Scientific Verification of Deterministic River Stage Forecasts

EDWIN WELLES

Systems Engineering Center, National Weather Service, Silver Spring, Maryland

SOROOSH SOROOSHIAN

Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California

(Manuscript received 19 March 2008, in final form 12 September 2008)

ABSTRACT

One element of a complete verification system is the ability to determine why forecasts behave as they do. This paper describes and demonstrates an operationally feasible method for conducting this type of diagnostic verification analysis. Hindcasts are generated using different configurations of the forecast system and then the skill of the generated hindcasts is compared. The hindcasts and comparisons are constructed to isolate individual elements of the forecast process. The approach is used to evaluate the role of model calibration, model initial conditions, and precipitation forecasts in generating skill for deterministic river forecasts. The authors find that calibration and initial conditions provide skill for the short lead-time forecasts, with precipitation forecasts providing the majority of the skill in forecasts of high stages at longer lead times. At all lead times, this study shows model calibration is essential, as the calibration makes forecasts reliable.

1. Introduction

Recently, Welles et al. (2007) evaluated National Weather Service (NWS) river stage forecasts. They found the forecast skill may not have improved as much as expected because, as they suggested, forecast system updates were not driven by objective measures of forecast skill. Many people have studied elements of the forecast process—calibration, state updating, and precipitation forecasts—but the forecast process itself with the various elements linked together has not been studied. This paper presents a hindcasting experiment used to analyze stage forecasts that illustrates a systematic method for using the distributions-oriented verification of Murphy and Winkler (1987) to identify sources of forecast skill.

Standard meteorological verification metrics are applied to a set of hindcasts to address the following questions: What is the primary source of skill in the hindcasts at each lead time?[1] What is the role of calibration, initial conditions, and quantitative precipitation forecasts (QPFs) 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? This study focuses on precipitation-driven headwater basins with forecast lead times up to only three days. Numerous similar studies on downstream forecast locations, snow-covered basins, reservoir outflow points and the like will be required to build a robust understanding of hydrologic forecast skill and the associated uncertainties.

It is worth noting that in this article, we approach the hydrologic forecast problem from the perspective of the forecast process itself, as was done recently in Shi et al. (2008). Most studies aimed at improving hydrologic

[1] Lead time is the difference between the time a forecast is issued (the forecast basis time) and the time that forecast is valid (the forecast valid time). NWS-issued hydrologic warnings are issued from model output. One way to increase the warning lead time is to make the models more accurate at longer lead times. Therefore, it is important to understand how the skill characteristics of the models change with lead time.

Corresponding author address: Dr. Edwin Welles, Deltares USA, Inc., 1010 Wayne Ave., Suite 800, Silver Spring, MD 20910. E-mail: edwin.welles@deltares-usa.us

DOI: 10.1175/2008JHM1022.1

forecasts focus on analyzing and modeling of basin processes, taking the processes as the object of study. In this research, the forecast process itself is the object of study. By focusing on the forecast process itself, a new line of inquiry for improving hydrologic forecasts can be opened. In fact, current assumptions regarding the most effective methods for improving hydrologic forecasts can be evaluated. For example, the assumption that improved hydrologic modeling will always improve forecast skill can be validated. A review of this assumption is provided in section 4.

a. Error and skill in hydrologic forecasts

The uncertainty in hydrologic forecasts is traditionally divided into two categories: *meteorological* error and *hydrologic* error. Meteorological error refers to the error in hydrologic forecasts caused by error in the meteorological forecasts. Hydrologic error consists of the errors caused by the hydrologic modeling. This study focuses on the meteorological error resulting from 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 (refer to National Research Council 2006) when evaluated at the short modeling time steps and the fine spatial scales used for hydrologic models, even if those models run at 6-h time steps over lumped basins hundreds of square kilometers 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.

Within the broad category of hydrologic errors, 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. In particular, this study focuses on the error from model calibration and the model initial conditions. The hydrologic and meteorological error terms are interrelated, and errors of one type may exaggerate or mask errors of another type. In addition, the spatial–temporal scales of the forecast area will affect the interaction between the errors. On large basins, where most of the river flow is routed water, the affect of meteorological error will be attenuated, while on small basins the affect of meteorological error will be considerable. This study focuses on small basins and, as will be seen in the hindcasts, the interaction between the types of error changes with lead time, which is an important element in understanding the sources of error and skill in the hindcasts.

b. Hydrologic hindcast experiments

With the growing availability of inexpensive computing power and disk space, hindcasting, which aims at retroactively generating forecasts using a fixed forecast scenario, is becoming a more usable tool for analyzing forecasts. The experimenter sets up a system to reforecast a set of events based upon the prior observations (obs) and forecasts (fcst). The forecast model is run with observed precipitation up to a date marked as ''present;'' the initial conditions for the model system are stored and then the model is restarted with forecast precipitation. The reforecast is computed and stored, and the model is run forward with observed precipitation to a new date marked as ''present.'' Each reforecast is called a hindcast. For the hindcasts to be valid, it is critical no observation be used in the calculations during the ''forecast'' period. During the hindcasting process, the computational methods, 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 as a result of annual variability in the local climate.

A few previous authors have used hindcasts to analyze hydrologic forecasts. Krzysztofowicz and Herr (2001) and Krzysztofowicz and Maranzano (2004) used hindcasts to parameterize their Bayesian Forecast System (BFS), which integrates the hydrologic and meteorological uncertainty into a single probability forecast. 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 National Weather Service (NWS) Ensemble Streamflow Prediction System from these reconstructed initial conditions. Werner et al. (2004) used hindcasts to evaluate several methods of computing temperature ensembles for use in mid- to long-range hydrologic forecasts. Using the methods of Franz et al. (2003), they reconstructed initial conditions for past years and compared seasonal volume hindcasts from various temperature ensembles. Demargne et al. (2007) analyzed the affect of two sets of input ensembles—climatology and QPF-based precipitation ensembles—on the quality of streamflow ensembles. They used both observed and simulated flows to divide the total uncertainty into the input uncertainty and hydrologic uncertainty. Deterministic forecasts are studied here, but the hindcast methods apply equally well to ensemble forecasts, as was demonstrated by Franz et al. (2003) and Werner et al. (2004), or by Krzysztofowicz and Herr (2001) and

Krzysztofowicz and Maranzano (2004). The results of a deterministic study like this one can be used to parameterize a model of forecast uncertainty. As was recommended by the National Research Council (2006), though hindcasting has not been used extensively in hydrology, it can be an effective tool for analyzing hydrologic forecasts.

c. Diagnostic verification

The verification method demonstrated here follows the diagnostic approach of Murphy and Winkler (1987). Murphy and Winkler suggested an approach based on the concepts of *discrimination* and *reliability,* which are derived from factoring the joint distribution of forecasts and observations $p(f,o)$ into the conditional distributions $p(f|o)$ for discrimination, or $p(o|f)$ for reliability. Each conditional distribution yields different information about the relations between the forecasts and the observations. Within the diagnostic framework, discrimination refers to the ability of the forecasts to distinguish between future events. Reliability refers to the forecasts' ability to forecast an event correctly; that is, if an event was forecast, did it occur.

When applying this diagnostic approach, the forecasts are sorted into discrete subsets and then each subset is evaluated. For example, when sorting stage forecasts into two categories, as was done in this analysis, the distributions to be evaluated when assessing discrimination skill are $p[(f,o)|o < T]$ for the low stage category and $p[(f,o)|o \geq T]$ 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 subsetted based upon the forecast value. The distributions to be assessed are then $p[(f,o)|f < T]$ for the low stage category and $p[(f,o)|f \geq T]$ 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 et al. (1989), which is 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.

2. The hindcast experiment

a. Algorithms used to compute the hindcasts

The forecast process to be analyzed here is the typical NWS short-term, deterministic river stage forecast process. For precipitation-driven headwater basins, the NWS generally uses a calibrated Sacramento model (Burnash 1995) at 6-h time steps to compute runoff from rainfall, a unit hydrograph to route runoff to the basin outlet (Linsley et al. 1975), and 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 h and zero is used after 24 h. The hydrologic model output is postprocessed using a simple linear difference scheme (National Weather Service 2002) to remove current model biases. The forecast flows are then converted to stages with a rating curve and the stage time series is issued as the forecast. For a more detailed description of the NWS short-term hydrologic forecast process, refer to Welles et al. (2007). 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. For the basins studied here, the initial conditions are soil wetness and channel flow, as described by the states of the Sacramento model and the unit hydrograph.

The forecast process cannot be reproduced exactly in the hindcast process because of three differences. Most obviously, the manual state updating cannot be recreated, as it would be too expensive and nonobjective. The variational assimilation method (VAR) proposed by Seo et al. (2003) is used to update these 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 that include snow, upstream routed flows, or reservoir operations. 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 that the simulation postprocessing is not used in the hindcasts, as it obscures the differences between the hindcast scenarios. The postprocessing algorithm forces the simulations to run through the last observed value; therefore, if the postprocessing were 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. Like the many physical basin characteristics requiring analysis (refer to the introduction), elements of the forecast process itself require analysis and the affect of postprocessing on the forecast skill is identified for a future study.

The third difference between the actual forecast operations and the hindcasts is the forecast issuance time. The actual forecasts are issued once daily, at 1200 UTC, unless flooding is imminent, in which case the forecasts are issued on an as-needed basis. The hindcasts are

''issued'' twice daily, at 0000 and 1200 UTC, and the schedule is not changed even if there is flooding.

b. The data

One obstacle to effective hindcasting is data archiving. Without a proper archive of the input to the original forecasts, they cannot be recreated. Three basins for which there was a suitable archive of the input data were found: the Illinois River at Watts, Oklahoma, and the Blue River at Blue, Oklahoma; and the Elk River at Tiff City, Missouri. These basins have been used in the Distributed Model Intercomparison Project (DMIP; Smith et al. 2004) and in testing the VAR (Seo et al. 2003), and they are selected here for the same reasons they have been selected previously: good data and well-understood hydrology. The basin locations are mapped in Fig. 1. The three basins range in size from 1230 to 2250 km^2, typical sizes for NWS forecast operations. Annual rainfall is approximately 1200 with 350 mm annual runoff. The topography is rolling hills, resulting in moderately fast hydrograph responses, with the Blue River having steeper hydrograph recessions than the Elk and Illinois Rivers. For a more detailed description of the basin geo–hydrology, refer to Smith et al. (2004). The observed precipitation used for the hindcasts was taken from the NWS Stage III grids (Young et al. 2000) computed 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 U.S. Geological Survey (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.

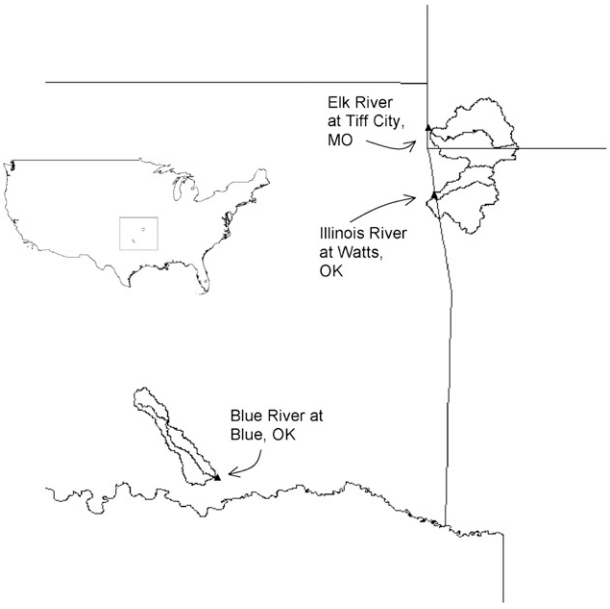

FIG. 1. The location of the basins used in this study.

c. The hindcast scenarios

Three forecast process elements were studied here: calibration, state updating, and QPF. For each forecast process element, a ''skilled'' implementation and an ''unskilled'' implementation was developed. For the skilled calibration, parameters were derived by NWS experts within the NWS Hydrology Laboratory using manual calibration methods described in the NWS calibration handbook (Anderson 2002). For the unskilled calibration, model parameters were derived from the pedological equations of Koren et al. (2003), with no additional manual calibration performed on the pedological results. These parameters are commonly used as an initial parameter set to begin the manual calibration process and are referred to as the uncalibrated or a priori parameters. There are a number of methods for calibrating hydrologic models—manual, automated, and semiautomated—in addition to postprocessing techniques used to account for model bias. A comparison of these different techniques to determine the one most suitable for skillful forecasting merits additional hindcast analysis but is beyond the scope of this discussion.

The skilled and unskilled state updating was computed by running the hindcasts with the VAR turned on for the skilled implementation and turned off for the unskilled implementation. Three QPF implementations were used: skilled, unskilled, and perfect. For the unskilled implementation, the QPF is set to zero for the entire forecast period; this is called the zero QPF scenario. For the skilled implementation, the operationally modeled QPF is used for the first 24 h and 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 QPF scenario, the observed precipitation is used as the QPF. The first two QPF implementations are commonly used in the NWS operational forecast process. Although the zero QPF scenario may not appear as a reasonable QPF alternative, it has been commonly used in forecast operations for many years. The calibration, the state updating, and the QPF types are matched for a total of 12 hindcast scenarios on each basin. Table 1 lists each hindcast scenario. Persistence (pers) hindcasts were also generated and are used to provide a perspective on the hindcast skill. Persistence is defined as the observation at the basis time[2] of the forecast.

[2] The basis time of a forecast is the time when the forecast is issued.

TABLE 1. The names of the hindcast scenarios.

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 uncalibrated parameters	P-V-U
Perfect QPF without VAR and uncalibrated parameters	P-NV-U
Real QPF with VAR and uncalibrated parameters	R-V-U
Real QPF without VAR and uncalibrated parameters	R-NV-U
Zero QPF with VAR and uncalibrated parameters	Z-V-U
Zero QPF without VAR and uncalibrated parameters	Z-NV-U

d. Hindcast analysis process

Once the hindcasts have been generated, they are sorted into subsets according to the purpose of the analysis, for example, by lead time, season, or basin size. The subsets of hindcasts can then be compared using a variety of statistical and verification measures. Distributions can be compared directly through parametric or nonparametric tests, or verification metrics can be computed and then compared. There is considerable latitude within this process to allow for analyses of numerous forecast types and characteristics.

For the purpose of this study, which is focused on flood forecast skill, the forecasts and observations were sorted into two subsets: high and low stages. It is possible to sort into finer categories, and when this was done, the characterization of the hindcast–observed relation was similar to that for the two-category sorting. Therefore, a stage just below the NWS alert stage was selected for each basin to ensure sufficient sample sizes in the high stage category (4.0 m for the Illinois River at Watts, OK; 6.1 m for the Blue River at Blue, OK; and 3.7 m for the Elk River at Tiff City, MO). In addition to sorting by stage height, the hindcasts were sorted into lead times at 6-h time steps. Statistics for each subset were computed on the forecast observation pairs collected from all three basins and then these statistics were compared to isolate the changes in skill provided by each forecast process element.

TABLE 2. A priori parameter scenarios compared to calibrated parameter scenarios to assess calibration affects.

Scenarios compared
ΔRMSE(P,V) = RMSE(P-V-U) – RMSE(P-V-C)
ΔRMSE(P,NV) = RMSE(P-NV-U) – RMSE(P-NV-C)
ΔRMSE(R,V) = RMSE(R-V-U) – RMSE(R-V-C)
ΔRMSE(R,NV) = RMSE(R-NV-U) – RMSE(R-NV-C)
ΔRMSE(Z,V) = RMSE(Z-V-U) – RMSE(Z-V-C)
ΔRMSE(Z,NV) = RMSE(Z-NV-U) – RMSE(Z-NV-C)

TABLE 3. No variational assimilation scenarios compared to variational assimilation scenarios to assess data updating affects.

Scenarios compared
ΔRMSE(P,C) = RMSE(P-NV-C) – RMSE(P-V-C)
ΔRMSE(P,U) = RMSE(P-NV-U) – RMSE(P-V-U)
ΔRMSE(R,C) = RMSE(R-NV-C) – RMSE(R-V-C)
ΔRMSE(R,U) = RMSE(R-NV-U) – RMSE(R-V-U)
ΔRMSE(Z,C) = RMSE(Z-NV-C) – RMSE(Z-V-C)
ΔRMSE(Z,U) = RMSE(Z-NV-U) – RMSE(Z-V-U)

The sets were characterized with the following summary statistics: the mean absolute error, the root-mean-square error (RMSE), the mean error (ME), the false alarm ratio (FAR), the probability of detection (POD), the critical success index (CSI), the area under the relative operating characteristics (ROC) curve, a ROC discrimination distance, and the Pearson correlation coefficient (R). It was found the measures themselves were not the key to understanding the error in the hindcasts but rather the comparisons among the hindcasts and subsets made the verification meaningful. Therefore, the RMSE is used in the presentation of the hindcast comparisons. For the description of the calibrations, the ME and R are also reported.

For each hindcast scenario, including the persistence hindcasts, and for each lead time, the RMSE is computed for the high stage and low stage reliability and discrimination subsets across all three locations. For

TABLE 4. QPF scenarios compared to assess precipitation forecast affects.

Scenarios compared
ΔRMSE(RVC,PVC) = RMSE(R-V-C) – RMSE(P-V-C)
ΔRMSE(RNVC,PNVC) = RMSE(R-NV-C) – RMSE(P-NV-C)
ΔRMSE(ZVC,PVC) = RMSE(Z-V-C) – RMSE(P-V-C)
ΔRMSE(ZNVC,PNVC) = RMSE(Z-NV-C) – RMSE(P-NV-C)
ΔRMSE(ZVC,RVC) = RMSE(Z-V-C) – RMSE(R-V-C)
ΔRMSE(ZNVC,RNVC) = RMSE(Z-NV-C) – RMSE(R-NV-C)
ΔRMSE(RVU,PVU) = RMSE(R-V-U) – RMSE(P-V-U)
ΔRMSE(RNVU,PNVU) = RMSE(R-NV-U) – RMSE(P-NV-U)
ΔRMSE(ZVU,PVU) = RMSE(Z-V-U) – RMSE(P-V-U)
ΔRMSE(ZNVU,PNVU) = RMSE(Z-NV-U) – RMSE(P-NV-U)
ΔRMSE(ZVU,RVU) = RMSE(Z-V-U) – RMSE(R-V-U)
ΔRMSE(ZNVU,RNVU) = RMSE(Z-NV-U) – RMSE(R-NV-U)

each forecast process element, the scenarios that are similar except for the forecast process element of interest were compared. For example, to evaluate the contribution to the hindcast skill from the calibration, the hindcasts with the skilled and unskilled calibration but the same QPF and updating treatments were compared. The same was done to isolate the contribution of the initial conditions to the hindcast skill: the hindcasts with skilled and unskilled updating but the same QPF and calibration treatments were compared. For the QPF, the same procedure was followed: the state updating and the calibration were held constant and the different QPF scenarios were compared. A list of the comparisons is provided in Tables 2–4. The hindcast results are introduced by reporting the RMSE for the scenarios (Figs. 2–5). The comparisons between the scenarios are presented as differences (Figs. 6–13). That is, the RMSE for the skilled hindcast is subtracted from the RMSE for the unskilled hindcast, resulting in a delta RMSE, noted as ΔRMSE. Positive ΔRMSE indicates an improvement in the forecast RMSE when moving from the unskilled to the skilled method. Negative ΔRMSE indicates there was no improvement when moving from the unskilled to the skilled method. Sample sizes are used to indicate confidence in the metrics (Figs. 14 and 15). Developing constructive methods for computing confidence intervals for these metrics is an area requiring further research. Interested readers may find an initial approach described in Welles (2003).

3. Results

a. The two calibrations

Because calibration is such an important aspect of hydrologic modeling, the skilled and the unskilled calibration are described and compared. The perfect QPF hindcast with no state updating (P-NV-C and P-NV-U) is the same as a standard calibration simulation: there is no state updating and observed precipitation is used to drive the models. It is customary within the NWS to use the ME to evaluate a calibration; therefore, the ME is reported in addition to the RMSE. For completeness, the R is also reported. The statistics computed from this hindcast scenario for the calibrated and uncalibrated parameters are summarized in Table 5.

For the low stage discrimination and reliability, both the calibrated and the uncalibrated parameters have almost no ME. The uncalibrated low stage discrimination RMSE (1.07 m), however, is more than twice the calibrated RMSE (0.40 m), and the uncalibrated 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 (0.85 for discrimination and 0.91 m for reliability). The uncalibrated model, on the other hand, tends to overforecast the observed high stages (discrimination ME of 0.98 m), and it tends to forecast too many high stages (reliability ME of 3.20 m). In addition, the high stage discrimination and reliability correlations for the uncalibrated 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 that the expert calibration provides a considerable improvement to the simulations.

b. The QPF skill

A short summary of the QPF skill is presented to help explain the behavior of the hindcasts. For the QPF, the ME is reported in addition to the RMSE 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. In addition, it was found there is little variation in the QPF skill across the four lead times, so the 6-h forecasts were pooled into a single sample set to simplify the reporting. (They were not added together to produce a single 24-h QPF; they were collected into a single sample set for the 24-h period.) The metrics reported in Table 6 are for the 24-h collection.

For both discrimination and reliability, the NWS–ABRFC issued forecasts have lower RMSE and ME than the zero QPF, demonstrating that the NWS–ABRFC 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) 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 underforecast 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 hindcasts.

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

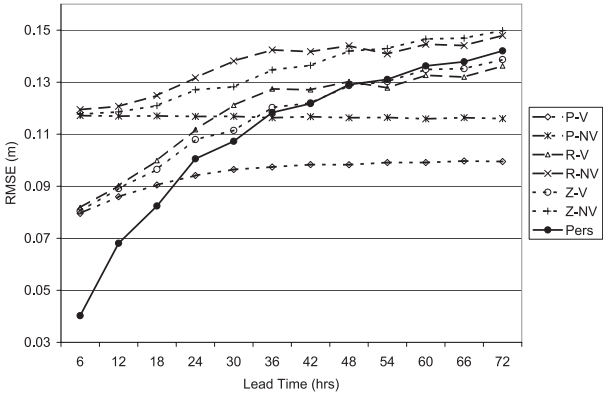

FIG. 2. Calibrated parameters, discrimination statistics for low stage events. Scenarios shown are P-V, P-NV, R-V, R-NV, Z-V, Z-NV, and pers. Refer to Table 1.

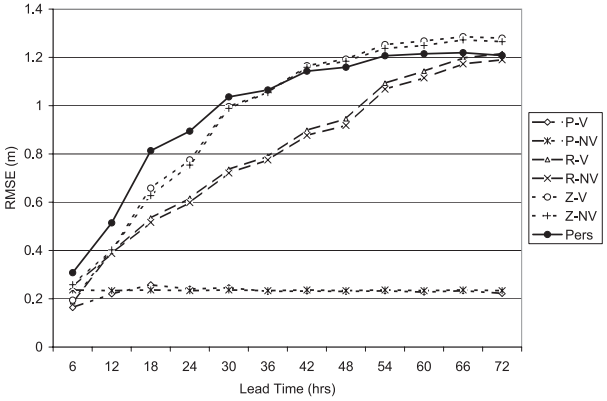

FIG. 4. Same as Fig. 2 but for calibrated parameters, discrimination statistics for high stage events.

places in the United States), the statistics from the NWS National Precipitation Verification Unit (NPVU; McDonald et al. 2000; available online at http://www.hpc.ncep.noaa.gov/npvu/index.shtml) are provided in Table 7. 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 uncertainty in the QPFs across the country, and the error in the hydrologic simulations caused by the QPF in the hindcasts representative of the QPF-driven error elsewhere in the United States.

c. *Hindcasts and persistence*

The persistence provides a baseline for comparison to the hindcast skill. For low stage discrimination and reliability, the only hindcasts that perform better than persistence are the well-calibrated scenarios with perfect QPF at lead times greater than 18 h for the VAR and 30 h for nonVAR scenarios (see Fig. 2). The uncalibrated parameters for both the low stage discrimination and reliability never perform better than persistence; even for the perfect QPF scenarios (see Fig. 3).

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 persistence. For the high stage discrimination, all the calibrated scenarios (Fig. 4) perform better than the persistence for the first 24 h. After 30 h, 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 high stage reliability (not shown), the RMSE for the calibrated parameters is almost a factor of 2 smaller than the persistence RMSE, while the uncalibrated RMSE (Fig. 5) is larger than the persistence RMSE. By comparing the relations to persistence, it can be seen that the forecast process adds more reliability to the forecasts than it adds discrimination because, as will be seen

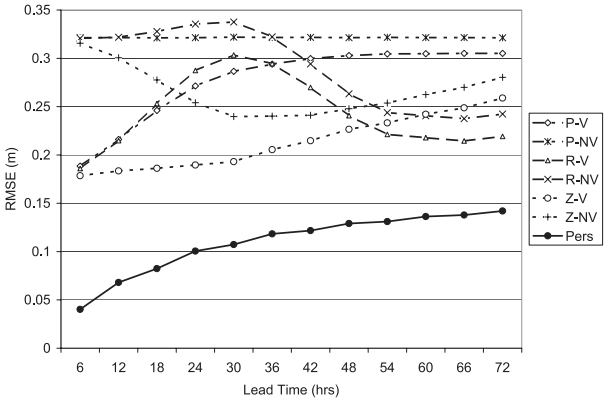

FIG. 3. Same as Fig. 2, but for a priori parameters.

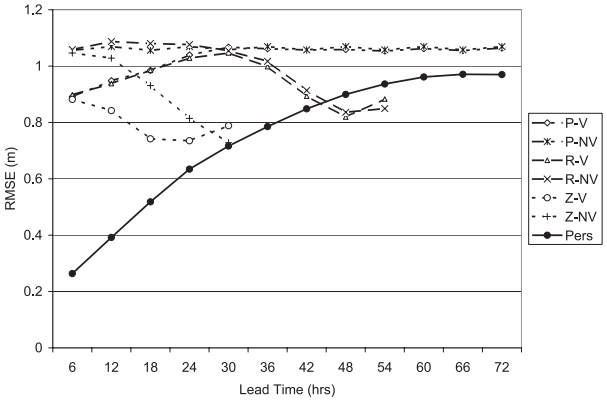

FIG. 5. Same as Fig. 2, but for a priori parameters, reliability statistics for high stage events.

TABLE 5. Summary statistics to compare the model calibrations (P-NV-C and P-NV-U).

	Discrimination			Reliability		
	RMSE (m)	ME (m)	*R*	RMSE (m)	ME (m)	*R*
Low: Calibrated	0.40	0.061	0.85	0.40	0.61	0.85
Low: Uncalibrated	1.07	−0.30	0.51	0.79	−0.18	0.50
High: Calibrated	0.85	0.0	0.65	0.91	0.0	0.75
High: Uncalibrated	1.77	0.98	0.55	3.50	3.20	0.35

later, most discrimination skill comes from QPF, while the model calibration adds reliability skill when the calibration is done well.

d. Hindcast skill from calibration

The results of comparing the skill of the calibrated and uncalibrated hindcast scenarios indicate calibration is important for the low stage skill and for the high stages when the lead time is <1 day. However, the skill calibration can provide to the high stages at lead times >1 day is limited when the QPF is poor. The differences between the hindcast RMSEs for the low stage discrimination (Fig. 6) indicate the calibration provides considerable improvement to the hindcasts, reducing the RMSE to half of the original uncalibrated RMSE (shown in Fig. 3). 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 that must be converted to runoff. The improvement to the real QPF scenario matches the improvement to the perfect QPF scenario until the real QPF turns to zero (at 24 h) 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.

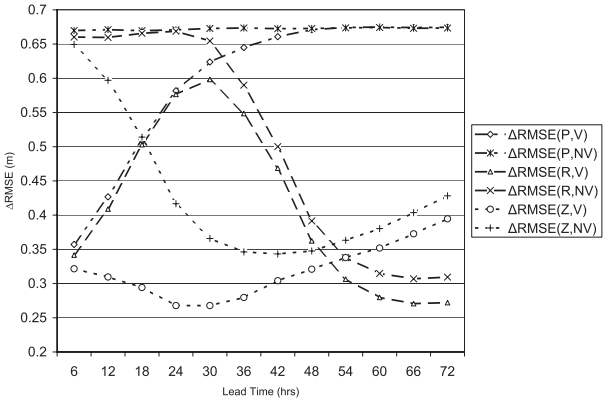

FIG. 6. Differences between calibration scenarios, discrimination statistics for low stage events. Refer to Table 2 for scenarios.

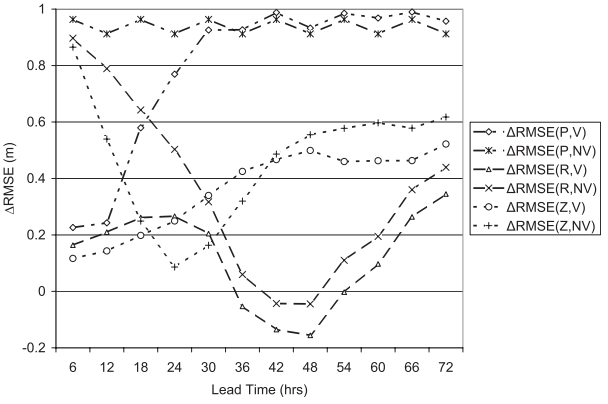

FIG. 7. Same as Fig. 6, but for high stage events.

As was noted above (section 3a), the expert calibration provides an improvement of slightly more than 0.85 m to the high stage simulations, half the a priori RMSE. The hindcasts without state updating realize this 0.85 m improvement in the first time steps (Fig. 7), but the updated scenarios benefit much less (0.3 m). As with the low stages, this difference indicates the calibration provides skill through good initial conditions, and, as will be seen in the next section, the calibration and the 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 h and then becomes negative. 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 lead times, it only resolves a small portion of the total discrimination uncertainty at the longer lead times because at longer lead times, the uncertainties in the input overwhelm the model, no matter how well structured and well calibrated the model is.

Another important result seen in Fig. 7 is the complex interaction between the hydrologic and meteorological errors. For example, the dipping and rising pattern for the real QPF scenarios in days 2 and 3 is caused by the tendency of the QPF to underforecast the interaction with the tendency of the uncalibrated hydrologic model to overforecast. During the early lead times, the over forecast precipitation in the real QPF causes the uncalibrated model to overreact and improving the calibration can improve the hindcasts. After the first 24 h, zeros are used in the real QPF and the zeros tend to mitigate the tendency of the uncalibrated model to overforecast, while at the same time causing the calibrated model to underforecast. Therefore, calibrating the models does not improve the hindcasts after the first 24 h. At the longest lead times, when the QPF has been zero for 24 h, the

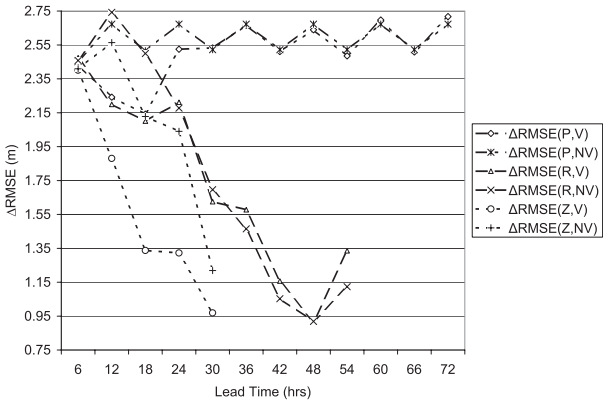

FIG. 8. Same as Fig. 6, but for reliability statistics for high stage events.

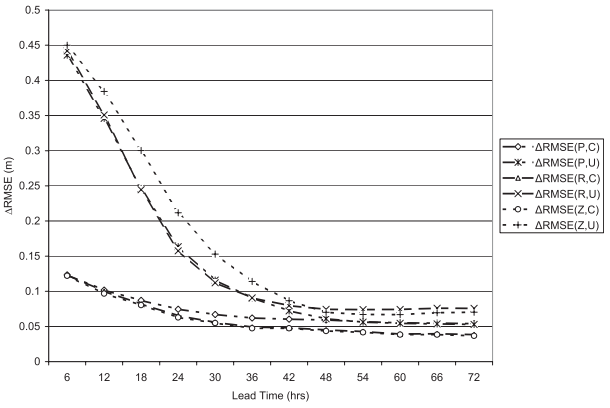

FIG. 9. Differences between state updating scenarios, discrimination statistics for low stage events. Refer to Table 3 for scenarios.

overforecasting in the hydrologic models does not mitigate the underforecast QPF and the calibration begins to add skill again. This rising and falling is a clear indication the hydrologic and the meteorological errors are neither independent nor additive. The common assumption that forecasts will always improve when an element of the forecast process is enhanced turns out not to be true. Therefore, it is imperative that proposed improvements to the forecast process be considered within the context of the entire forecast process and not independently.

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 skill by the calibration (Fig. 8) is more than half the total error of the a priori parameters. For the zero QPF scenarios, this improvement falls quickly. For the real QPF scenario, on the other hand, the improvement holds up for the first 24 h before it begins to fall because the QPF becomes zero and the value of the calibration is reduced, as there is no rain to convert to runoff. A good calibration contributes reliability skill but little discrimination skill, partially explaining the results seen in the comparison to persistence.

e. Hindcast skill from state updating

As one might expect, the comparison of the state updating procedures shows the uncalibrated model benefits the most from the skilled state updating. For the low stage discrimination (Fig. 9), the hindcasts group themselves by the type of calibration. This is the same phenomenon seen in the calibration comparisons, with the updated and the nonupdated 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 0.2 m at 42 h. 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.

The patterns seen in Figs. 9 and 10 indicate the skill derived from the initial conditions is independent of the QPF skill and comparable to the skill derived from a good calibration. For the low stages, the calibration provides slightly more improvement (0.7 m) to the 6-h lead time than the state updating (0.45 m); however, for the high stages, the calibration and the state updating bring an equal amount of improvement (0.9 m) to the 6-h lead time. For the high stage reliability, the pattern of improvement is the same, but the magnitude is less (0.6 m). For the high stage discrimination (see Fig. 10), the same pattern is apparent. The way the scenarios are grouped, by calibration not by QPF, indicates a good state updating scheme, and a good calibration can provide similar skill to the hindcasts through the initial conditions.

f. Hindcast skill from QPF

For the low stage discrimination (Fig. 11), the type of QPF makes little difference to the hindcasts; the improvement to the RMSE due to improving the QPF stays below 0.1 m for all the scenarios and only reaches 0.1 m in day 3. This is much less than the minimum 0.3-m improvement provided by the calibration and the state updating in the early lead times. The nonlinear interaction of the meteorological and hydrologic errors is, again, visible in these comparisons. Changing the QPF from the zero QPF to the real QPF results in near-zero change in the RMSE of the calibrated model. On the other hand, in the case of uncalibrated model, improving the QPF from the zero QPF scenario to either the real or the perfect QPF actually harms the hindcast RMSE (improvement of −0.37 m) at hour 30. These rises and falls can be traced to the changes in the

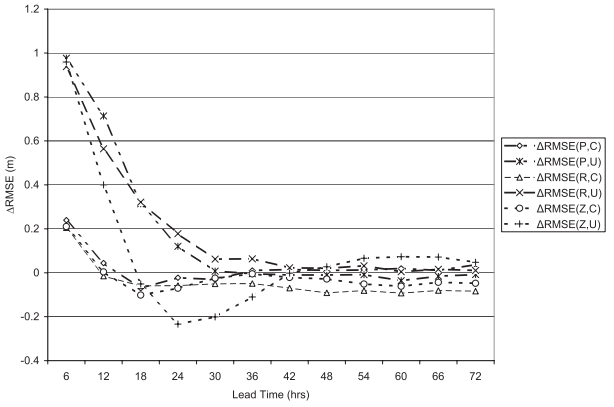

FIG. 10. Same as Fig. 9, but for high stage events.

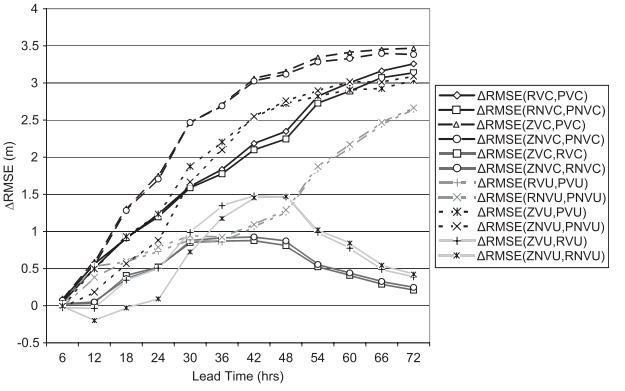

FIG. 12. Same as Fig. 11, but for high stage events and ΔRMSE(ZVU,RVU) and ΔRMSE(ZNVU,RNVU) instead of ΔRMSE(ZVU,PVU) and ΔRMSE(ZNVU,PNVU).

forecast variance, as the uncalibrated model responds too strongly to the nonzero QPF where 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 (Fig. 12), 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. It is worth noting that the comparisons between the real QPF and the zero QPF scenarios fall toward zero after 24 h for the calibrated model [see comparisons (Z-NV-C, R-NV-C) and (Z-V-C, R-V-C)]. These comparisons fall toward zero because the QPF is zero after 24 h in the real QPF scenario. This fall confirms the importance of the modeled QPF to the forecast skill. The transition to the perfect QPF shows the large potential improvement possible from improving the QPF.

While the improvement to the discrimination skill from the three QPF scenarios was similar for the calibrated and uncalibrated models, the improvement to the high stage reliability (Fig. 13) shows marked differences between the calibrated and the uncalibrated results. For the calibrated high stages, switching between the QPF types makes no change to the hindcast reliability for the first 18 h. After 18 h switching to perfect QPF improves the hindcast reliability but switching from the zero to real QPF causes a negative change. That is, when switching to a more skillful QPF scenario, the skill of the uncalibrated model falls, again demonstrating the complexity of the interaction between the hydrologic and meteorological errors. In addition, these results show the importance of a good calibration. A good calibration ensures the forecasts will improve as the QPF

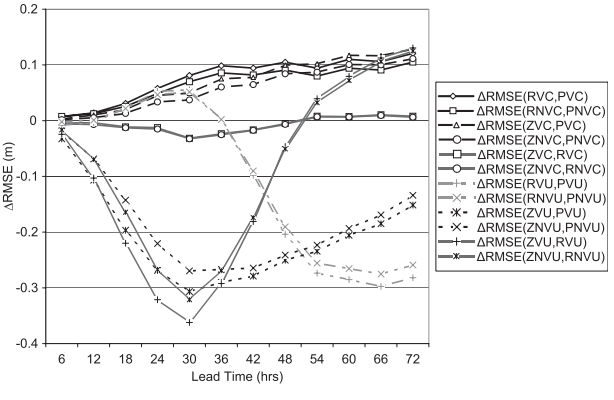

FIG. 11. Differences between QPF scenarios, discrimination statistics for low stage events. Comparisons of scenarios ΔRMSE(RVC,PVC), ΔRMSE(RNVC,PNVC); ΔRMSE(ZVC,PVC), ΔRMSE(ZNVC,PNVC); ΔRMSE(ZVC,PVC), ΔRMSE(ZNVC, PNVC); and scenarios with ΔRMSE(RVU,PVU), ΔRMSE(RNVU, PNVU); ΔRMSE(ZVU,PVU), ΔRMSE(ZNVU,PNVU); and ΔRMSE(ZVU,PVU), ΔRMSE(ZNVU,PNVU). Refer to Table 4.

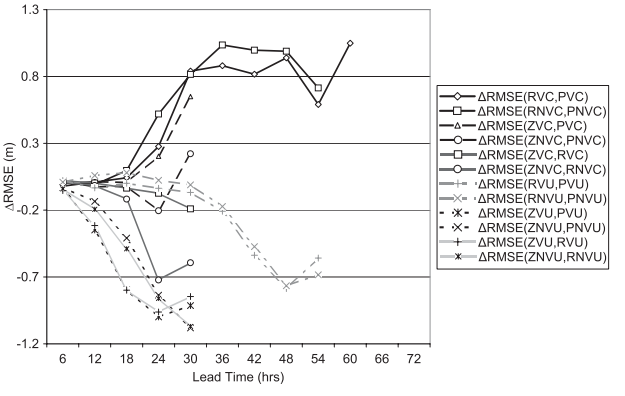

FIG. 13. Same as 12, but for reliability statistics.

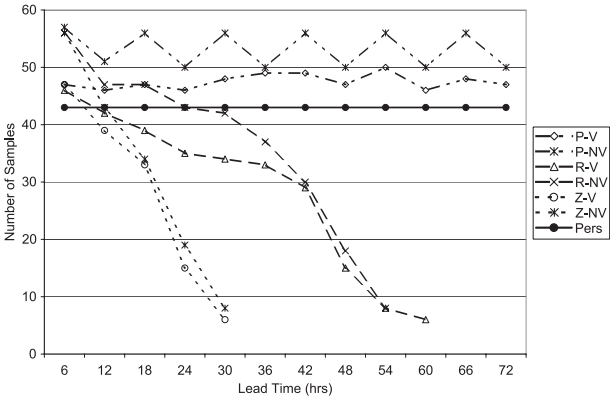

FIG. 14. Sample sizes of calibrated parameter scenarios for high stage reliability statistics. Same scenarios as for Fig. 2.

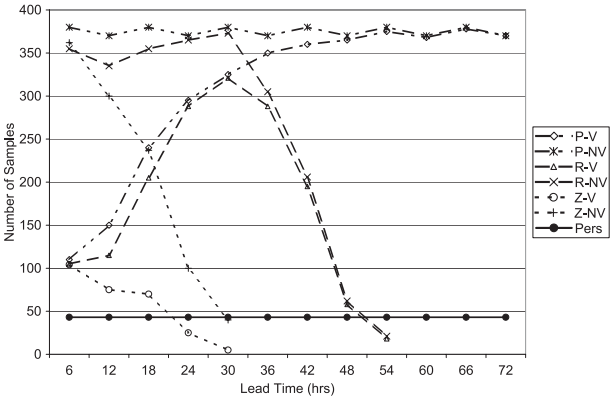

FIG. 15. Same as Fig. 14, but for a priori parameter scenarios.

skill improves, validating the consistent drive by hydrologic forecasters to calibrate their models.

g. *Hindcast sample sizes*

The sample sizes can 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 (Figs. 14 and 15) vary from reasonably high (400) to very small (5). 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 ordering from the comparisons, though the absolute magnitudes of the metrics may be uncertain. The higher stage threshold was used in this presentation because the primary purpose of any operational hydrologic service is flood forecasting; therefore, it is the high stages that are critical.

4. Discussion

a. *Hindcast skill and the forecast process*

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 h or fewer), the discrimination skill for both the high and the low stages is controlled by the initial conditions (Figs. 7, 9, and 10). 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 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 in the large

TABLE 6. The actual QPF compared to the zero QPF for the three hindcast basins.

	(mm)	ME by obs (mm)	RMSE by obs (mm)	ME by fcst (mm)	RMSE by fcst (mm)	Samples by obs (No. samples)	Samples by fcst (No. samples)
Actual QPF	≤25	0.2	2.5	0.1	3.0	31 800	31 920
Zero QPF	≤25	−0.7	2.6	−0.7	3.5	31 800	31 946
Actual QPF	>25	−22.8	25.7	15.5	20.3	146	26
Zero QPF	>25	−33.2	34.7	NA	NA	146	NA

TABLE 7. NPVU QPF statistics and the QPF statistics for the three hindcast basins.

	Category (mm)	ME by obs (mm)	RMSE by obs (mm)	ME by fcst (mm)	RMSE by fcst (mm)	FAR	POD
National	≤25	0.1	2.4	0.1	2.4	—	—
Hindcast basins	≤25	0.2	2.5	0.1	3.0	—	—
National	>25	−24.7	29.1	16.0	23.0	0.76	0.10
Hindcast basins	>25	−22.8	25.7	15.5	20.3	0.77	0.04

differences between the perfect and the real QPF scenarios (Figs. 4 and 12) and at the same time the much smaller differences between the well-calibrated and uncalibrated models when using zero or real QPF (Fig. 7). 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; therefore, 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 (Fig. 7). At the same time, improving the QPF may degrade the forecasts if the calibration is biased; 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 (Fig. 12). This interaction between the hydrologic and the meteorological errors is the same phenomenon noted by Krzysztofowicz (1999) when he found the common notion that the hydrologic and meteorological errors are additive was false.

Model calibration and improved hydrologic process modeling have been the traditional focus for improving forecast skill. Similar to the Shi et al. (2008) study, the results here indicate hydrologic model accuracy (derived via model calibration or improved process descriptions) may not improve hydrologic forecast skill on small headwater basins, as is usually expected. Rather, ensemble techniques used to capture precipitation uncertainty (e.g., Schaake et al. 2007) may be more likely to yield improved forecast skill. It is important to note though, while the present day QPF skill limits the improvement possible in the hindcasts as a result of 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. In addition, the reliability skill is controlled by the calibration, indicating an accurate, well-calibrated hydrologic model is an important foundation for a skillful hydrologic forecast system.

b. Implications for hydrologic verification

The analysis presented here also provides some direction for the integration of verification analyses with real-time forecasting. Using simple comparisons of forecasts combined with analysis of the input forecasts can be used to establish objective insight into the sources of forecast skill and error. Therefore, storing a persistence baseline and the perfect QPF simulation without state updating will provide substantial objective information to support operational forecasters. The persistence forecast will provide 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

expand the objective description of the forecast skill beyond the precipitation-driven headwaters studied here.

5. Conclusions

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

- First, 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.
- Second, 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.
- Third, it is essential that all the input forecasts to the system are verified alongside 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 will have a substantial influence on the forecast skill.
- Fourth, and perhaps most importantly, hydrologists need to do more studies like this one. 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 (nonheadwater locations) is important and requires study as well. Unfortunately, such studies are hampered by the cost of developing the infrastructure to compute hindcasts along 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 that include analysis procedures, such as the one illustrated here, are at the center of developing this comprehensive understanding.

Acknowledgments. Both authors would like to thank two reviewers from the National Weather Service's Office of Hydrology (OHD): Gary Carter, director, and Dr. Julie Demargne, research scientist. In addition, thanks are due to Dr. Holly Hartmann, who reviewed an early version of this manuscript. Many thanks also to Billy Olsen, hydrologist in charge, and Bill Lawrence, Development and Operations hydrologist at the National Weather Service, Arkansas–Red Basin River Forecast Center. Dr. D. J. Seo provided the code to run the variational assimilation hindcasts, so we thank him as well. The second author also would like to thank the National Weather Service (Grants NA87WHO582 and NA07WHO144) and the National Science Foundation STC, Sustainability of Semi-Arid and Hydrology and Riparian Areas (SAHRA) at the University of Arizona in Tucson (Grant EAR-9876800) for their support. The views expressed in this paper are those of the authors and do not necessarily represent those of the National Weather Service.

REFERENCES

Anderson, E. A., 2002: Calibration of conceptual hydrologic models for use in river forecasting. NOAA Tech. Rep. NWS 45 Hydrology Laboratory, 247 pp.

Burnash, R. J. C., 1995: The NWS River Forecast System—Catchment modeling. *Computer Models of Watershed Hydrology*, V. P. Singh, Ed., Water Resources Publications, 311–366.

Demargne, J., L. Wu, D.-J. Seo, and J. Schaake, 2007: Experimental hydrometeorological and hydrological ensemble forecasts and their verification in the US National Weather Service. *Quantification and Reduction of Predictive Uncertainty for Sustainable Water Resources Management*, IAHS Publication 313, 177–187.

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. *J. Hydrometeor.,* **4,** 1105–1118.

Koren, V., M. Smith, and Q. Duan, 2003: Use of a priori parameters estimate in the derivation of spatially consistent parameters sets of rainfall-runoff models. *Calibration of Watershed Models*, Q. Duan et al., Eds., Water Science and Applications Series, Vol. 6, Amer. Geophys. Union, 239–254.

Krzysztofowicz, R., 1999: Bayesian theory of probabilistic forecasting via a deterministic hydrologic model. *Water Resour. Res.,* **35,** 2739–2750.

——, and H. D. Herr, 2001: Hydrologic uncertainty processor for probabilistic river stage forecasting: Precipitation-dependent model. *J. Hydrol.,* **249,** 46–68.

——, and C. J. Maranzano, 2004: Hydrologic uncertainty processor for probabilistic stage transition forecasting. *J. Hydrol.,* **293,** 57–73.

Linsley, R. K., M. A. Kohler, and J. L. H. Paulhus, 1975: *Hydrology for Engineers*. McGraw Hill, 508 pp.

McDonald, B. E., T. Graziano, and C. K. Kluepfel, 2000: The NWS national QPF verification program. Preprints, *15th Conf. on Hydrology*, Long Beach, CA, Amer. Meteor. Soc., 247–250. [Available online at http://ams.confex.com/ams/annual2000/techprogram/paper_6441.htm.]

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev,* **115,** 1330–1338.

——, B. Brown, and Y.-S. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting,* **4,** 485–501.

National Research Council, 2006: *Toward a New Advanced Hydrologic Prediction Service (AHPS)*. National Academies Press, 74 pp.

National Weather Service, 2002: Adjust-Q: Adjust simulated discharge operation. NWS River Forecast System Manual, V.3.3., 10 pp.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark, 2007: HEPEX: The Hydrological Ensemble Prediction Experiment. *Bull. Amer. Meteor. Soc.,* **88,** 1541–1547.

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

Shi, X., A. W. Wood, and D. P. Lettenmaier, 2008: How essential is hydrologic model calibration to seasonal streamflow forecasting? *J. Hydrometeor.,* **9,** 1364–1377.

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

Welles, E., 2003: Verification of river stage forecasts. Ph.D. dissertation, University of Arizona, 155 pp. [Available from University Microfilm, 305 N. Zeeb Rd., Ann Arbor, MI 48106.]

——, S. Sorooshian, G. Carter, and B. Olsen, 2007: Hydrologic verification: A call for action and collaboration. *Bull. Amer. Meteor. Soc.,* **88,** 503–511.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay, 2004: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts. *J. Hydrometeor.,* **5,** 1076–1090.

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.