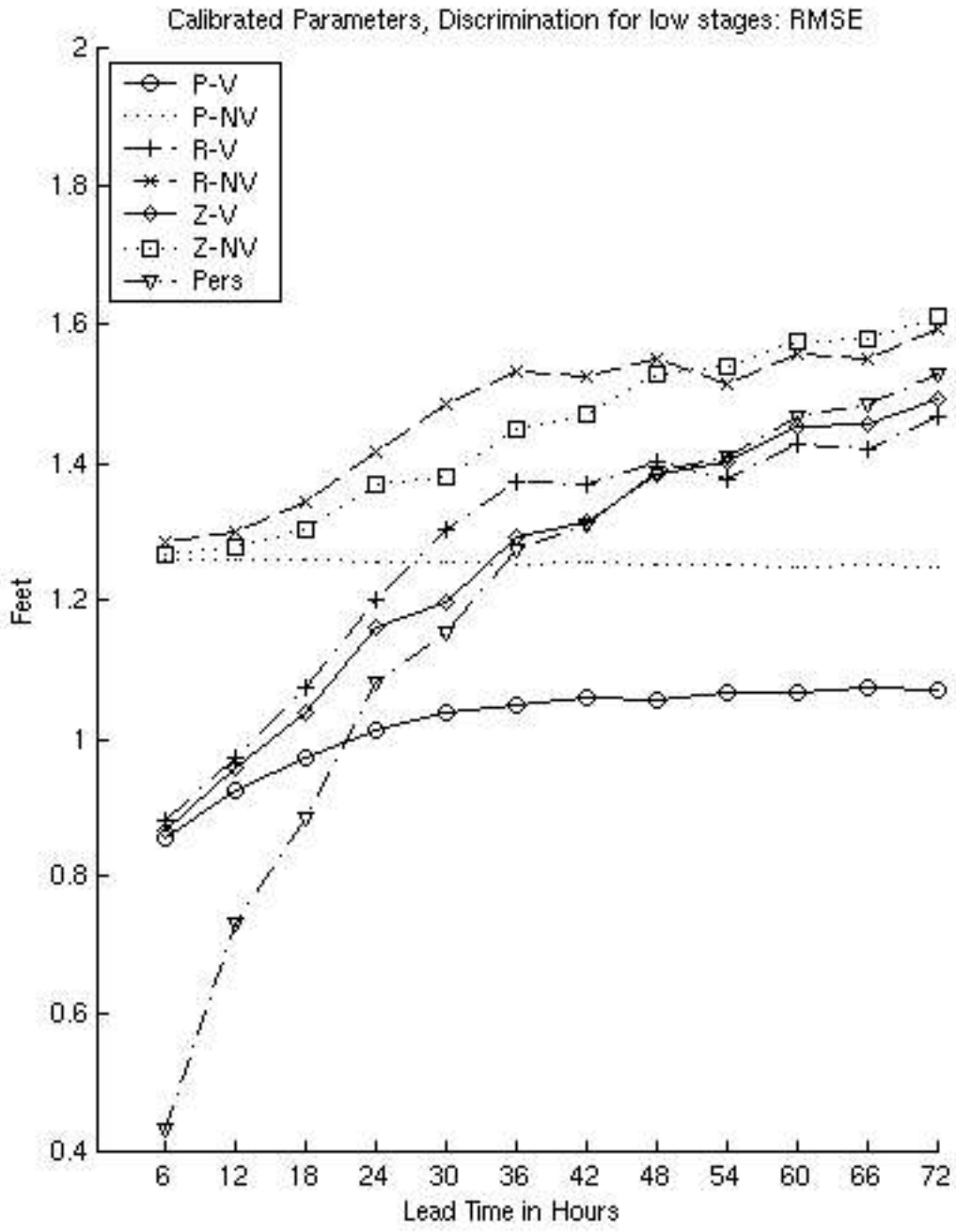

Figure 12: Low stage discrimination RMSE for the calibrated parameters.

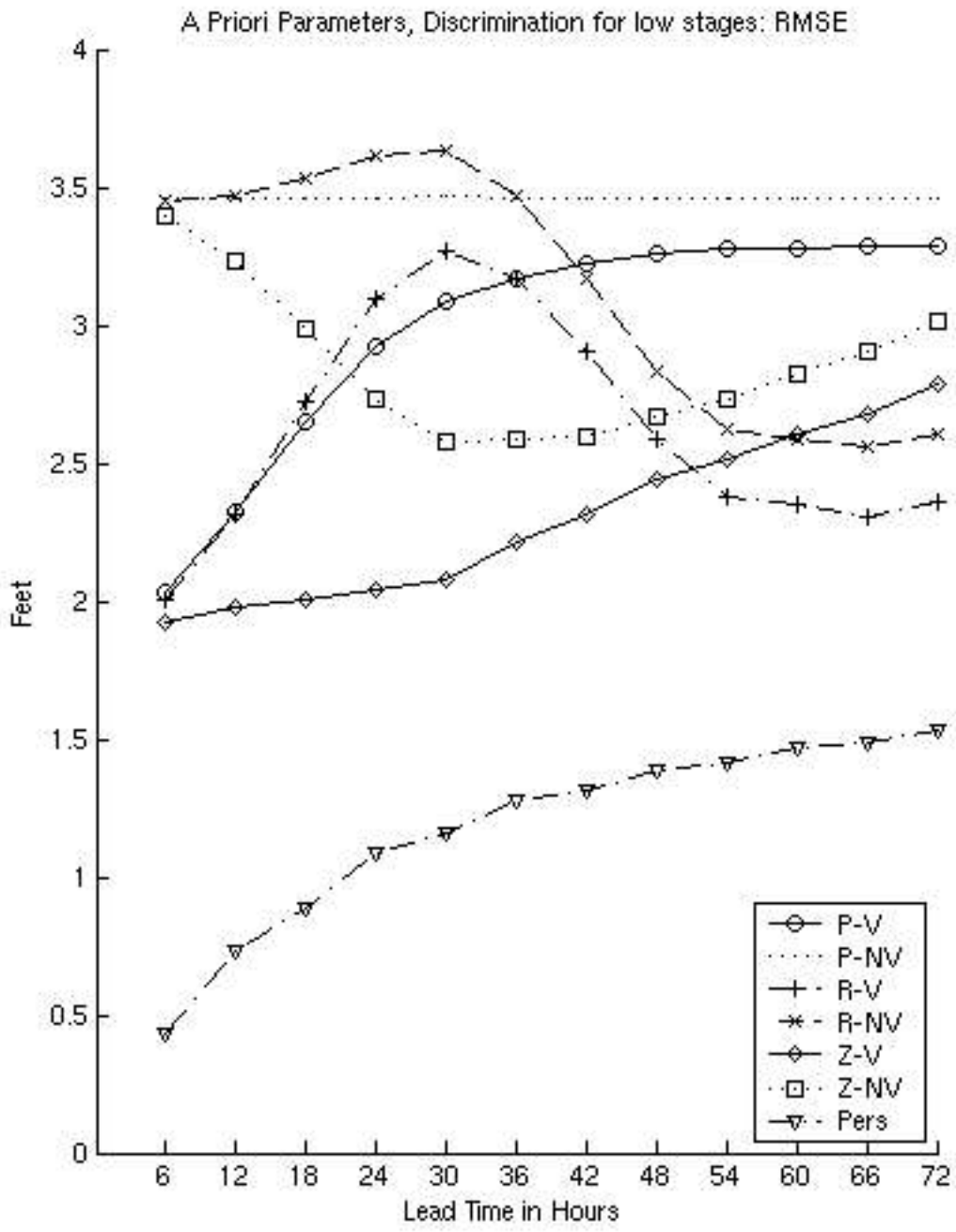

Figure 13: Low stage discrimination RMSE for the a priori parameters.

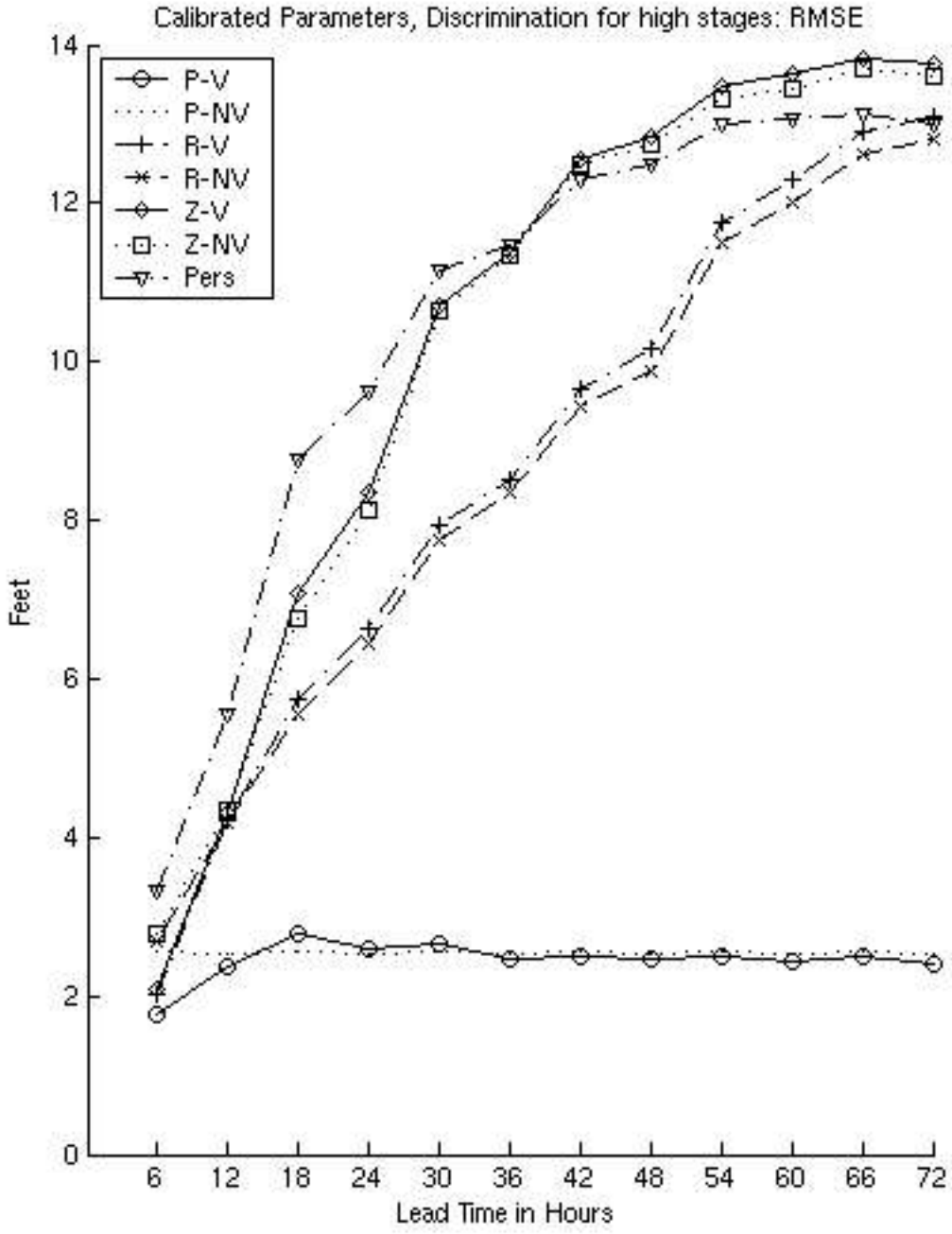

Figure 14: High stage discrimination RMSE for the calibrated parameters.

Discrimination, all the calibrated scenarios (Figure 14) perform better than the persistence for the first 24 hours. After 30 hours, the Zero QPF scenarios converge to the persistence and the Real QPF scenarios converge to persistence at hour 72. The Perfect QPF scenarios perform better than the persistence for all lead-times. For the un-calibrated parameters, the convergence to persistence is slightly faster with the Zero QPF scenarios converging after 24 hours and the Real QPF scenarios after 54 hours. For the high stage Reliability, the RMSE for the calibrated parameters are almost a factor of two smaller than the persistence RMSE, while the un-calibrated RMSE (Figure 15), are larger than the persistence RMSE. By comparing their relation to persistence, it can be seen the hindcasts have greater *Reliability* skill than *Discrimination* skill.

4.5.4 The effect of the calibration upon the hindcast skill

The results of comparing the skill of the calibrated and un-calibrated hindcast scenarios indicate the calibration is important for the low stage skill and for the high stages when the leadtime is less than a day, but the skill the calibration can provide to the high stages at leadtimes greater than a day is limited when the QPF is poor. The differences between the hindcast RMSEs for the low stage *Discrimination* (Figure 16) indicate the calibration provides considerable improvement to the hindcasts, reducing the RMSE to a half of the original un-calibrated RMSE. The calibration provides the most improvement to the Perfect and Real QPF scenarios as opposed to the Zero QPF scenarios because both the Real and the Perfect QPF include precipitation which must be converted to runoff. The

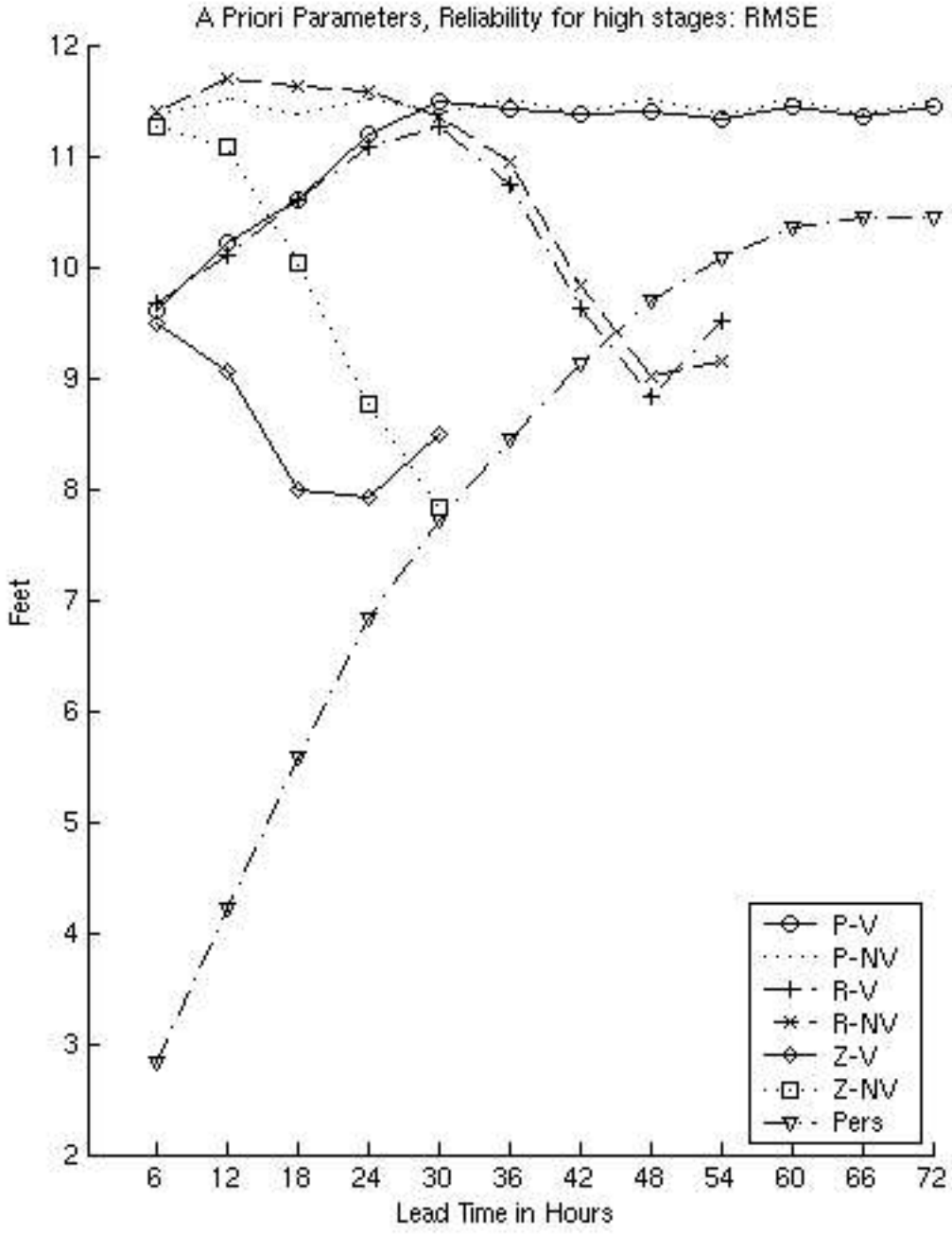

Figure 15: High stage reliability RMSE for the a priori parameters.

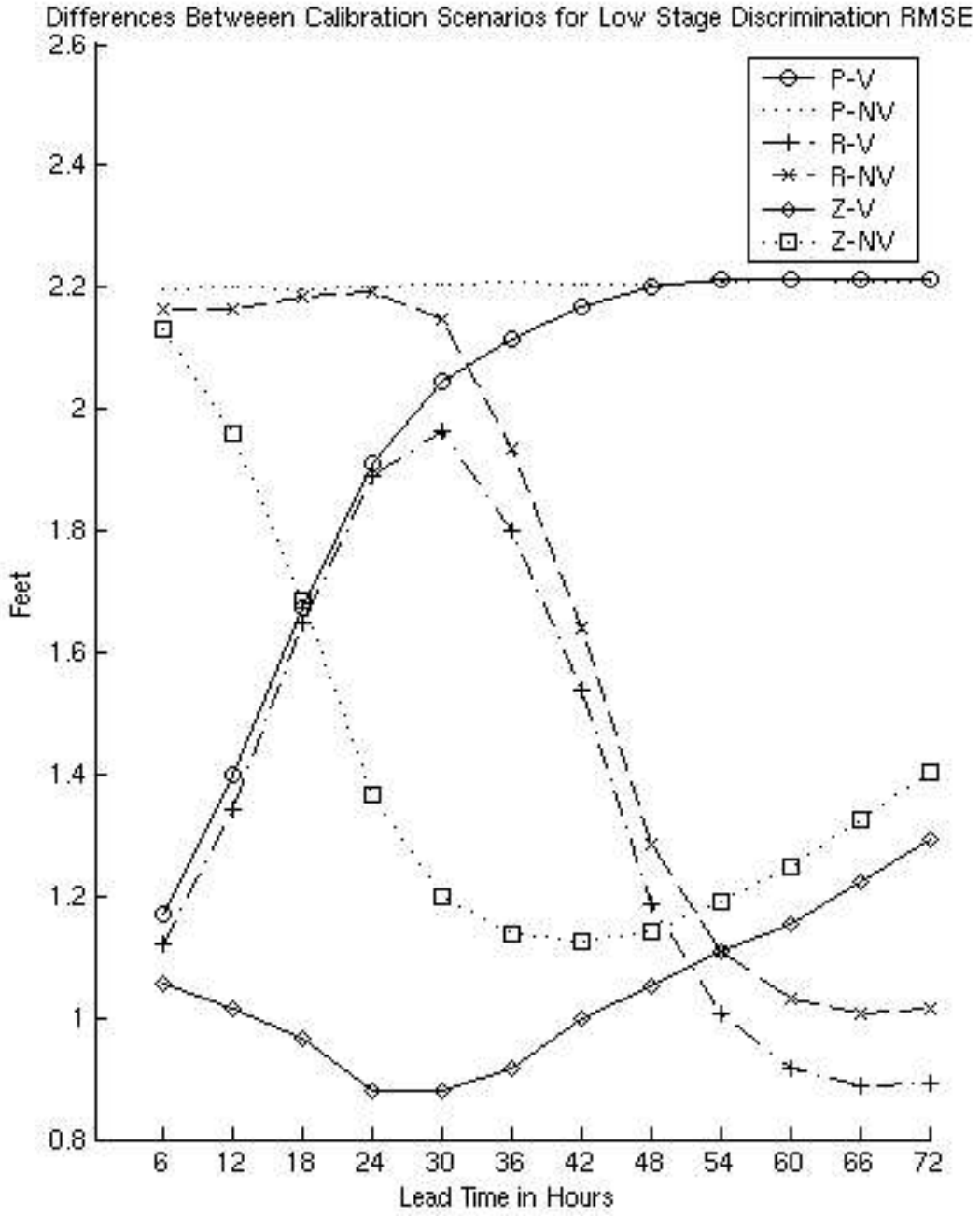

Figure 16: Differences between calibration scenarios, low stage discrimination RMSE.

improvement to the Real QPF scenario matches the improvement to the Perfect QPF scenario until the Real QPF turns to zero (at 24 hours) and then the Real QPF scenarios parallel the Zero QPF scenario. The Zero QPF scenario sees little benefit from the calibration except in the early lead-times because there is no rainfall to convert to runoff. The calibration improves the hindcasts without the state updating more than those with the state updating indicating the calibration and the state updating contribute similar improved initial conditions to the hindcasts.

As was noted above (Section 4.5.1), the expert calibration provides an improvement of a little over 3 feet, half the a-priori RMSE, to the high stages for the Perfect QPF hindcasts. The hindcasts without state updating realize this 3 foot improvement in the first time steps (Figure 17), but the updated scenarios benefit much less (1 foot). As with the low stages, this difference indicates the calibration can provide skill through good initial conditions, and, as will be seen in the next section, the calibration and the skilled state updating provide comparable skill in the first day. Unfortunately, at the later lead-times, when the Real or the Zero QPF is used instead of the Perfect QPF, the magnitude of the improvement to the *Discrimination* skill from the calibration falls to zero at 36 hours and then becomes negative, though it does rise back up at the end of the third day. This fall in the benefit of the calibration is caused by the meteorological error overwhelming the value of the skilled calibration. Though calibration provides skill to the short leadtimes,

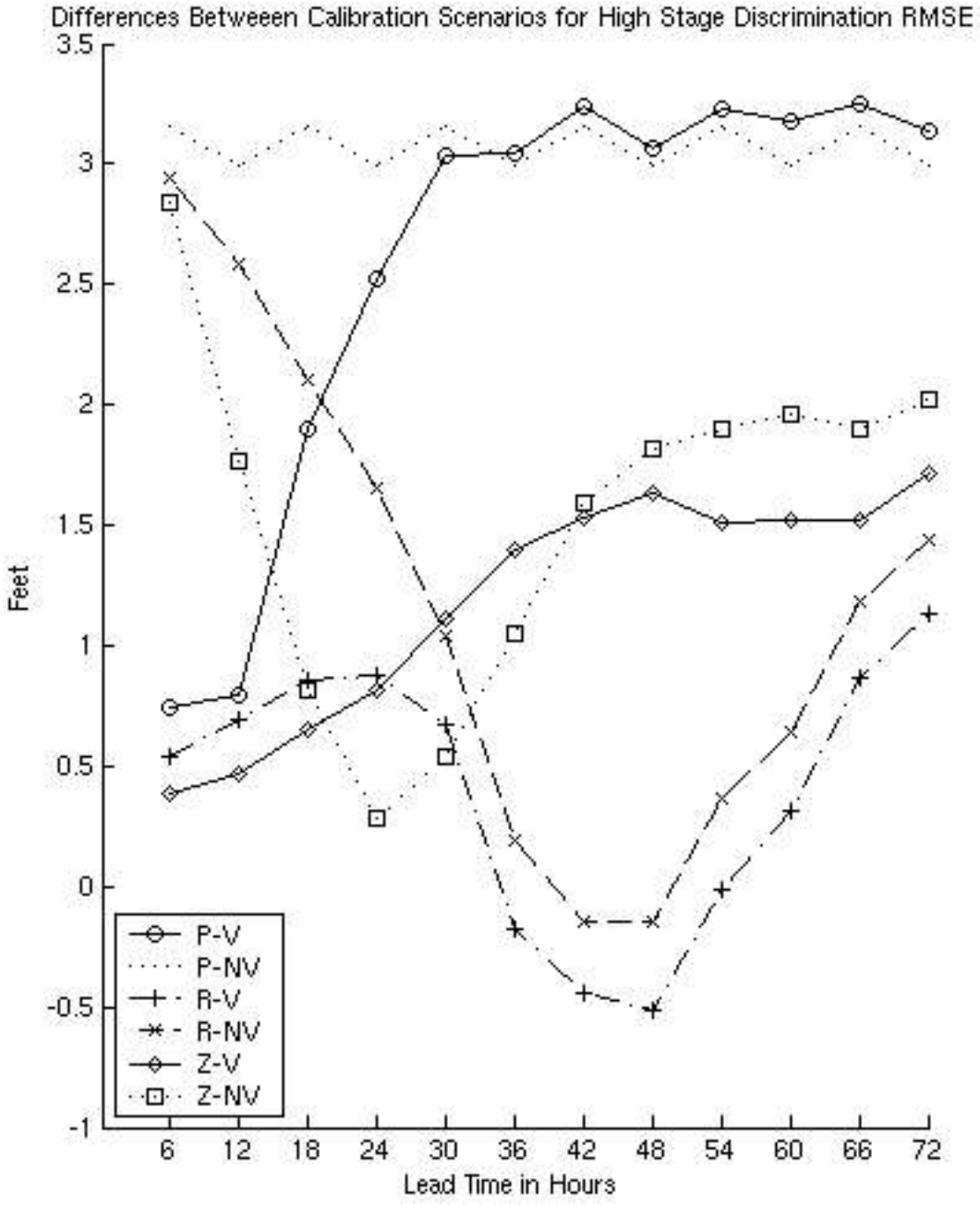

Figure 17: Differences between calibration scenarios, high stage discrimination RMSE.

it only resolves a small portion of the total *Discrimination* uncertainty at the longer lead-times.

The dipping and rising pattern seen in Figure 17 for the Real QPF scenarios in days 2 and 3 is caused by the tendency of the QPF to under-forecast interacting with the tendency of the un-calibrated hydrologic model to over-forecast. During the early lead-times, the Real QPF causes the un-calibrated model to over-react and improving the calibration can improve the hindcasts. After the first 24 hours, zeros are used in the Real QPF and the zeros tend to mitigate the tendency of the un-calibrated model to over-forecast, while at the same time causing the calibrated model to under-forecast. Therefore, calibrating the models does not improve the hindcasts. At the longest lead-times there is no longer any over-forecasting to mitigate because the QPF has been zero for the past 24 hours; therefore the calibration can begin to add skill again. This rising and falling is a clear indication that the hydrologic and the meteorological errors are neither independent nor additive.

While the *Discrimination* skill is only slightly sensitive to the calibration, the *Reliability* skill is very sensitive to the calibration. The improvement provided to the high stage *Reliability* by the calibration (Figure 18) is more than half the total error for the a-priori parameters. For the Zero QPF scenarios, this improvement falls quickly, but for the Real QPF scenario, the improvement holds up for the first 24 hours before it begins to fall.

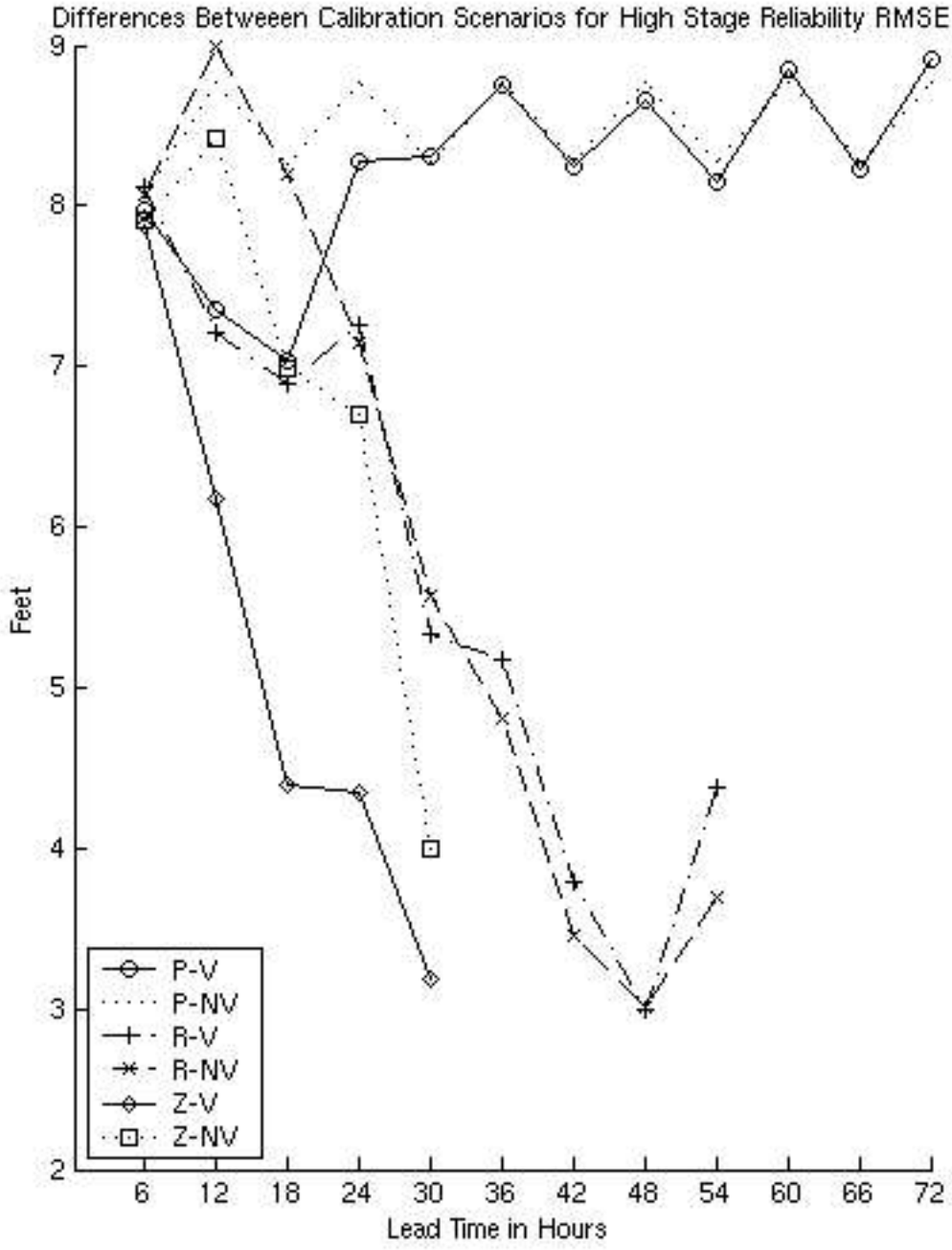

Figure 18: Differences between calibration scenarios, high stage reliability RMSE.

Again the value of the calibration is reduced when the QPF is zero and there is no rain to convert to runoff.

4.5.5 The effect of updating and not updating the initial model states on the hindcast skill

In the comparisons between the "skilled" and "un-skilled" state updating for the low stage *Discrimination* (Figure 19), the hindcasts group themselves by the calibration scenario, with the un-calibrated scenarios showing greater improvement than the calibrated scenarios. This is the same phenomenon seen in the calibration comparisons with the updated and the non-updated scenarios grouping themselves. The improvement provided by the initial conditions drops steeply until the end of day 1. Although the improvement does not drop all the way to zero it flattens to less than half a foot at 42 hours. While the calibration and the state updating interact with one another, the QPF treatment has little influence upon the value of the state updating as there is little distinction between the QPF scenarios.

For the high stage *Discrimination* (see Figure 20) the same pattern is apparent: the scenarios group themselves by the calibration, not by the QPF. The fact that the QPF does not influence the differences between the "skilled" and "un-skilled" initial condition scenarios does not imply the QPF does not change the hindcast hydrographs. It only implies that the skill brought to the hindcasts by the initial conditions is independent of the skill of the QPF. By the same token, the greater improvement for the un-calibrated

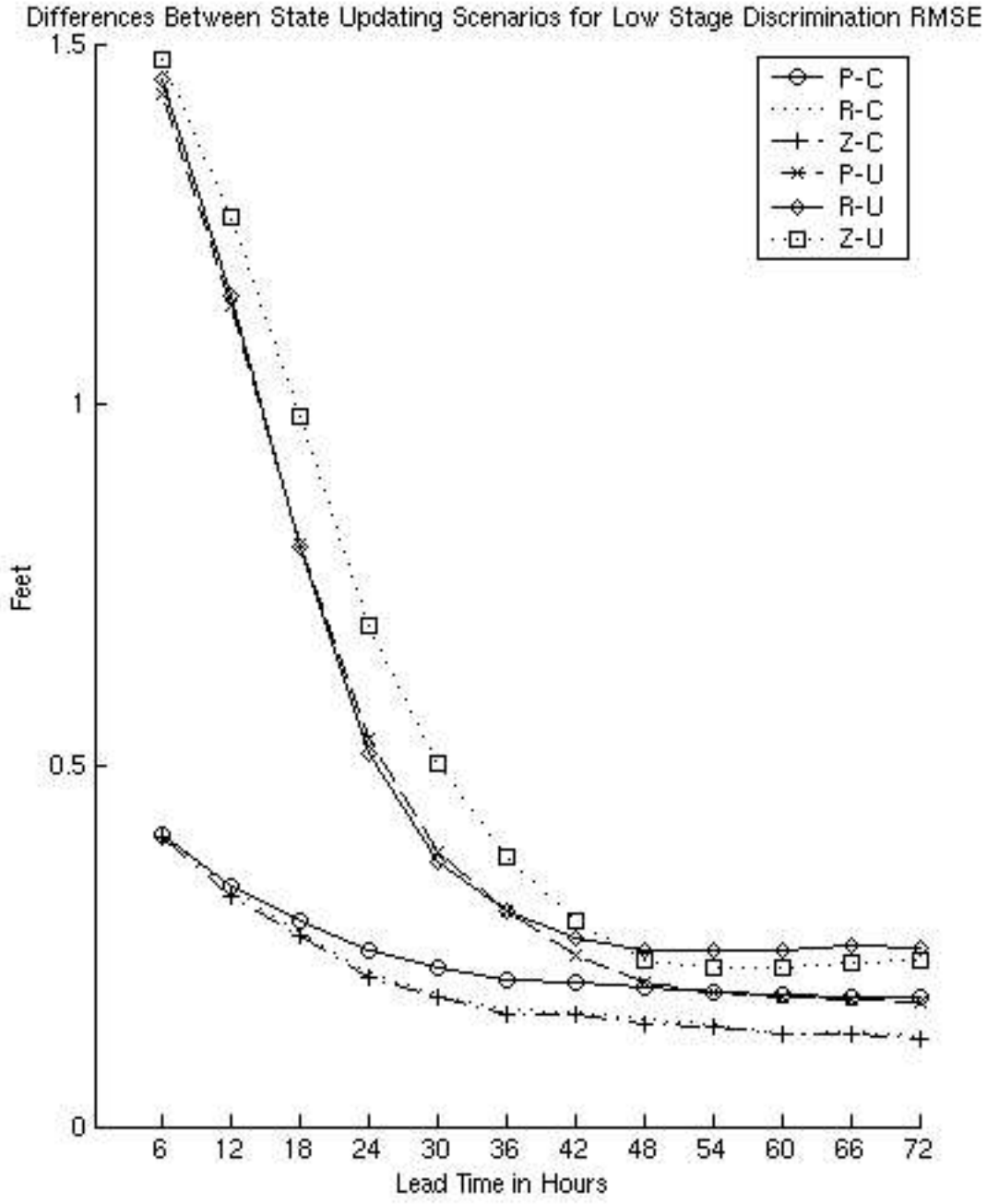

Figure 19: Differences between state updating scenarios, low stage discrimination RMSE.

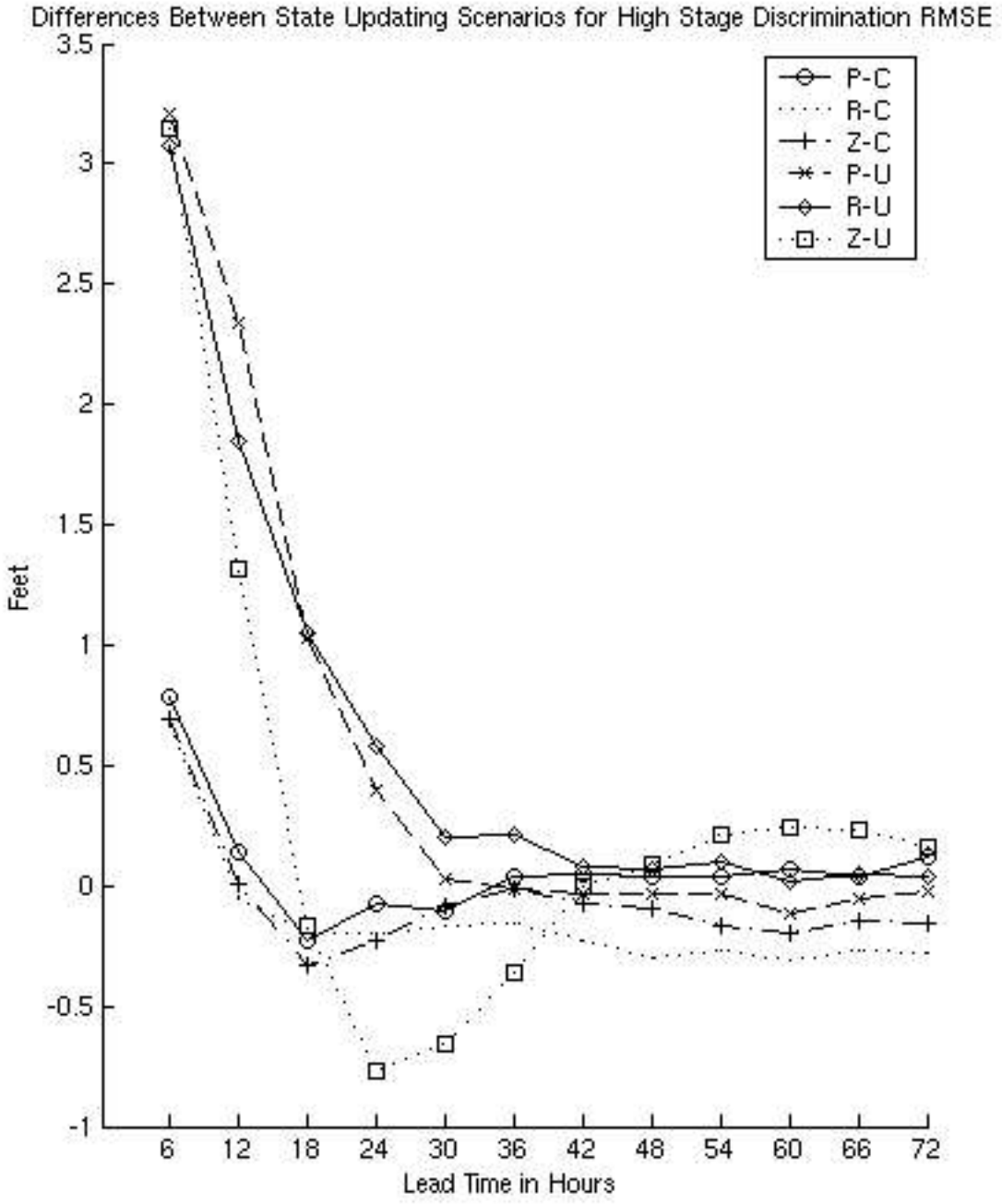

Figure 20: Differences between state updating scenarios, high stage discrimination RMSE

scenarios does not indicate the calibration harms the state updating, rather it indicates there are multiple methods of improving the initial conditions from which the hindcasts are generated. For the low stages, the calibration provides slightly more improvement(2.2 ft.) to the 6 hour lead-time than the state updating (1.5 ft.), but for the high stages, the calibration and the state updating bring an equal amount of improvement (3 ft) to the 6 hour lead-time. For the high stage *Reliability,* the pattern of improvement is the same, but the magnitude is less (2 ft.).

4.5.6 The effect of improving the QPF on the hindcast skill

For the low stage *Discrimination* (Figure 21) the type of QPF makes little difference to the hindcasts; the improvement to the RMSE due to improving the QPF stays below 0.4 feet for all the scenarios and only reaches 0.4 feet in day 3. This is much less than the minimum one foot improvement provided by the calibration and the state updating in the early lead-times. The non-linear interaction of the meteorological and hydrologic error is again visible in these comparisons. Changing the QPF from the Zero QPF to the Real QPF makes close to zero change in the RMSE of the calibrated model. While for the un-calibrated model, improving the QPF from the Zero QPF scenario to either the Real or the Perfect QPF actually harms the hindcast RMSE (improvement of -1.2 feet) at hour 30. After hour 30, the Real QPF scenario shows improvement even though the QPF is set to zero for days 2 and 3. These rises and falls can be traced to the changes in the forecast variance as the un-calibrated model responds too strongly to the non-zero QPF where

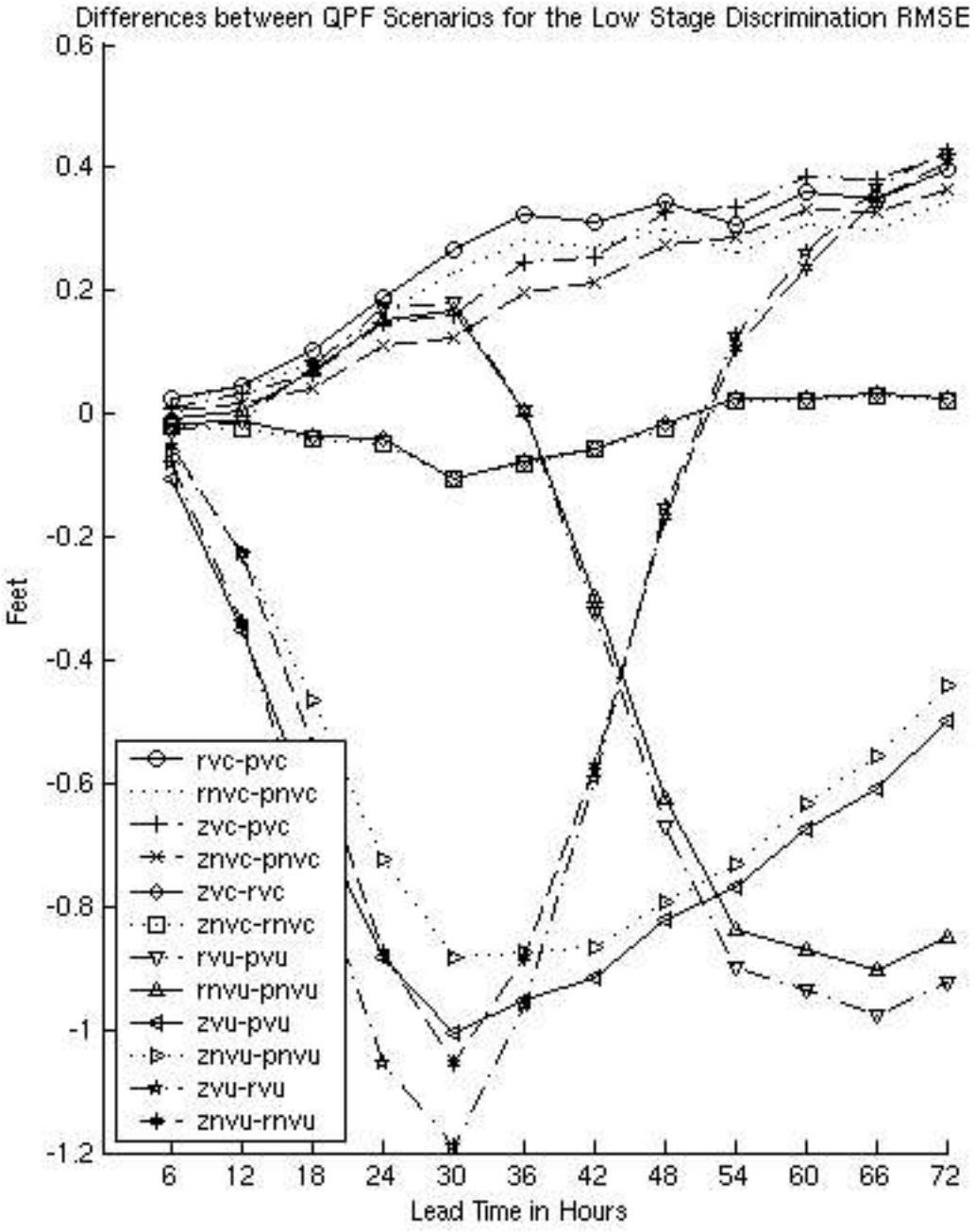

Figure 21: Differences between the QPF scenarios for the low stage discrimination RMSE.

previously the zero QPF had mitigated this tendency. The same pattern is visible in the low stage *Reliability* statistics, though it is muted.

For the high stage *Discrimination* (Figure 22), the QPF plays a central role in the success of the hindcasts with all the scenarios showing improvements due to improved QPF. Like the low stages, the QPF improvement does not depend upon the initial conditions as all the scenarios begin near zero for the first lead-time. The value of the Real QPF can be seen in the way the improvement derived from converting to the Real QPF from the Zero QPF falls toward zero after the Real QPF has switched to zeros (leadtimes greater than 24 hours). The transition to the Perfect QPF shows the potential improvement due to improving the QPF. This potential is large, especially for the long lead-times where it is over 10 feet.

While the improvement to the *Discrimination* skill from the three QPF scenarios was similar for the calibrated and the un-calibrated model, the improvement to the high stage *Reliability* (Figure 23) shows marked differences between the calibrated and the un-calibrated results. For the calibrated high stages, switching between the QPF types makes no change to the hindcast *Reliability* for the first 18 hours. After 18 hours switching to Perfect QPF improves the hindcast *Reliability*, but switching from the Zero to Real QPF causes a negative improvement. For the un-calibrated parameters the hindcasts degrade when using improved QPF with the Real to Perfect QPF scenario falling after 30 hours

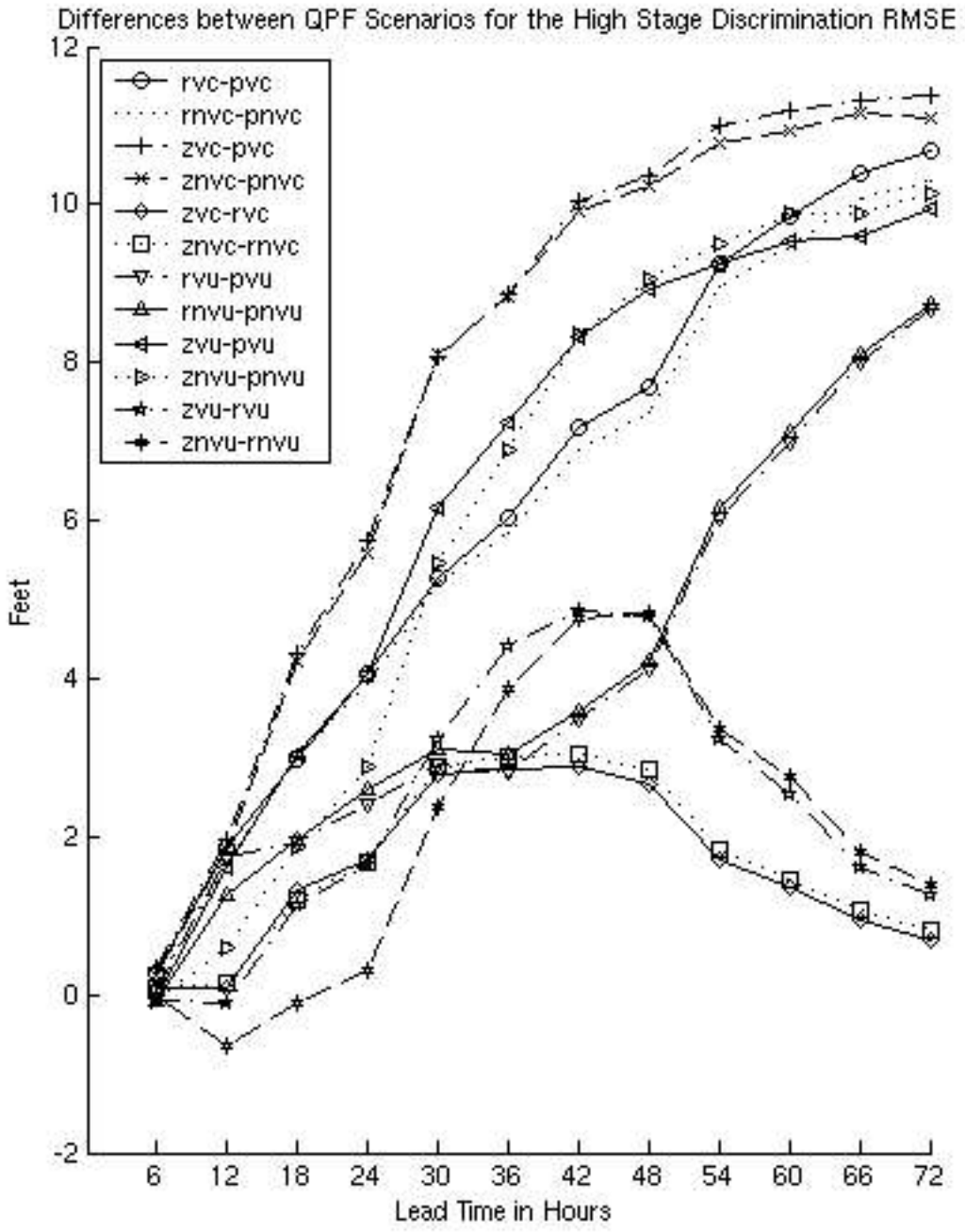

Figure 22: Differences between the QPF scenarios for the high stage discrimination RMSE.

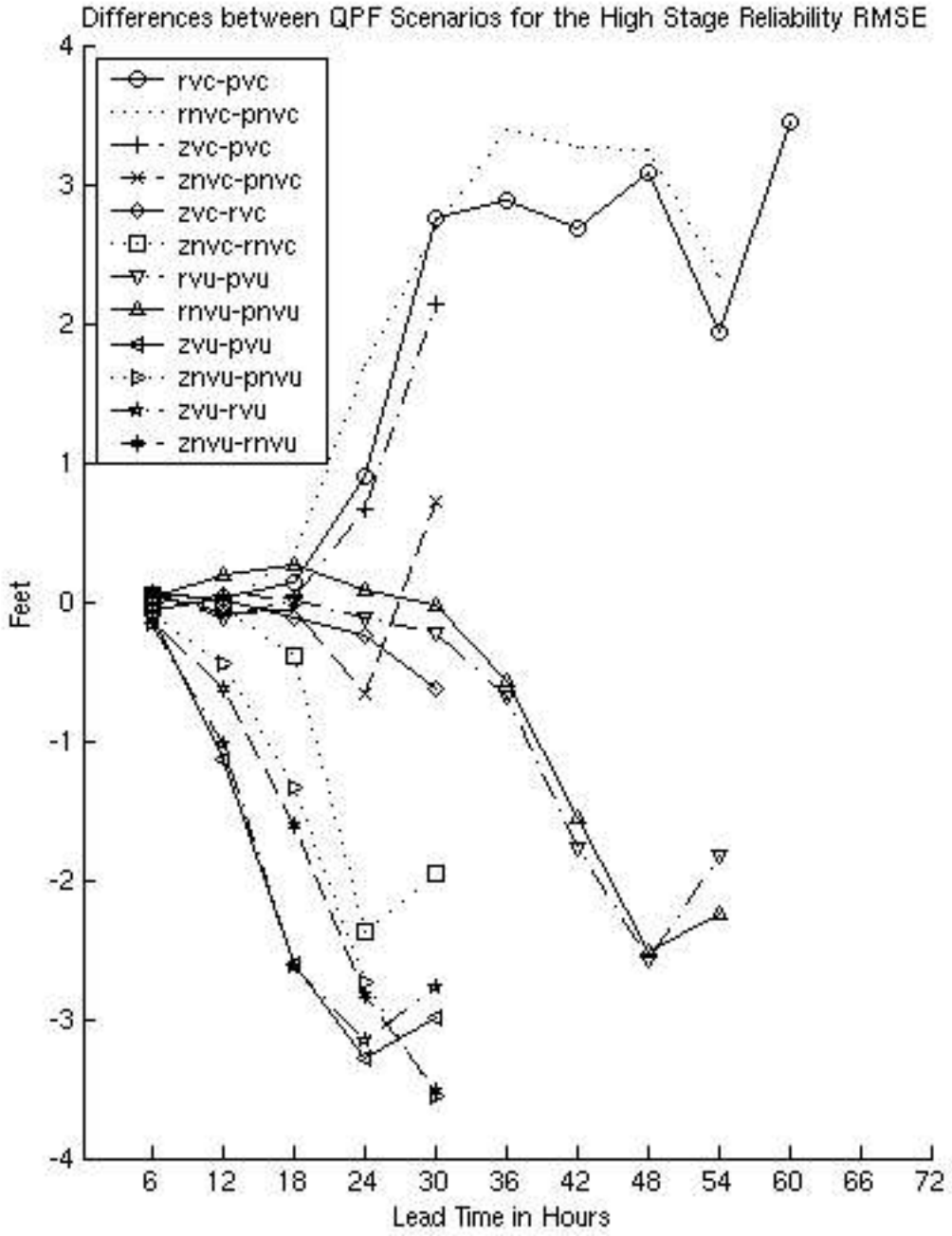

Figure 23: Differences between the QPF scenarios for the high stage reliability RMSE.

and the others falling right from the start. These negative improvements indicate the need to have a well calibrated model to take advantage of improved QPF.

4.5.7 Hindcast Sample Sizes

The hindcast sample sizes reflect the description of the skill derived with the RMSE. The un-calibrated model's tendency to over-forecast is seen in the large *Reliability* sample sizes (Figures 24 and 25) for the high stage category in the early lead-times. The steep drop in skill of the high stage forecasts with lead-time is seen in the steep drop of the Sample Sizes for the high stage *Reliability.*

The sample sizes can also be used to assess the uncertainty in the computed metrics. The low stage category Sample Sizes for both *Discrimination* and *Reliability* are all above 7500 samples at each time step. Even though there is serial correlation between the samples, this large number of samples provides confidence to the low stage metrics. The high stage metrics, on the other hand, are computed from many fewer samples. For the *Discrimination* metrics, the sample sizes for each time step are all greater than 39 samples. The *Reliability* Sample Sizes (Figures 24 and 25) vary from reasonably high (400 samples) to very small (5 samples). Clearly there is much greater uncertainty associated with the high stage category metrics. Several experiments were conducted with changes to the threshold between the high and low stages. The ordering and patterns seen in the metrics remained consistent. It seems reasonable therefore, to believe the

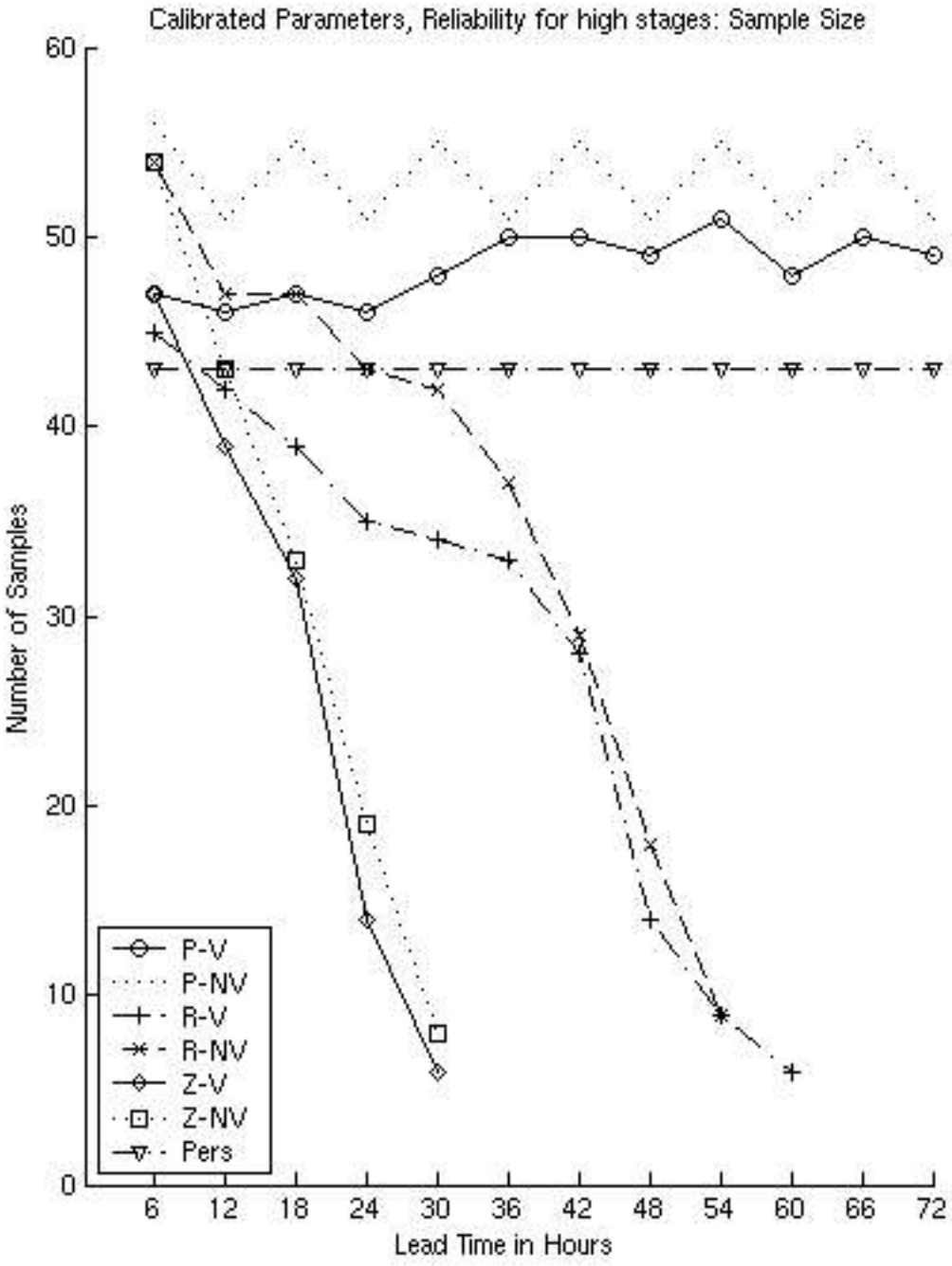

Figure 24: Hindcast sample sizes for calibrated parameters and the high stage reliability.

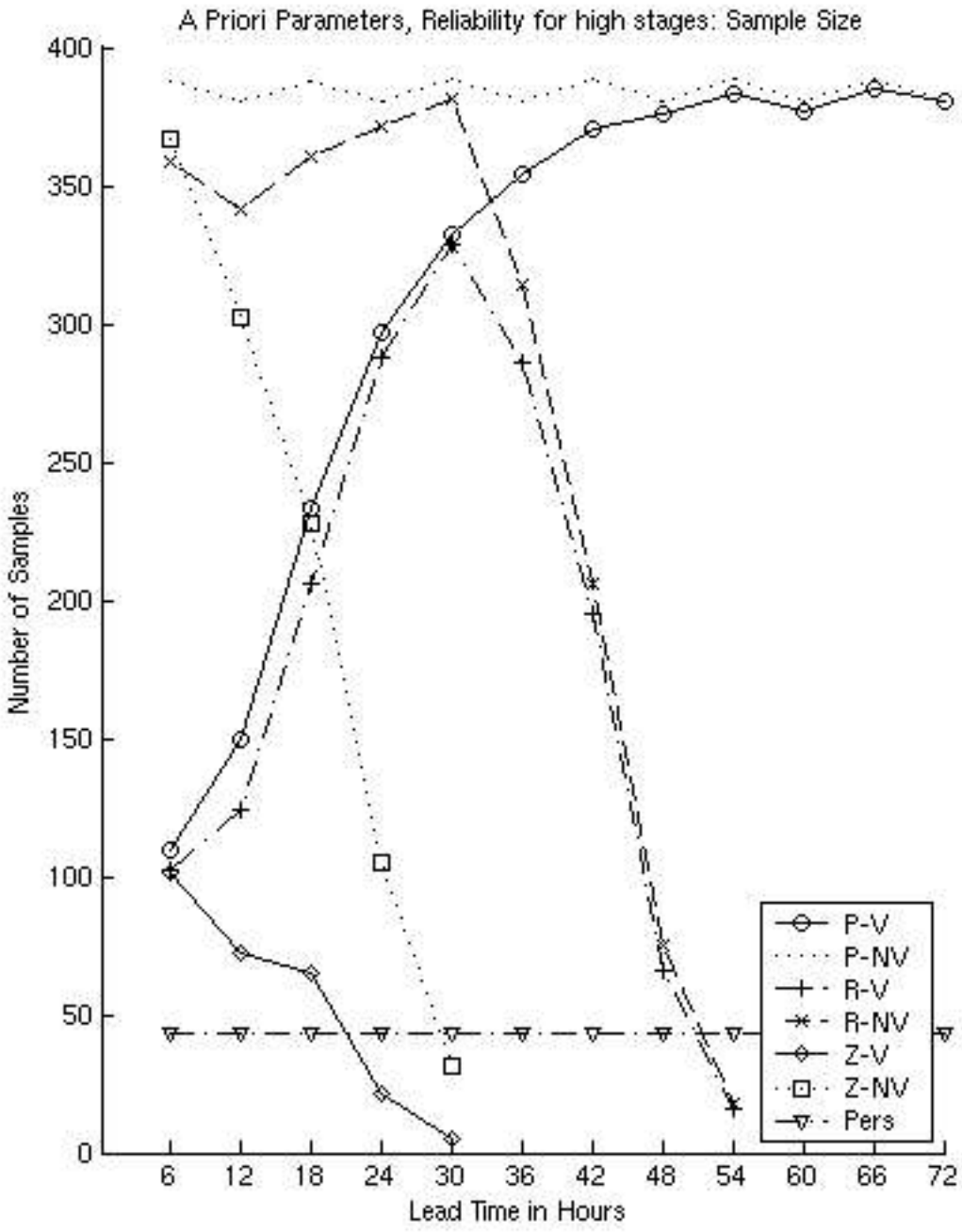

Figure 25: Hindcast sample sizes for a priori parameters and the high stage reliability.

purpose of the NWS forecasts is flood forecasting, and therefore it is the high stageswhich are critical to the success of the NWS mission.

4.6 Discussion of the hindcast experiment results

The hindcast experiment provides two categories of insight. First, it provides insight into the sources of skill in the forecasts studied here and the relations between those sources of skill. Second, it offers insight into the requirements of a comprehensive verification program for hydrologic forecasts.

4.6.1 The role of calibration, initial conditions and QPF in forecast skill with lead-time

The role of the three forecast process elements in contributing to the skill of the hindcasts changes with lead-time and with the type of skill being measured. For the very short lead times (18 hours or less) the *Discrimination* skill for both the high and the low stages is controlled by the initial conditions. Good initial conditions can be derived from a good calibration or from effective state updating procedures. While improved initial conditions lead to improved hindcasts, these improvements are necessarily limited to the first few time steps when the initial conditions can influence the skill of the forecasts. In addition, the initial states control the forecast skill at these short lead times irrespective of the QPF and the calibration, indicating it is possible to take advantage of good initial conditions even with the present QPF skill and without extensive model calibration. For the poorly calibrated model, the state updating provides greater benefit, because there is

more error to be corrected. The duration of the improvement continues for longer with the poorly calibrated model as well, again, because the well calibrated model requires less correction.

At the longer lead-times, uncertain meteorological input is the largest source of uncertainty in the hindcast *Discrimination* skill. This can be seen by the large differences between the Perfect QPF and the Real QPF scenarios and at the same time the very small differences between the well calibrated and the un-calibrated model when using Zero or Real QPF. Although the QPF is the largest source of error in the hindcasts at the longer lead-times for the high stage *Discrimination* skill, the control of the forecast skill is not limited to the QPF, but rather a mix of the QPF and the calibration. Neither one of them controls the skill independently of the other and no assumption can be made with respect to the likely result in the *Discrimination* skill when changes are made to one or the other. Improving the calibration may have little influence upon the forecast skill if the QPF has little skill as was seen in the calibration comparisons for the Zero QPF scenarios (Figure 17). At the same time, improving the QPF may degrade the forecasts if the calibration is biased and this bias is accounting for errors in the QPF as was seen in the QPF comparisons for the transition from the Zero QPF to the Real QPF (Figure 22). This interaction between the hydrologic error and the meteorological error is the same phenomenon noted by Krzysztofowicz (Krzysztofowicz, 1999b) when he found the common notion that the hydrologic and meteorological error are additive was false.

While the present day QPF skill limits the improvement possible in the hindcasts due to improving the model calibrations, this does not mean calibration is not an essential element of the hydrologic forecast model implementation; in these hindcasts, improving the QPF improved the hindcast *Discrimination* skill most with a well calibrated model. The *Reliability* skill, on the other hand, is not controlled by the QPF and initial conditions, but by the calibration. However, the *Reliability* itself appears unreliable as it vanishes at the longer lead-times when the hindcasts no longer rise up into the higher stage category.

4.6.2 Implications for hydrologic verification

This analysis provides two fundamental building blocks to the development of hydrologic verification. First, this analysis offers an initial insight into the connections between verification metrics or subsets and the forecast process. Second a simple method for analysing the forecasts has been demonstrated to be successful. From these building blocks several core recommendations for hydrologic verification systems can be derived.

By identifying the sources of error in these hindcasts connections between verification metrics and the forecast process elements can be identified. The low stage *Discrimination* metrics are connected to the calibration and the initial conditions. *Discrimination* metrics computed for the early lead-times and high stages are connected

to the initial conditions. The high stage *Discrimination* metrics at longer lead-times are connected to the QPF. The high stage *Reliability* metrics are also connected to the calibration. Once these connections have been identified they can be used to identify appropriate measures to quantify the change in forecast skill resulting from enhancements to the forecast system. To quantify improvements provided by calibration, the *Reliability* of the forecasts should be measured. Improvements in the state updating procedures should be reflected in improvements in the short term forecast *Discrimination* skill. Upgrades to observing networks, which will improve the initial conditions, will also be reflected in the short term *Discrimination* skill. Improvements to the QPF are best measured by the *Discrimination* skill past day 1 when the model is well calibrated.

Even in this small set of homogeneous hindcasts there was considerable variety in the characteristics of the forecast skill, and in the way that forecast skill changed with changes in the forecast process. This variety of skill attributes points out the difficulty of finding a single representative metric or subset to assess forecast skill across a region, much less the Nation. Any single metric would necessarily capture only some of the important changes to the forecast skill.

The analyses presented here, also demonstrate that simple comparisons of control, baseline and actual forecasts combined with analysis of the input forecasts can be used to establish objective insight into the sources of forecast skill and error. Error analysis is a

critical element of verification, and therefore any operational verification system must include sufficient control forecasts to capture the changes in skill provided by individual forecast process elements. For example, just having a persistence baseline and the Perfect QPF simulation without state updating may be sufficient for an operational system. The persistence forecast provides an objective baseline for minimum forecast performance while the Perfect QPF simulation allows the forecast verifier to distinguish between model calibration error and error in the initial conditions or the QPF depending upon the lead-time. A well performing forecast system will show better skill than persistence at all lead-times. If the Perfect QPF simulation for the high stages is not as good as the persistence, this indicates there is a problem with the calibration. At the short lead-times the *Discrimination* skill of the actual forecasts should be better than the Perfect QPF scenario for the high and the low stages. If the early periods of the actual forecasts are not better than the Perfect QPF simulation, then the initial state updating is not adding much skill. If the initial state updating does not add much skill, it may be the result of having a good calibration, or a poor state updating procedure. Comparisons to the persistence or *Reliability* metrics can be used to determine which is the case. If the initial state updating adds substantial skill to the forecasts, then it is likely the model calibration could be improved. As the initial conditions become less influential, and the QPF becomes important, the *Discrimination* statistics for the actual forecasts will perform less well than the Perfect QPF scenario. At the longer lead-times, the magnitude of the difference between the metrics for the forecasts and the Perfect QPF scenario is an

indicator of the size of the error caused by erroneous QPF. The insight from these types of simple comparisons can provide the means for hydrologists to gain the understanding needed to develop an objective description of the forecast skill for all types of hydrologic forecasts and to track changes to the forecast skill.

This analysis also demonstrates the importance of developing well defined verification procedures (like those which already exist for calibration). One common question considered by most operational forecasters is "should QPF be used?" The common conclusion derived from years of experience and some verification results, is that the QPF is not useful. As an example, the Arkansas Basin River Forecast Center has published the RMSE at monthly time-steps across all their forecasts for two forecast scenarios, one with and one without QPF (ABRFC, 2004). Though, the forecasts with QPF are better than the non-QPF forecasts in 69% of the months since 2000, they are only very slightly better, 3% overall. Because the ABRFC metrics are an aggregate of all the forecasts, the metrics reflect the characteristics of the low stage category and the ABRFC result is the same result seen in this study for the low stage category where the Real and the Zero QPF provided comparable skill to the hindcasts. However, for the high stage category in this study, the QPF was an important element of the forecast skill. If the ABRFC metrics were computed for two categories, the metrics might indicate the importance of using QPF was much greater than it is currently considered to be.

4.7 Conclusions from the hindcast experiment

From this hindcast experiment, several fundamental elements of a hydrologic forecast verification process can be defined.

- First, forecast quality is multi-faceted and no single metric will describe the many ways a forecast can be good (or bad). Attempts to construct single universal metrics will result in metrics which obscure the causes for the changes in forecast skill.
- Second, sorting forecasts into appropriate subsets is a necessary and effective means of determining elements of the forecast skill. Different methods of sorting expose different characteristics of the forecasts.
- Third in order to support effective error analysis, both control and unskilled baseline forecasts are required to make the verification meaningful. Without these additional forecasts there is not sufficient background information to determine sources of error or skill or areas which require improvement.
- Fourth, it is essential that all the input forecasts to the system be verified along side the hydrologic forecasts. This requirement is likely to become more important as verification analyses move downstream into more complex basin configurations where reservoir outflow forecasts are likely to have a substantial influence on the forecast skill.
- The fifth, and perhaps most importantly, more studies like this are needed. This initial study provides only a start on the larger project of developing an objective description of hydrologic forecast skill. Analysis of the error at downstream forecast locations

(non-headwater locations) is important and requires study as well. Unfortunately, such studies are hampered by the cost of developing the infrastructure to compute hindcasts down the length of a large river across hundreds of basins. However, such studies are needed if a complete understanding of the hydrologic forecast process is to be established.

Developing an objective and comprehensive understanding of the forecast error and skill sources is an essential step toward improving hydrologic forecasts. Well designed verification systems which include analysis procedures such as the one illustrated here are at the center of developing this comprehensive understanding. A proposal for just such a system is presented in the next section.

5 A PROPOSAL FOR STANDARDIZED EVALUATION PROCEDURES

This section proposes standard forecast evaluation procedures.. The purpose of this proposal is to initiate the discussion and implementation of verification systems throughout the hydrology community. As was noted in Section 2, Literature Review, the notion of standardized verification systems is not new. The meteorological community has developed numerous sets of standards in order to ensure verification is conducted according to accepted scientific practices. The hydrology community needs to follow suit.

Standardized verification procedures offer the hydrologic forecasting community (both forecasters and forecast researchers) a number of benefits. They enhance the communication between people who are using, developing, managing and constructing forecasts (WMO, 2002). Using standard verification procedures facilitates the comparison of forecast methods developed and implemented in widely disparate locales. Standardized procedures also provide a template to operational agencies for the implementation of their verification systems. By using an established set of standards, operational agencies can be certain their methods will be scientifically sound and well understood by the community of people interested in the forecasts. One final benefit of standardized procedures is they can be treated as a baseline for additional research into improved methods for verifying forecasts. New methods can be described in terms of the improvement they make over the documented standards.

The procedures proposed here are expected to evolve as the understanding of the hydrologic forecast process grows, as the forecast process itself evolves, and as the hydrologic community develops insight into informative verification techniques. The examples in this paper are drawn from the NWS, but these procedures are intended to be applicable to hydrologic forecast processes in general whether they be university research projects, privately run, or based in other nations.

5.1 The method for designing these proposed Standardized Evaluation Procedures

When the meteorological community has developed standard verification procedures, they have done so through a committee of experts. For example, the Standardized Verification System for Long Range Forecasts published by the World Meteorological Organization (2002) was the work of an11 member team of experts. In the recent past, when the National Weather Service has established standardized verification practices, they have convened a committee to do so: once in 1939 with a three member Special Committee on Forecast Verification (NWS, 1939) and then again in 1980 the NWS convened a National Verification Team with 18 participants to identify the procedures currently in use for weather forecasts (NWS, 1982). Convening a committee, however, and then letting the committee do the work of developing procedures, takes a long time. For example, the NWS team convened in 1980 took two years to produce a final report. The need to establish verification methods for hydrologic forecasts is acute, and therefore,

the development of hydrologic verification must be *jump-started*. To do so, a set of procedures are proposed here with the hope that at some time in the future, the work of a committee of hydrologic forecast verification experts will supersede the suggestions put forth here.

The method used to develop the Standardized Evaluation Procedures proposed here is to first define the purposes of the verification procedures to be defined. With the purposes clearly defined, methodologies to meet those purposes are collected from well established meteorological methods, from existing hydrologic verification systems and from the experience of the verification completed in this thesis.

5.1.1 Purposes of these Standardized Evaluation Procedures

Forecasts may be verified for a variety of reasons which Brier and Allen (1951) sorted into three categories *Administrative*, *Scientific* and *Economic*. The purposes of the verification procedures to be described here are *Administrative* and *Scientific*. They are designed to provide all people who are interested in forecasting with a standard means of quantifying the characteristics of a forecast system. These quantified characteristics are intended to encompass all aspects of a forecast service with the goal of being useful to forecasters, to forecast managers, to forecast users and to forecast researchers.

The second purpose of the procedures described here is to provide a means of identifying the sources of error and skill in a set of forecasts. Analyzing the sources of forecast skill is a critical step in identifying implementation needs, identifying research needs and in identifying the value of past research and implementation efforts. The requirement to conduct analyses of these sorts was clearly identified in the results presented in Section 3, Administrative Verification of Deterministic River Stage Forecasts.

A third purpose of these procedures is to provide the operational hydrologic forecast community with a template for a baseline system that can be implemented to track the skill of their hydrologic forecasts. To that end, after this proposal has been peer reviewed, it will be used as the basis for the NWS hydrologic verification program implementation.

5.1.2 Evaluating Standardized Procedures

One reason a committee of experts have defined verification schemes in the past is it is very difficult to evaluate the quality of verification procedures using objective measures. The evaluation of the procedures is necessarily heuristic. For example, one identified benefit of Standardized Evaluation Procedures is improved communication, yet there is no clear method by which the procedures can be reviewed to determine if they will or will not improve communication. In the case of these procedures, the standard committee

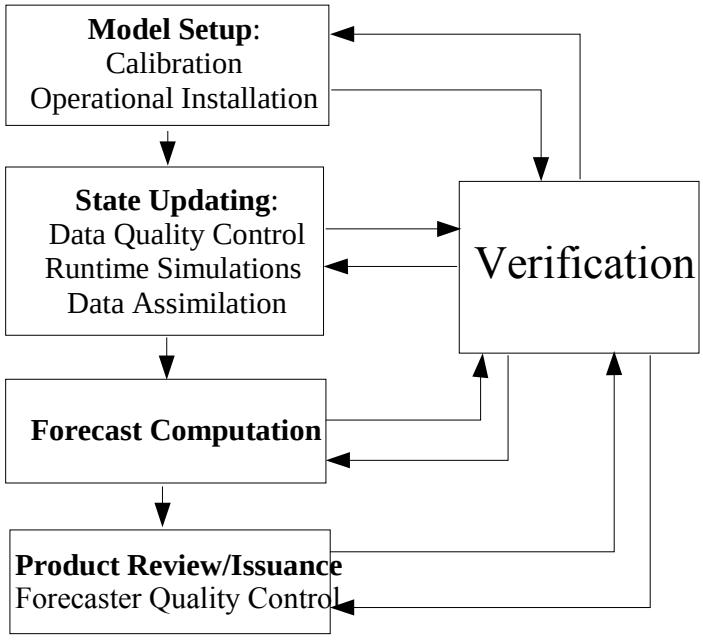

Figure 26. The role of verification in the forecast process.

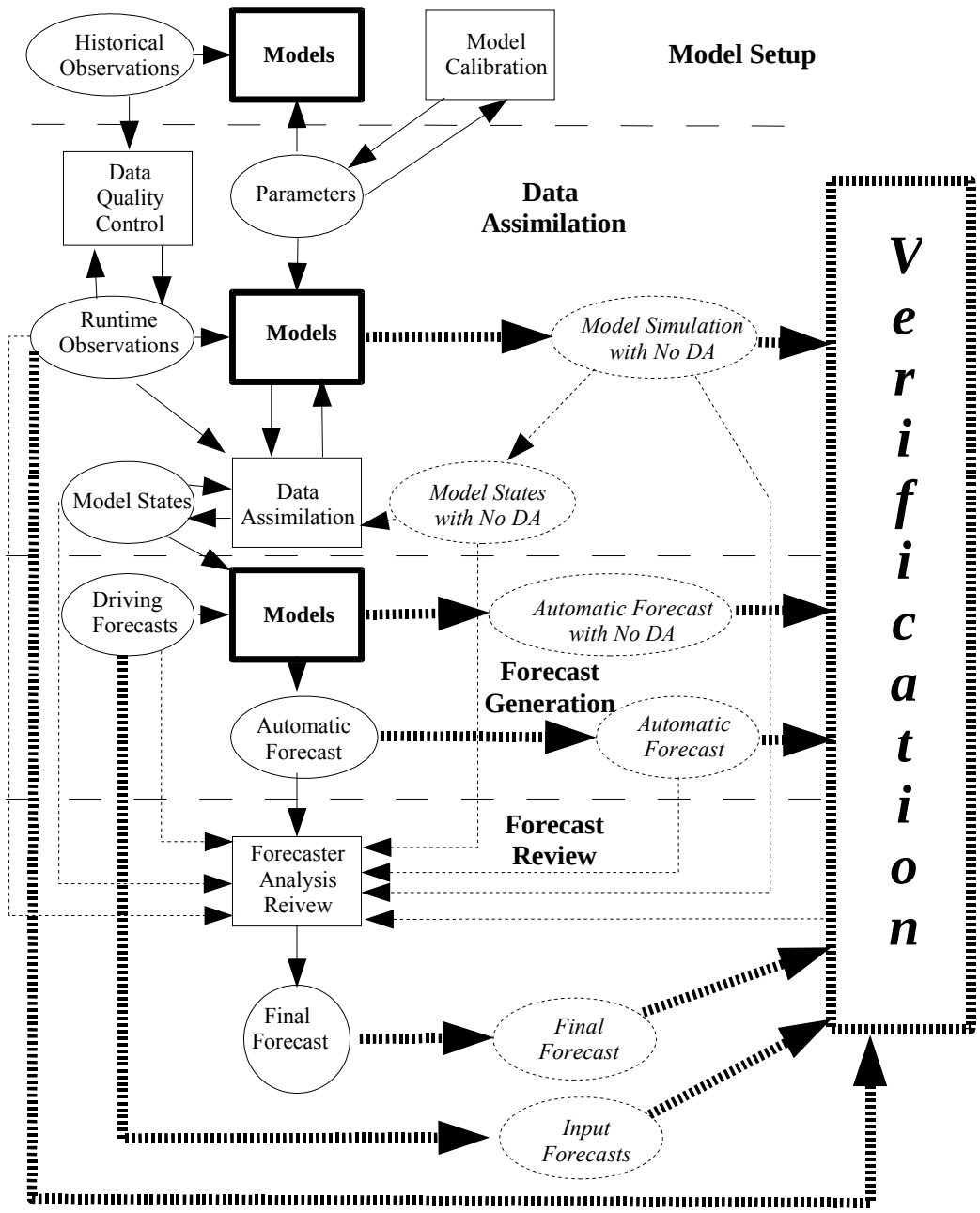

Figure 27. Detailed depiction of the role of verification in the forecast process.

work will be replaced by the peer review process and the evaluation of the procedures is completed by the collective agreement of the individual reviewers.

5.2 The Hydrologic Forecast Process

Verification procedures are shaped by the forecast process they evaluate. Therefore, prior to proposing verification standards, a generalized hydrologic forecast process is described. Although there can be many implementations of the hydrologic forecast process, all implementations can be described in terms of the following five steps outlined here: 1) model setup, 2) state updating (which includes runtime data quality control) 3) forecast computation using input forecasts, 4) product review and issuance, and 5) verification to close the forecasting loop. Figure 26 shows a general schematic of this process and Figure 27 shows a detailed schematic of this process.

Model setup is the process of selecting, parameterizing and linking a suite of models to simulate the hydrologic system to be forecast. Historical data is collected, quality controlled and corrected as needed. A set of simulation characteristics are selected (e.g. peaks, volumes, baseflow) and the models are tuned, by adjusting their parameters, to match the characteristics of the observed hydrograph record as well as possible. In some cases, the tuning is done manually (the NWS recommended method (NWS 2002b)), in other cases, it is done automatically using optimal search algorithms and objective functions.

State updating refers to the process by which the model states are adjusted to ensure the models have the best possible initial conditions with which to begin forecasting. To conduct the data assimilation, observations are collected, quality controlled and fed to the models during the observed period; the model parameters, the model states, and/or the input time series are then adjusted to make the model performance match the observations. The actual adjustments and search for the best fit to the observations may be done manually (as is done at the NWS) or automatically through statistical techniques.

Forecast data is then used to drive the models into the future to the desired lead time. A variety of input forecasts may be included depending upon the methods employed to model a basin's hydrologic system. In most cases Quantitative Precipitation Forecasts will be used, though temperatures, reservoir releases, lock and dam schedules, or upstream flows can be important. In some cases, the skill of the local hydrologic modeling will be overwhelmed by the skill (or non skill) of the input forecasts, and verifying the hydrologic forecasts becomes a matter of verifying the input forecasts. This phenomenon was seen in the role of the QPF in the hindcast skill presented in Section 4, Scientific Verification of Deterministic River Stage Forecasts.

Before issuing the forecast, a forecaster must review the model output and construct a forecast. The human quality control of the forecasts is critical to the forecast process as

computational procedures may arrive at un-realistic solutions to simulations of complex hydrologic systems. Ideally, this review should be made in the light of up-to-date verification metrics.

The fifth and final step in the forecast process is to evaluate the forecast performance. Although, each step in the forecast development process is assumed to contribute skill to the final forecast product, the contribution of the forecaster, the forecast models and the input forecasts to the forecast skill is not known, nor can it be assessed without a comprehensive verification system in place. The procedures proposed here provide a framework for a comprehensive verification system.

5.3 Proposed verification methods

An effective verification process must quantify the characteristics of the forecast system and offer a means to analyze why the forecasts behave as they do. The methods for characterizing and analyzing the forecasts are described in three parts: characterizing the logistical aspects of the forecasts, characterizing the skill of the forecasts, and analyzing the skill and error sources in the forecasts.

5.3.1 Logistical characterization of the forecast system

A forecast service is like a forecast, it cannot be described in terms of a single attribute (e.g. forecast skill). It is important to measure the non-skill attributes of the service

because service enhancements are often directed at improving the logistical properties of the service not the skill characteristics of the forecasts themselves. For example, adding forecast locations cannot be measured by improvements to forecast skill metrics. There are few published logistical metrics for hydrologic forecast systems. The implementation of Advanced Hydrologic Prediction Services (AHPS) by the NWS is an exception, as the primary tracking metric for the AHPS is the number of forecast locations where new Services have been implemented.

The purpose for collecting the logistical information is to answer questions like the following. What new types of forecasts have been developed? Is the number of forecast locations increasing or decreasing? Have computational improvements reduced the effort to issue a forecast? Have methodological improvements reduced the time it takes to prepare a basin for forecasting? To answer these types of questions, the folowing logistical measures are proposed:

1. the types of forecasts issued,
2. the frequency with which each type of forecast is issued,
3. the number of each type of forecast issued,
4. the locations at which each type of forecast is issued,
5. the person effort to setup a basin for forecasting, and
6. the person effort required to issue each type of forecast.

For any forecast service, these characteristics can be measured with simple counting metrics except for the person effort metrics (items 5 and 6). The person effort to setup a basin for forecasting (item 5) must be clearly defined if it is to be measured. The following definition is proposed: the effort required to collect historical data, quality control that data, select and set up models, calibrate the models and then install those models into the operational data stream. The NWS currently estimates this time to be 4 days for basins without reservoirs, and 10 days for basins with reservoirs.

The person effort required of a forecaster to issue a forecast (item 6) must also be defined if it is to be measured. The proposed definition is: the time it takes to quality control input data, run the models, quality control the model output, construct the forecasts, and make them available to the forecast users. It does not include the time it takes to generate the input driving forecasts nor the time it takes to setup the models. The effort required to issue a forecast varies considerably from simple no-rain, baseflow forecasts to flood events. In addition, forecasters are required to issue their forecasts on a set schedule and the time they have to spend on a basin may be limited by these schedules. Current NWS estimates for simple baseflow forecasts of deterministic forecasts are 1 minute per location, while the times for flood forecasts range between 5 and 30 minutes per location.

5.3.2 Characterizing forecast quality

Unlike measuring the logistical characteristics of forecast systems, measuring forecast quality characteristics is a much discussed problem. Alan Murphy described forecast "goodness" in terms of *consistency*, *value*, and *quality (*Murphy, 1993). By *consistency* he meant the consistency between the forecasters' best estimate of the future event and the forecast. V*alue* refers to the expense the users avoid or the benefits they accrue as a result of the forecasts. *Quality* refers to the correspondence between the forecasts and the events they forecast. The verification procedures described here address this last type of goodness, forecast *quality*. The proposed metrics and procedures are intended to provide the forecast verifier with information regarding the manner in which the forecasts correspond or do not correspond with the observations.

5.3.2.1 Persistence as a no-skill baseline forecast

A no-skill baseline forecast is essential to understanding the forecast metrics. This need was seen in the evaluation of the NWS forecasts presented in Section 3, as the persistence baseline was essential to understanding the values of the forecast skill metrics. A persistence forecast, where persistence is the observation at the forecast issuance time, is proposed as the no-skill baseline for these procedures.

5.3.2.2 Pairing the forecasts and observations

The first step in verifying forecasts, is to pair them with observations. In many cases the observations will not be reported for times which are identical to the forecast valid time, and the observations must be selected within some window of the forecast valid-time. The NWS national verification scheme uses ± one hour and the same window was used in the forecast assessment presented in Section 3. This window is proposed as the standard.

Forecasts for which there is no observation cannot be verified. The reverse, however, is not always true; observations for which there is no forecast may need to be verified. Within the NWS there is a class of forecast locations for which forecasts are issued only if flooding is imminent. These are called flood only locations. When a flood occurs, but no forecast was issued, a forecast placeholder must be inserted into the forecast observation pair [3]. One simple placeholder, which is proposed as a standard is the minimum value of the rating curve.

5.3.2.3 Sorting and aggregating the forecast observation pairs

One essential function in any verification system is sorting and aggregating the forecast-observation pairs into informative subsets. The most basic types of sorting or aggregating are by location and by time of issuance (month, season, year, etc.), and by lead-time. In addition to these types of sorting, it is proposed that any verification system support the

3 Cases where the forecast was issued but not properly archived may be exempted from this requirement.

diagnostic approach of Murphy and Winkler (1987) and offer functions to sort by the observed and forecast stage height. This simple sorting process has proved useful in the prior two Sections, and has been demonstrated to be useful in the wider meteorological community. More complex sorting procedures, such as by event and by the slope of the hydrograph are proposed as recommendations for more advanced verification systems.

When sorting by the stage height, the selection of the category boundaries will depend upon the purpose of the analysis. For regional administrative purposes, a single threshold like the NWS Flood Stage may be suitable. For analyzing sources of error, finer thresholds may be appropriate. No specific recommendations are proposed for the category boundaries except that they be clearly identified, and the means of selecting them be clearly identified.

One characteristic of basins currently in use by the NWS to aggregate metrics into informative collections is the river size. The basin response time is used to sort forecast locations into small, medium and large basins. As was seen in the review of past NWS forecasts (Section 3), the size of the river makes a considerable difference in the verification results. However, specific recommendations are proposed as requirements for sorting by the river size as this may not be appropriate in all circumstances.

5.3.2.4 Proposed metrics

The following list of metrics are proposed to characterize each subset. Short summaries explaining the reason for including each metric in this proposal are provided. The characteristics of these metrics are well described by others (e.g. Joliffe and Stephenson, 2004) and those descriptions are not repeated here.

1. The Probability of Detection (POD), the False Alarm Ratio (FAR), the Critical Success Index (CSI), the Pierce Skill Score (PSS) and the Gerrity Score (GS),
2. The Root Mean Square Error (RMSE), the Mean Absolute Error (MAE) and the Mean Error (ME),
3. The Pearson correlation coefficient (CC_P),
4. The Variance of the observations (OVAR) and the Variance of the forecasts (FVAR),
5. The Root Mean Squared Error persistence skill score ($SS\text{-}RMSE_{pers}$),
6. The Relative Operating Characteristics Area (A) and
7. The Sample Sizes for each set of forecast-observation pairs upon which the metrics are computed.

This list is not a comprehensive list of all useful verification metrics; it constitutes a proposed minimum set to be included in any hydrologic verification program.

The POD, the FAR and the CSI are recommended because they are used in most meteorological verification programs and so can provide a link to those existing programs. The POD measures the frequency with which an event was foreseen by the forecasts, while the FAR measures the frequency with which the forecasts warned of a non-existent event. The CSI merges information in the FAR and the POD. The PSS is an equitable skill score and can be used to compute the Gerrity score. The Gerrity score is useful for comparing multi-category contingency tables. These categorical metrics provide only coarse measures of the forecast skill because effective hydrologic forecasting requires more than forecasting a category correctly. It requires the correct river height be forecast as well.

The RMSE, the MAE, ME and the CC_P are standard metrics used for hydrologic model calibration and model development studies. River stage is a continuous variable and not a categorical variable, as noted above, so these four metrics provide an informative estimate of the expected error at a location or across a group of locations. Because it is dimensionless, the CCp offers some capacity to compare locations and times, but differences in the CCp may be caused by the changes in the forecast problem and not in changes in the capacity of a forecaster or a forecast system to forecast. By including the OVAR and the FVAR all the factors of the Mean Squared Error are included, and they can be used to determine what aspects of the forecasts and the observations are causing a change in the RMSE.

The SS-RMSE$_{pers}$ provides a summary of the relation between the persistence baseline and the actual forecasts. The SS-RMSE$_{pers}$ normalizes for the difficulty of the forecast problem to allow comparisons between locations and times. However, the SS-RMSE$_{pers}$ should always be reported in conjunction with the RMSE so that large errors in the persistence do not mask large errors in the actual forecasts leading to a false sense of success.

The ROC Area, A, provides an additional dimensionless measure of skill. It is included because it can also be computed for probability forecasts, and thus provide a link to allow comparisons between single- and multi-valued forecasts. Sample Size is included to provide information regarding the sampling error in the metrics. Explicitly computed confidence intervals for each metric are also proposed and a method for computing them is described below.

5.3.2.5 Computing confidence intervals

When metrics are computed, it is important to compute confidence intervals to assist those interpreting the metrics. Reporting the sample size is one simple way to provide the users with information about the likely significance of the metrics. Confidence intervals offer more explicit estimates of the uncertainty in the metrics. However, computing the confidence intervals is itself an uncertain process for two reasons: one, there is strong

serial correlation in the samples, and two, analytic methods are not available for all the metrics considered here. To address the problem of serial correlation, Livezey (1999) recommends re-sampling the original set of forecast-observation pairs to construct uncorrelated sub-samples. A correlation length can be computed as described by Trenberth (1984) and by Thiebaux and Zweirs (1984) and the original time-series can then re-sampled at time-steps longer than this correlation length to create uncorrelated sub-samples. Where analytic techniques are not available for computing confidence intervals, one alternative is to use Monte Carlo bootstrapping techniques. A methodology to compute confidence intervals which employs the re-sampling and then bootstrap techniques is proposed below.

The correlation length can be computed from the following equation: (Livezey, 1999).

$$T_o = 1 + 2\sum_{\Delta=1}^{N}(1-\frac{\Delta}{n})\rho_\Delta$$

T_0 is the effective correlation length,

N is the maximum number of lags,

Δ is the lag, delta ≤ N,

n is the sample size, and

$\rho\,\Delta$ is the correlation at lag Δ .

When this equation is applied to the Watts, OK dataset introduced in Sections 3 and 4, the correlation length(T_o) for the observations is 30 days and for the forecasts is 35 days

when computed with six hour river stage time series. For the Saint Charles, MO forecast location, the correlation length is 235 days for the forecasts and the observations when computed with the daily river stage observations and forecasts from the Missouri mainstem. A correlation length of 235 days is too long to be useful, unless a very long time series is to be analyzed. Even 30 and 35 days are long if you want to analyze a year of forecasts Re-sampling a year long time series at 30 or 35 day intervals results in sample sizes on the order of 12 samples per uncorrelated sub-sample. One solution is to re-sample at the largest possible distance and still have sample sizes in the sub samples on the order of 30. When sampling intervals are used that do not meet the independence requirement, they should be identified.

This resampling to construct uncorrelated sub-samples can be done on the entire original sample set, or upon the original sorted subsets of the larger sample set. That is, the re-sampling to create uncorrelated sub-samples can be done before or after the forecast-observation pairs are sorted into stage height categories. However, the correlation structure of the sorted subsets of the forecast observation pairs (e.g. above and below Flood Stage subsets) after they have been sorted will be more complex than the correlation structure in the original sample because the sorted subsets consist of short time-series from disconnected periods of time. Until further evidence indicates an alternate route, it seems most reasonable that the re-sampling to create the uncorrelated

sub-samples should be done on the entire original sample and then the sub-setting by stage height done on the uncorrelated sub-samples.

The proposed bootstrapping process then consists of first constructing uncorrelated sub-samples of pairs from the original sample set. Each uncorrelated sub-sample is then re-sampled with replacement to construct numerous additional re-sampled sets. Each bootstrap generated re-sampled set is then sorted into subsets (by stage height, etc) and the metrics of interest are computed. Distributions of the metrics can then be constructed and the confidence intervals extracted from the distribution based upon some probability range of interest. With the assumption that all of the computed metrics are independent estimates of the actual value, all the computed metrics from the bootstrap generated re-sampled sets from all the independent sub-samples should be used collectively in estimating the distribution.

The importance of using the uncorrelated sub-samples as opposed to the original correlated sample to compute the confidence intervals is demonstrated with data from Watts, OK basin. The effective correlation length (T_0) for the forecast time series has been reported above as 30 days. The Day 3 MAE was computed for the entire period of record from the entire set of forecasts and observations. Confidence intervals (95%) were then computed using standard bootstrap techniques applied to the entire original correlated sample set. Confidence intervals were then computed using the un-correlated

sub-samples as described above. These values are reported in Table 11, and as can be seen there, when the correlation in the sample set is not addressed, the estimated confidence intervals are much smaller than the confidence intervals when the confidence intervals are constructed on uncorrelated samples. It is worth noting however, that the sample sizes for the high stages in the uncorrelated sub-samples are very small (see Table 12). This is one limitation of this approach. These example confidence intervals were computed using 1200 re-samplings in the bootstrap experiments, following the suggestion of Livezey (1999). The number of re-samplings required to correctly estimate the distribution is discussed next.

This example data set is also used to confirm Livezey's (1999) suggestion that 1000 re-samplings in the bootstrap process are sufficient to estimate the distribution of the metric. To conduct this assessment, the confidence intervals are recomputed with 120, 600, 1200, and 6000 resamples. The plots for the high stage and the low stage distributions are presented in Figure 28. These plots show that with 600 resamples the distribution is reasonably well described for the low stages as the distribution is smooth and it does not change when 6000 resamples are used. The high stage distribution does not change between the 600 and 6000 re-samples either. However, even with 6000 samples the distribution is not smooth. This is a result of the small sample sizes for the high stages.

MAE for Day 3 at Watts, OK.	<FS Lower 95% Limit	<FS Upper 95% Limit	>FS Lower 95% Limit	>FS Upper 95% Limit
Sample Values	0.48		12.5	
Original correlated sample set	0.46	0.49	11.9	13.0
Uncorrelated sub-sample sets	0.29	0.80	8.0	15.8

Table 11: The confidence intervals computed for the MAE using several bootstrap experiments.

SAMPLE SIZES	<FS Smallest	<FS Largest	>FS Smallest	>FS Largest
Original Sample	13124		76	
Uncorrelated sub-sample sets	107	110	1	3

Table 12: The sample sizes for the bootstrap experiments.

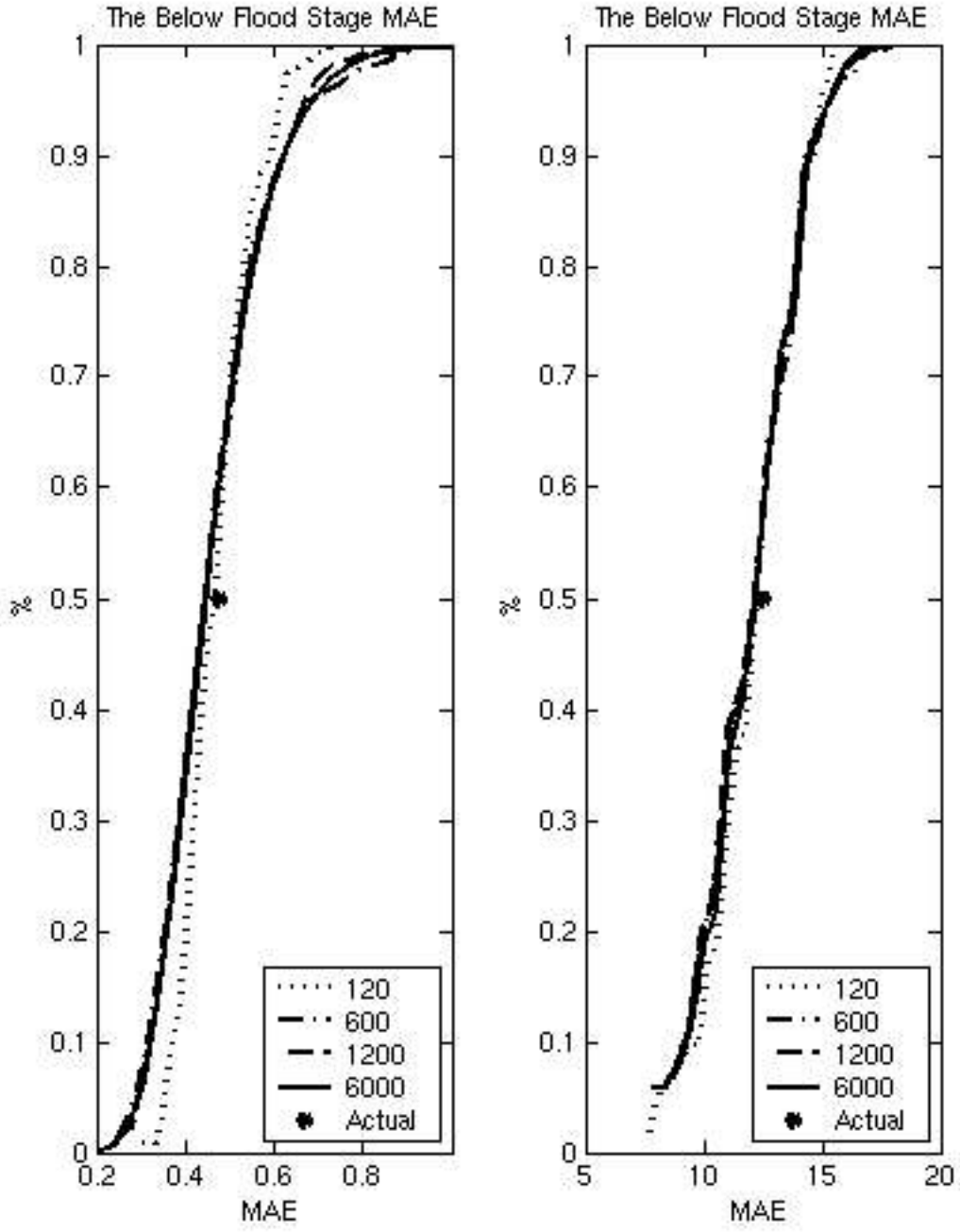

Figure 28. Comparison of the distributions with different numbers of bootstrap resamples.

5.3.3 Forecast system error analysis

The third element of a comprehensive verification system is analysis of the the forecast process to determine which elements of the forecast process provide skill and which introduce error. The procedures proposed here follow the method described in Section 4, Scientific Verification of Deterministic River Stage Forecasts. They provide a means to partition the skill and error between the four forecast process steps outlined earlier. As the hydrologic forecast process becomes better understood, it is expected these procedures will be updated. Figures 26 and 27 provide diagrams of the forecast process with the runtime verification providing an objective framework to support analyses of the forecast process.

The analysis process consists of comparing carefully selected sets of forecasts to determine how each step in the forecast process affects the forecast skill. The information to be gathered from the comparisons depends upon the differences between the sources of the forecasts in the comparison. The notion of one set of forecasts being *sufficient* (Degroot and Feinberg, 1982; Ehrendorfer and Murphy, 1988) for another offers an appealing framework for comparisons. However, Murphy (1997) points out that one set of forecasts is rarely *sufficient* for another, and therefore the *sufficiency* relation has limited utility. In lieu of the *sufficiency* criteria, the suite of metrics described earlier can be used to characterize each subset to construct a comprehensive view of the differences between forecast sets.

A parsimonious set of intermediate output from the forecast steps must be collected to provide objective control on the forecast process without creating a burden on the forecasters in maintaining an overly complex system. The model setup is the first step to implementing a forecast system; it is the foundation upon which all other forecast procedures are built. In order to provide a continuous monitoring of the model setup, a simulation which does not include state updating must be computed. This simulation[4] can also serve as a useful reference for comparisons with forecasts to assess the error the input forecasts are inserting into the forecasts. This control forecast is labeled "*Model Simulation with no DA*" in Figure 27. The second major source of skill in hydrologic forecasts is the initial conditions. Collecting a set of forecasts which do not include state updating and comparing them to the forecasts computed with updated states allows the forecast verifier to determine the skill integrated into the forecasts by the updating process. In Figure 27, this control forecast is labeled "*Automatic Forecast with no DA*".

Understanding the error contributed by the input forecasts requires two steps. First, the input forecasts themselves should be verified by comparing them to their corresponding observations. Second, by comparing the forecast with no state updating (*Automatic Forecast with no DA*) to the simulation with no state updating (*Model Simulation with no DA*) it is possible to determine the error contributed to the forecasts by the input forecasts. To assess the skill contributed to the forecasts by the input forecasts, a reasonable

4 Simulation means the models are driven by observations.

alternative to the input forecast must be selected and the the forecasts computed with this alternative, for example, comparing forecasts computed with modeled precipitation forecasts to forecasts that use zero for the QPF. In basins where multiple input forecasts drive the models, the permutations of possible alternate forecasts becomes large and hindcasting experiments are more suitable than runtime controls. The final step in the forecast development process is the forecaster quality control of the forecasts. The improvement to the forecasts provided by the forecaster can be evaluated by comparing the "*Automatic Forecasts*" (so labeled in Figure 27) with the Final Forecasts. (In the case of the NWS, the automatic and the final forecasts are almost identical because the forecasters have manually updated the forecasts as a part of the data assimilation process.)

In addition to runtime control forecasts, an important component of a complete system for analyzing the sources of error in the forecasts is hindcasting. At this early stage in the development of the hydrologic forecast process, it is essential the data required to conduct hindcast experiments be collected. The depth of the hindcasting one hopes to conduct will determine the detail of the data storage requirements. At a minimum, it is proposed that the input forecasts and observations must be stored in a manner which provides for easy retrieval and formatting for hindcasting.

5.4 Communication of Evaluation Results

Verification metrics are like forecasts; they are only valuable if they are used. Publication of verification metrics is as critical as computing them: publication to the general public, to forecasters, to administrators and to researchers. Each group has different requirements for the type of information they need. For the administrators, summary reports describing logistical characteristics, skill trends, and current skill status aggregated over large areas will be required. Users need forecast skill information for individual locations, though sample sizes may limit the information which can be provided. The researcher or forecaster require detailed analyses of forecast skill and the data to permit research into the forecast process. These data requirements include the forecasts and observations to conduct additional verification analyses as well as the data to support hindcasting. As the hydrologic forecast process matures, it is expected the information needs of the managers, users, and researchers will change and the descriptions of the report contents provided below will be updated.

The needs of particular administrators or users must be assessed on an individual basis. Therefore three general requirements are proposed as standards for communicating verification results. The first, proposed requirement is the verification results be publicly disseminated at regular intervals. One benefit of standardizing procedures is to enhance communication. This will only happen if the communication is permitted to occur. The second, proposed requirement is the forecasts and observations be made publicly

available to encourage detailed research into the forecast characteristics. This will encourage the growth of a robust understanding of the hydrologic forecast skill. The third proposed requirement is that sufficient data be made publicly available to support hindcasting studies. If the hydrologic forecast process is to develop, hindcasting studies are essential.

5.5 Conclusions for Standardized Procedures

In order to promote the development of the hydrologic forecast process, standardized verification procedures for deterministic hydrologic forecasts have been proposed. Clearly defined verification procedures will facilitate communication between disparate groups working on forecast methodologies through concise, well understood descriptions of forecast skill. If similar verification processes are used by hydrologic forecasters, then it will be possible to describe advances in the science of hydrologic forecasting and have them understood across the globe. Without a common language of verification, it will be difficult to identify that a forecasting innovation employed in say Bangladesh, is likely to help forecasts in say the US.

This proposal is intended to serve as the basis for a discussion on appropriate methods for hydrologic forecast verification. It is not intended to serve as the final word on hydrologic forecast verification. This author looks forward to seeing a robust discussion of hydrologic forecast verification develop in all hydrology related publications. A

comprehensive verification program will enhance the ability of the research community to support the development of the hydrologic forecast process.

6 SUMMARY OF CONTRIBUTION TO HYDROLOGY

This work contributes to the science and practice of hydrology in three ways. First this work identifies an important hydrologic problem, which is that little is objectively known about the skill of hydrologic forecasts. This point was demonstrated through a verification analysis of past NWS forecasts on the MM and A/O datasets. Two findings from that study (presented in Section 3) were that hydrologic forecast skill appear not to have improved over the past decades, and forecasts for high stages (e.g. floods) have little skill beyond Day 1. These results are new and have never been previously documented indicating that even the most basic characterization of the hydrologic forecast skill is not being carried out. As a result, it seems that the hydrologic research and operations community have been operating under the assumption that the forecasts were skillful and that we had been making progressive improvements to forecast skill. The fact that neither of these assumptions appears to be true leads to the conclusion that hydrologists know little about the skill of their forecasts.

The second contribution this work makes to the science and practice of hydrology is to propose simply that hydrologists adopt an objective process for verifying forecasts. A review of the literature. While this very simple suggestion appears trivial, it is a new concept for hydrologists to consider. As described in the review of the existing literature and in the review of existing operational procedures (presented in Section 2), as well as discussions with hydrologic scientists and practitioners suggest that this seemingly trivial

suggestion is a navel concept for hydrologists since verification is not actively practiced within the hydrology community. In fact, quite the contrary is pre-supposed as was noted by Joliffe and Stephenson (2004) when they described hydrologic forecasts as being too hard to verify. The work presented in this dissertation demonstrates that this latter presumption is false. Forecasts can be assessed as was done in the review of NWS forecasts, the sources of skill and error can be identified as was done in the hindcasting study (Section 4), and it is possible to define Standard Verification procedures to support the entire river forecast process as was shown in Section 5. Knowing the skill characteristics of the forecasts through verification will help to focus the development of new procedures towards improvement of forecast skill, as was shown in the way the results of the hindcast analysis were used to identify the roles of the three forecast process elements. An initial methodology for improving the forecast skill is proposed here – through the use of verification metrics to drive the research and implementation efforts aimed at improving the forecasts. At present the selection of both the research and the implementation projects directed toward improving hydrologic forecasts is primarily driven and evaluated by the opinions of experts. However, beliefs, even good ones, do not form a solid foundation for scientific development. Instead, it is proposed here that the development of the hydrologic forecast process be based upon an objective approach using documented verification metrics.

In addition, simply knowing the skill characteristics of the forecasts is useful in and of it itself, and can lead to more skillful forecasts because, as was noted in the Introduction (Section 1), Krzysztofowicz and Sigrest (1999a) found that providing forecasters with objective verification metrics improved their forecasts.

The third contribution of this work to the science and practice of hydrology is to initiate the development of hydrologic verification procedures by proposing a method for forecast error analysis, and standards for operational agencies to follow in implementing verification programs. If hydrologists are to be successful verifying their forecasts, they need to establish methods by which the forecasts can be evaluated to determine why they perform as they do. An approach to forecast error analysis is proposed in Section 4 of this dissertation. Another key to successful verification is to ensure that the operational agencies issuing the forecasts conduct their verification using well documented and scientifically valid methods. When the decisions for updating the forecast process follow the guidance of objective verification metrics, new science will be more objectively judged by its ability to improve operational skill. The metrics computed by the operational agencies will become central to the communications between the research and operations communities. Therefore, the operational verification procedures must follow well understood and accepted standard procedures.

Ironically, it is not possible to use objective measures to determine a-priori whether or not the proposed method for improving the hydrologic forecasts will or will not succeed. (In time, the verification metrics will answer this question.) However, the proposed method is based on the model used by the meteorological community which has worked well for them[5]. Given that the current approaches have not succeeded; given that there are numerous interesting forecast problems that can lead to productive research; given that the forecast process is a scientific endeavor and the proposed method creates a scientific process based on objective data; given that the meteorologists have succeeded with the proposed approach; and given that the pressure to measure all forecast performance is growing, it seems reasonable to insist that hydrologists at least begin to verify their forecasts and consider verification metrics in their analyses of hydrologic forecast performance. By focusing hydrologic research on the task of verifying itself, a new branch of the hydrologic discipline, an *Hydrologic Forecast Science*[6], can be developed to support the complex task of providing forecasts for the nations waterways.

5 I found an interesting (but unpublished) document in the NOAA Library here in Silver Spring entitled "Selecting Research Problems for Forecast Improvement" written by Charles Lester Bristor in 1952. He makes the same suggestion for meteorologists.

6 The term, Hydrologic Forecast Science was first used by D.J. Seo in his presentation *Toward Improved Hydrologic Forecasts* at the Office of Hydrologic Development, in Silver Spring, MD in November of 2003.

REFERENCES

Ahnert, P., M. Hudlow, E. Johnson, D. Greene, and M. Rosa Dias, 1983: Proposed "on-site" precipitation processing system for NEXRAD. Preprints, 21st Conf. on Radar Meteor., Edmonton, AB, Canada, Amer. Meteor. Soc., 378-385.

Ahnert, P., W. Krajewski, and E. Johnson, 1986: Kalman filter estimation of radar-rainfall field bias. Preprints, 23rd Conf. On Radar Meteor., Snowmass, CO, Amer. Meteor. Soc., JP33-JP37.

Anderson, E.A., 1973, National Weather Service River Forecast System - Snow Accumulation and Ablation Model, NOAA Technical Memorandum NWS HYDRO-17, U.S. Dept. of Commerce, Silver Spring, MD.

Arkansas Red River Basin River Forecast Center (ABRFC), 2004: ABRFC Verification, RMS Error, at http://www.srh.noaa.gov/abrfc/fcstver/images/versum1.gif

Breidenbach, J. P., D.-J. Seo, P. Tilles, and K. Roy, 1999: Accounting for radar beam blockage patterns in radar-derived precipitation mosaics for River Forecast Centers, Preprints, 15th Conf. on IIPS, Amer. Meteorol. Soc., 5.22, Dallas, TX.

Brier, G.W. and R.A. Allen, 1951: Verification of Weather Forecasts. *Compendium of Meteorology*, T.F. Malone, Ed., Amer. Meteor. Soc., 841-848.

Brooks, H.E., A. Witt and M.D. Eilts, 1997: Verification of public weather forecasts available via the media, *Bulletin of the American Meteorological Society*, **78**, 2167-2177.

Burnash, R.J.C. and R.L. Ferral, and R.A. McGuire, 1973: A Generalized Streamflow Simulation System - Conceptual Modeling for Digital Computers, Joint Federal-State River Forecast Center, Sacramento, California, 204 pp.

Deque, M., 2003: Continuous Variables. *Forecast Verification, A Practitioners Guide in Atmospheric Science*. I.T. Joliffe, and D.B. Stephenson, Ed. Wiley, 97-119

Degroot, M. and S. Fienberg, 1983: The comparison and evaluation of forecasters, *The Statistician*, 32, 12-22

Ehrendorfer, M. and A. Murphy, 1988: Comparative evaluation of weather forecasting systems: sufficiency, quality and accuracy. *Monthly Weather Review*, 1757-1770

Fritsch, J.M. and R.E Carbone, 2004: Improving quantitative precipitation forecasts in the warm seasons, a USWRP research and development strategy, *Bulletin of the American Meteorological Society*, 955-965.

Finley, J.P., 1884: Tornado Predictions. *American Meteorological Journal*, **1**, 85-88

Franz, K.J., H.C. Hartmann, S. Sorooshian and R. Bales, 2003: Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin. *Journal of Hydrometeorology*. **4**, 1105-1118.

Gandin, L.S. and A.H. Murphy, 1992: Equitable scores for categorical forecasts. *Monthly Weather Review*, **120**, 361-370.

Gilbert, G.K., 1884: Finley's tornado predictions, American Meteorological Journal, Vol 1, 166-172

Glahn, Bob, 2004: Discussion of Verification Concepts in Forecast Verification: a Practitioner's Guide in Atmospheric Sciences, *Weather and Forecasting*, Vol. 19 pp 769-775.

Hartmann, H.C., T. Pagano, S. Sorooshian, R. Bales, 2002, Confidence builders, evaluating seasonal climates forecasts for user perspectives, Bulletin of the American Meteorological Society, Vol. 83, No. 5, pp. 683–698.

Hudlow, M. D., 1988: Technological developments in real-time operational hydrologic forecasting in the United States. *J. Hydrol*., 102, 69-92.

Joliffe, I.T. and D.B. Stephenson, 2003: Introduction. *Forecast Verification, A Practitioners Guide in Atmospheric Science*. I.T. Joliffe, and D.B. Stephenson, Ed. Wiley.

Koren V., M. Smith, Q. Duan, 2003: Use of a priori parameters estimate in the derivation of spatially consistent parameters sets of rainfall-runoff models, *in Calibration of Watershed Models* Q. Duan, H. Gupta, S. Sorooshian, A. Rouseau, R. Turcotte, American Geophysical Union, Washington DC

Krzysztofowicz, R. and A.A. Sigrest, 1999a: Calibration of probabilistic quantitative precipitation forecasts, *Weather and Forecasting*, **14**, 427-442

Krzysztofowicz, R., 1999b: Bayesian theory of probabilistic forecasting via a deterministic hydrologic model, *Water Resources Research*, **35**, 2739-2750

Krzysztofowicz, R. and H. Herr, 2001: Hydrologic uncertainty processor for probabilistic river stage forecasting: precipitation dependent model, *Journal of Hydrology*, **249**, 46-68.

Krzysztofowicz, R. and C.J. Maranzano, 2004: Hydrologic uncertainty processor for probabilistic stage transition forecasting, *Journal of Hydrology*, **293**, 57-73.

Linsley, R.K., M.A. Kohler and J.L.H Paulhus, 1975: *Hydrology for Engineers*, McGraw, New York

Livezey, R.E. 1999: Field Intercomparison, *in Analysis of Climate Variability*, H. von Storch and A. Navarra, eds. Springer-Verlag, Berlin

Livezey, R.E. and S. W. Jamison, 1977: A skill analysis of soviet seasonal weather forecasts, *Monthly Weather Review*, Vol 105, No. 12, pp 1491-1500

Livezey, R.E., 2003: Categorical Events. *Forecast Verification, A Practitioners Guide in Atmospheric Science*. I.T. Joliffe, and D.B. Stephenson, Ed. Wiley, 77-96

Mason, I., 1980: Decision-theoretic evaluation of probabilistic predictions (using the relative operating characteristic). *Proc. WMO Symp. on Probabilistic and Statistical Methods in Weather Forecasting*, Nice, France, WMO, 219-228.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Meteor. Mag.*, **30**, 291-303

Mason, I.B., 2003: Binary Events. *Forecast Verification, A Practitioners Guide in Atmospheric Science*. I.T. Joliffe, and D.B. Stephenson, Ed. Wiley, 37-76

Morris, D., 1988: A Categorical, Event Oriented, Flood Forecast Verification System for National Weather Service Hydrology, *National Oceanographic and Atmospheric Administration Technical Memorandum, National Weather Service, HYDRO 43*. pp. 74.

McDonald, B. E., T. M. Graziano, and C. K. Kluepfel, 2000: The NWS National QPF Verification Program. Preprints, 15th Conference on Hydrology, Long Beach, CA, January 9-14, American Meteorological Society, p. 247-250.

Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, **8**, 281-293.

Murphy, A.H., 1996: The Finley Affair: A Signal Event in the History of Forecast Verification. *Monthly Weather Review*, **14**, 3-20.

Murphy, A.H., 1997: Forecast Verification, in Economic value of weather and climate forecasts, R. Katz and A. Murphy eds. Cambridge University Press, New York.

Murphy, A.H., B. Brown and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Weather and Forecasting*, **2**, 243-251.

Murphy, A.H. and R.L. Winkler: 1987, A general framework for forecast verification. *Monthly Weather Review,* **115**, 1330-1338

National Precipitation Verification Unit (NPVU) 2004: Web page with interactive access to verification statistics, http://www.hpc.ncep.noaa.gov/npvu/index.shtml.

National Research Council, (NRC) 1996: Assessment of hydrologic and hydrometeorological operations and services, National Weather Service Modernization Committee,National Research Council, National Academy Press, Washington D.C.

National Weather Service (NWS), 1939: Forecast Verification. Report to the Chief of the Weather Bureau, National Weather Service, Silver Spring, MD.

National Weather Service (NWS), 1982: National Verification Plan, Report of the National Verification Task Team, National Weather Service, Silver Spring, MD.

National Weather Service (NWS), 1999: Quantitative precipitation forecast process assessment final report, National Weather Service, Silver Spring, MD.

National Weather Service (NWS) 2001a, National Weather Service Operations Manual, National Weather Service Silver Spring, MD.

National Weather Service (NWS), 2001b: Service Hydrologists Manual, Glossary, National Weather Service, Silver Spring, MD

National Weather Service (NWS), 2002a: Adjust-Q, Adjust Simulated Discharge Operation, *National Weather Service River Forecast System Manual* **V.3.3**. National Weather Service, Silver Spring, MD.

National Weather Service (NWS), 2002b: Calibration of Conceptual Models for Use in River Forecasting, National Weather Service, Silver Spring, MD.

National Weather Service (NWS), 2003a: Forecast Component Operations, *National Weather Service River Forecast System Manual* **V.3.2**. National Weather Service, Silver Spring, MD.

National Weather Service (NWS), 2003b: Calibration System Mean Areal Potential Evaporation Computational Procedure, *National Weather Service River Forecast System Manual* **II.5**. National Weather Service, Silver Spring, MD.

National Weather Service (NWS), 2003c: The Kansas city antecedent precipitation index model, *National Weather Service River Forecast System Manual,* **V.3.3**. National Weather Service, Silver Spring, MD.

National Weather Service (NWS), 2003d: The National Weather Service Annual Operating Plan for 2004

Nurmi, P., M. Heiskanen, M. Frisk, 2004: Forecast Verification Report, Summer 2004, Finish Meteorological Institute, pp 48.

Olsen, B. and W. Lawrence, 2001: Southern region approach to verification, *National Weather Service RFC River Forecast Verification Workshop,* February, 2001, Silver Spring, MD

Pagano, T., D. Garen, S. Sorooshian: 2004 Evaluation of Official Western U.S. Seasonal Water Supply Outlooks, 1922–2002 *Journal of Hydrometeorology* Vol 5, pp 896-909

Schwein, N, 1996: The effect of quantitative precipitation forecasts on river forecasts, National Oceanic and Atmospheric Administration, Technical Memorandum, NWS CR-110, pp 39.

Seo, D.-J., 1998, Real-time estimation of rainfall fields using radar rainfall and rain gauge data. *J. Hydrol*., 208, 37-52.

Seo, D.-J., and J. P. Breidenbach, 2002: Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements, *Journal of Hydrometeorology:* Vol. 3, No. 2, pp. 93–111

Seo, D.-J., V. Koren, and N. Cajina, 2003: Real-time variational assimilation of hydrologic and hydrometeorological data into operational hydrologic forecasting *J. Hydrometeorology*, 4, 627-641.

Smith, M., D.-J. Seo, V. Koren, S. Reed, Z. Zhang, Q. Duan, F. Moreda, S. Cong, 2004: The distributed model intercomparison project (DMIP): motivation and experiment design, *J. Hydrology*, 298, 4-26.

Stephenson, David B. and I.T. Joliffe, 2004: Forecast verification: past, present and future *Forecast Verification, A Practitioners Guide in Atmospheric Science.* I.T. Joliffe, and D.B. Stephenson, Ed. Wiley, 189-201.

Stanski, H.R., L.J. Wilson, and W.R. Burrows, 1989: *Survey of common verification methods in meteorology*. World Weather Watch Tech. Rept. No.8, WMO/TD No.358, WMO, Geneva, 114 pp.

Thiebaux, H. and F. Zweirs, 1984: The interpretation and estimation of effective sample size, *Journal of Climate and Applied Meteorology*, 23 800-811

Trenberth, K., 1984: Some effects of finite sample size and persistence on meteorological statistics. Part 1: autocorrelation. *Monthly Weather Review*, 112 2359-2368

Vivoni, E.R., D. Entekhabi, R.L. Bras, V.Y. Ivanov, M.P. Van Horne, C. Grassotti and R.N. Hoffman, 2003: Quantitative flood forecasts using short-term radar nowcasting, *17th Conference on Hydrology, 83rd American Meteorological Society Annual Meeting,* Seattle WA, Feb. 2003.

Werner, K., D. Brandon, M. Clark, S. Gangopadhyay, 2004: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts, *Journal of Hydrometeorolog*, Vol5, pp 1076-1089.

Wilks, D.S. 1995. *Forecast Verification,* Section 7 in Statistical Methods in Atmospheric Sciences. Academic Press, pp 233-283.

Wobus, R.L. and E. Kalnay, 1995: Three years of operational prediction of forecast skill at NMC, Monthly Weather Review, Vol 123, pp 2132-2148.

Wood, E and D. Lettenmaier, 2002: An experimental operational West Wide water supply forecast system, Journal of Geophysical Research.

World Meteorological Organization (WMO), 1994: Guide to hydrological practices, fifth edition, publication no. 68, World Meteorological Organization, Geneva Switzerland.

World Meteorological Organization (WMO), 2002: Standardised Verification System for Long-Range Forecasts, attachment II-9 to the *Manual on the GDPS* (WMO-No. 485), Volume I, World Meteorological Organization, Geneva Switzerland.

World Meteorological Organization (WMO), 2004, WWRP/WGNE Joint Working Group on Verification *Forecast Verification - Issues, Methods and FAQ, web site,* *http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html*

Young, C. B., A. A. Bradley, W. F. Krajewski and A. Kruger, 2000: Evaluating NEXRAD multisensor precipitation estimates for operational hydrologic forecasting. J. Hydrometeor., 1, 241-254.