Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin

KRISTIE J. FRANZ,* HOLLY C. HARTMANN, SOROOSH SOROOSHIAN,* AND ROGER BALES+

Department of Hydrology and Water Resources, The University of Arizona, Tucson, Arizona

(Manuscript received 23 December 2002, in final form 22 May 2003)

ABSTRACT

The Ensemble Streamflow Prediction (ESP) system, developed by the National Weather Service (NWS), uses conceptual hydrologic models and historical data to generate a set, or ensemble, of possible streamflow scenarios conditioned on the initial states of a given basin. Using this approach, simulated historical probabilistic forecasts were generated for 14 forecast points in the Colorado River basin, and the statistical properties of the ensembles were evaluated. The median forecast traces were analyzed using ''traditional'' verification measures; these forecasts represented ''deterministic ESP forecasts.'' The minimum-error and historical traces were examined to evaluate the median forecasts and the forecast system. Distribution-oriented verification measures were used to analyze the probabilistic information contained in the entire forecast ensemble. Using a single-trace prediction, for example, the median, resulted in a loss of valuable uncertainty information about predicted seasonal volumes that is provided by the entire ensemble. The minimum-error and historical traces revealed that there are errors in the data, calibration, and models, which are part of the uncertainty provided by the probabilistic forecasts, but are not considered in the median forecast. The simulated ESP forecasts more accurately predicted future streamflow than climatology forecasts and, on average, provided useful information about the likelihood of future streamflow magnitude with a lead time of up to 7 months. Overall, the forecast provided stronger probability statements and became more reliable at shorter lead times. The distribution-oriented verification approach was shown to be applicable to ESP outlooks and appropriate for extracting detailed performance information, although interpretation of the results is complicated by inadequate sample sizes.

1. Introduction and scope

In the southwest United States, water supply outlooks of naturalized, or unimpaired, volumes are issued jointly by the National Weather Service (NWS) River Forecast Centers (RFCs) and the Natural Resources Conservation Service. Each agency generates forecasts individually and then meets with other interested forecasting parties to subjectively evaluate each forecast for combination into one product (Hartmann et al. 1999, 2002a). Water supply forecasts have been issued for many decades and are important for making a wide variety of decisions, including water allocation for urban and agricultural uses and reservoir operations.

Current water supply forecasts for the western United States are based largely on statistical regression equations that are developed mostly from monthly precipitation, recent snow-water equivalent, and past streamflow observations (Day 1985). Shafer and Huddleston (1984) conducted a comprehensive assessment of operational hydrologic forecasts in the west and concluded that, while the regression forecasts would continue to be useful, overall large improvement in the forecast accuracy could not be expected through the refinement of regression techniques. In addition, Day (1985) stated that the value of deterministic regression forecasts is limited because they do not provide information about the uncertainty of the predictions. To provide an objective means with which to generate streamflow forecasts with uncertainty, the NWS Ensemble Streamflow Prediction (ESP) method was developed (Day 1985). ESP uses physically based conceptual hydrologic models, with states set to current basin conditions, and multiple meteorological inputs to create a probabilistic outlook consisting of a distribution of possible future events.

Forecast verification is important for assessing forecast quality and performance trends, improving the forecasting procedures, and providing users with information helpful in applying the forecasts (Murphy and Winkler 1987). Verification is particularly important for understanding new forecasting methods. Because ESP

* Current affiliation: Department of Civil and Environmental Engineering, University of California, Irvine, Irvine, California.

+ Current affiliation: University of California, Merced, Merced, California.

Corresponding author address: Ms. Kristie J. Franz, Dept. of Civil and Environmental Engineering, University of California, Irvine, E/4130 Engineering Gateway, Irvine, CA 92697-2175.
E-mail: franzk@uci.edu

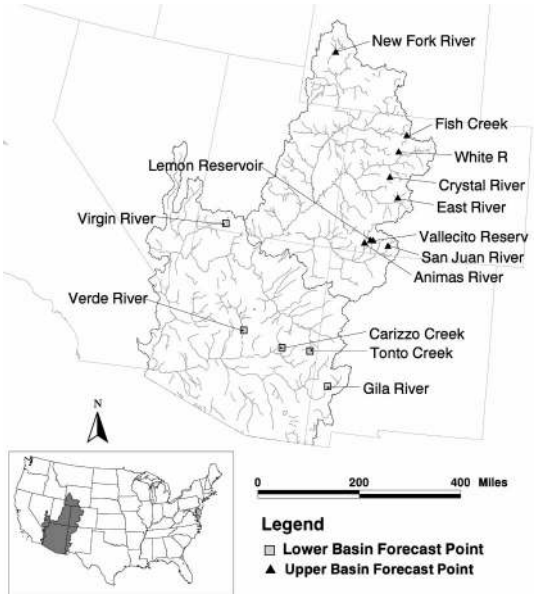

FIG. 1. The Colorado River basin and the forecast points used in this study.

is a fairly recent advancement in hydrologic forecasting, there is a limited database with which to test analytical methods and produce statistical information. Using the NWS Ensemble Streamflow Prediction Verification System (ESPVS) (Riverside Technology, Inc. 1999), simulated historical forecasts, or ''hindcasts'', were created for select locations in the Colorado River basin. The hindcasts allowed testing of several verification methods ranging from simple, more traditional measures, to more advanced distribution-oriented techniques, and an initial assessment of the ESP forecasts' information and quality. From these analyses, insight into information provided by ESP and appropriate verification techniques can be acquired.

The objectives of this paper are as follows:

- to examine the applicability of traditional statistical techniques and distribution-oriented measures for the evaluation of ESP forecasts, and
- to give insight into potential operational forecast performance based on simulated historical ESP forecasts for the Colorado River basin.

2. Methods

a. Forecast locations

The Colorado River basin is located in the western United States; its drainage area (242 000 square miles) covers one-fifth of the area of the country and includes seven states (Bureau of Reclamation 1998). In consultation with the Colorado Basin RFC (CBRFC), nine forecast points located in the upper basin (comprising Colorado, Nevada, Utah, and Wyoming) and five in the lower basin (comprising Arizona, California, and New Mexico) were chosen (Fig. 1 and Table 1). The locations were selected because all are headwater reaches with an unregulated flow record, calibrated and formatted for use in the ESP forecasting system by the NWS, and important water supply forecast points within the region. The stream systems of the upper basin forecast points used in this study experience continuous baseflow and low discharge variance compared to the lower basin forecast points, which may have zero baseflow at times and are highly variable (Table 2).

b. Forecast generation

The NWS CBRFC issues volumetric water supply forecasts of naturalized flows for the Colorado River basin bimonthly beginning 1 January of each year through the end of the snowmelt and spring/summer rainfall seasons (Table 3). A volumetric forecast reports

TABLE 1. Descriptions of the forecast points and forecast data.

Forecast point	Basin size, sq. mi.	Gauge elevation, ft	Median discharge, AF	Evaluation period	Traces per forecast	Forecasts evaluated
Upper basin			(Apr–Jul)			
Crystal River, near Redstone, CO	174	6900	166 000	1956–97	48	42
East River, near Almont, CO	285	8000	183 700	1949–97	48	49
Fish Creek, near Steamboat Springs, CO	26	7200	37 200	1983–94	45	12
New Fork River, near Big Piney, WY	1215	6800	373 400	1955–94	43	40
White River, near Buford, CO	255	7000	131 600	1952–94	45	43
Animas River, near Durango, CO	705	6500	352 800	1949–94	45	46
Lemon Reservoir, Florida River Valley, CO	68	8100	55 000	1966–94	45	29
San Juan River, Pagosa Springs, CO	286	7100	198 400	1949–94	45	46
Vallecito Reservoir, Los Pinos Rv, near Bayfield, CO	252	7600	179 000	1965–94	45	30
Lower basin			(Jan–May)			
Gila River, near Gila, AZ	1853	4700	52 600	1949–93	44	45
Carizzo Creek, near Show Low, AZ	491	5000	18 600	1968–98	47	31
Tonto Creek, near Roosevelt, AZ	729	2500	44 400	1951–98	47	48
Verde River, near Paulden, AZ	2161	4100	8500	1964–98	47	35
North Fork of Virgin River, near Springdale, UT	348	4000	33 000	1951–98	47	48

TABLE 2. Basin statistics averaged over all forecast points and years used in this study.

	Upper basin	Lower basin
Mean annual temperature [°C (°*F*)]	7 (44)	14 (57)
Mean precipitation (mm yr)*	304	1000
Discharge	(Apr–July)	(Jan–May)
Mean (AF)	200 000	51 100
Median (AF)	186 000	31 400
Standard deviation (AF)	75 600	52 000
Maximum (AF)	367 500	214 600
Minimum (AF)	56 100	7500
Coefficient of variance	0.37	1.06

* Sheppard et al. (1999)

the total volume of water that is predicted to pass through a specific point on the river over a specific seasonal period, or forecast window (Brandon 1998). Approximately 75% of the annual streamflow discharge in the western United States comes from melting of mountain snowpack during the spring and summer (Palmer 1988), supplying the majority of the yearly water supply for the region. The forecast window is designed to account for the seasonality of the snowmelt and thus the source of water supplies and to take advantage of winter and early spring snowpack measurements. Upper basin forecasts report a forecast window covering the spring months, and lower basin forecasts report both winter and spring runoff with a variable forecast window. The combination of the forecast window and lead time is referred to as the forecast period.

The NWS ESP system uses the Sacramento Soil Moisture Accounting Model (SAC-SMA) (Burnash et al. 1973; Burnash 1995) and the SNOW-17 model (Anderson 1973), along with streamflow routing algorithms to simulate streamflow. Starting the models at current basin conditions, historical sets of temperature and precipitation time series, assumed to be a sample of possible future events, are input into the models to produce an ensemble of streamflow outputs (traces) (Fig. 2). The input data span only the same calendar days as those of the forecast period (e.g., 1 Jan–31 Jul). A model run stops on the last day of the forecast period, and the model states are reset to the current initial conditions before the next historical year's data are inputted; thus, each trace is conditioned on the current basin states. Statistical analysis of the ensemble's distribution results in a probabilistic forecast (Day 1985).

Although the CBRFC has been archiving the input data required for ESP for many decades, ESP output has not been systematically archived until recently making forecast evaluation difficult. The ESPVS was designed to produce historical ESP forecasts (hindcasts) for verification purposes and was employed to reconstruct water supply outlooks for the 14 study locations discussed above. All available temperature and precipitation data (required for forecast generation) and discharge data (required for verification) were used to generate and evaluate as many hindcasts as possible (Table 3). There were an average of 47 traces per ensemble forecast and an average of 39 forecast years studied for each location. The ESP forecast trace values examined in this study are discrete totals of seasonal streamflow, aggregated from daily values. The use of the ESPVS, NWS River Forecast System (NWSRFS) models, data and site files obtained from CBRFC, and the same forecast dates and lead times produced by CBRFC allowed the historical forecast generation process to resemble, as closely as possible, operational procedures of the NWS. It should be noted that, in operational forecasting, the model states might undergo real-time manual adjustments to reflect short-term meteorological forecasts and/or recent observations. These adjustments have an unknown effect on forecast quality and were not included in these hindcasts, because records of such adjustments do not exist for the time periods studied.

TABLE 3. Forecast information. The Virgin River is an exception in the lower basin. (Dates given as month/day.)

Day forecast issued	Forecast window	Length of forecast (days)	Length of forecast period (days)
	Upper basin		
1/1	4/1–7/31	122	212
1/15	4/1–7/31	122	198
2/1	4/1–7/31	122	181
2/15	4/1–7/31	122	169
3/1	4/1–7/31	122	155
3/15	4/1–7/31	122	139
4/1	4/1–7/31	122	122
4/15	4/15–7/31	108	108
5/1	5/1–7/31	92	92
5/15	5/15–7/31	78	78
6/1	6/1–7/31	61	61
	Lower basin		
1/1	1/1–5/31	151	151
1/15	1/15–5/31	137	137
2/1	2/1–5/31	120	120
2/15	2/15–5/31	106	106
3/1	3/1–5/31	92	92
3/15	3/15–5/31	78	78
4/1	4/1–5/31	61	61
	Virgin River (Lower basin)		
1/1	4/1–7/31	122	212
1/15	4/1–7/31	122	198
2/1	4/1–7/31	122	181
2/15	4/1–7/31	122	169
3/1	4/1–7/31	122	155
3/15	4/1–7/31	122	139
4/1	4/1–7/31	122	122
4/15	4/15–7/31	108	108
5/1	5/1–7/31	92	92

A deterministic outlook is a forecast that predicts a single value of a variable (Croley 2000); with respect to ESP, a deterministic outlook could be obtained, for example, from choosing a single ensemble member. The streamflow volume forecast traces that were singled out for analyses were the median, the minimum-error, and the historical. The median trace was chosen from each ensemble to examine the effects of transforming the

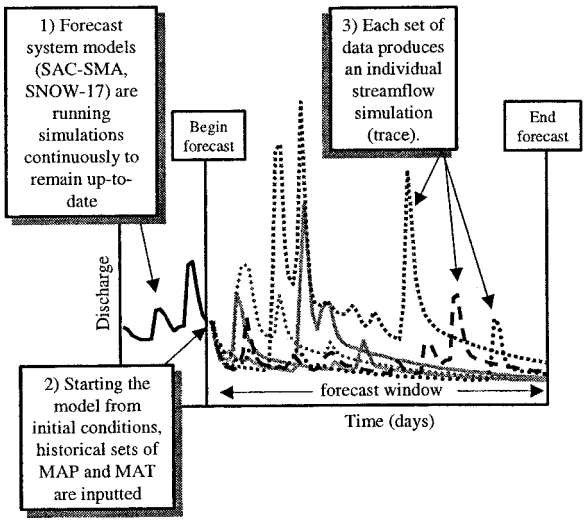

FIG. 2. Diagram of an ESP forecast and description of parts. MAP = mean areal precipitation. MAT = mean areal temperature. SAC-SMA = Sacramento Soil Moisture Accounting Model.

probabilistic ESP forecasts into deterministic forecasts and was considered the most appropriate ensemble statistic given the often-skewed distribution of streamflow. The minimum-error trace is the ensemble member that produces the lowest absolute error in predicted seasonal volume. Finally, the historical trace represents the modeling system accuracy when simulating actual observations. This trace is the result of running the models with temperature and precipitation observations from the forecast year itself. The accuracy, or inaccuracy, of the ''historical'' trace reflects errors in the input data, model structure, model calibration of parameters, and initial states at the beginning of the run.

A probabilistic forecast provides a predicted value, or values, of a variable and the associated distribution function that reflects the likelihood of the event (Croley 2000); a probabilistic ESP forecast results from considering the distribution of the entire ensemble. To generate forecast probability from the ESP ensembles, the cumulative distribution function of all available historical observations (climatology) was used to predetermine threshold streamflow values for non-exceedance probability categories. Based on the probability intervals traditionally referred to in historical official forecasts, the thresholds were set at 10%, 30%, 70%, and 90%, resulting in five intervals (0%–10%, >10%–30%, >30%–70%, >70%–90%, and >90%–100% nonexceedance). Forecast traces were placed into these categories according to their individual values. The probabilistic forecast for a given forecast period was obtained from calculating the relative frequency of the traces in each category.

c. Traditional statistical analysis

Three commonly used forecast evaluation statistics were used for analysis of the deterministic ESP forecasts. The mean absolute error (MAE) is a measure of the average correspondence between forecast and observed seasonal water supply values. The MAE was divided by the standard deviation of the respective observations [relative-mean absolute error (RMAE)] to allow comparison among forecast points; the optimal value of the RMAE is 0. Percent Bias (PBias) measures the difference between the average forecasted and the average observed seasonal water supply values (Wilks 1995). A PBias of 0% is desirable. A positive PBias indicates that forecasts tend to assign forecast values that are greater than the observations (overforecasting); a negative PBias indicates underforecasting. The correlation coefficient (R) is a measure of how the forecasts and observations vary together and is the ratio of the sample covariance to the product of their standard deviations. A perfect score of R equal to 1 indicates that the forecasts and observations vary linearly.

d. Probabilistic verification measures

Probabilistic verification methods have been used in the evaluation of meteorologic and climate forecasts (Murphy et al. 1989; Wilks 2000; Hartmann et al. 2002b); however, they have not been used extensively in the field of hydrology. There are many measures that can be used to evaluate forecasts (Wilks 2000; Croley 2000); however, a comparative analysis of all their attributes or application to the hindcasts is beyond the scope of this paper. In general, the statistics chosen show a progression in information content from simple measures to the more complicated distribution-oriented measures. The measures outlined in this section [ranked probability score (RPS), discrimination, and reliability] have been ''field-tested'' with stakeholders to ensure that users could understand the information that the verification measures provided and the practical implications for real-world decisions (Hartmann 2001).

RPS was used to assess the overall performance of the probabilistic forecasts (Epstein 1969; Wilks 1995). To calculate the RPS, the ensemble members were distributed into the streamflow nonexceedance categories discussed in part *b* of this section. The forecast cumulative distributions (F_m) were then calculated:

$$F_m = \sum_{j=1}^{m} f_j, \qquad m = 1, \ldots, J, \tag{1}$$

where f_j is the relative frequency of the forecast traces, and J is the number of nonexceedance categories (Wilks 1995). The observation (o) occurs in only one of the flow categories, which is given a value of 1; the remaining categories are given a value of 0. The cumulative distribution of the observations was then calculated (Wilks 1995):

$$O_m = \sum_{j=1}^{m} o_j, \qquad m = 1, \ldots, J. \tag{2}$$

The RPS for one forecast is the sum of the squared differences of the cumulative distributions:

$$\mathrm{RPS} = \sum_{m=1}^{J} (F_m - O_m)^2. \quad (3)$$

For a group of n forecasts, the RPS is the average ($\overline{\mathrm{RPS}}$) of the n RPSs:

$$\overline{\mathrm{RPS}} = \frac{1}{n} \sum_{k=1}^{n} \mathrm{RPS}_k. \quad (4)$$

A perfect forecast would assign all of the probability to the same streamflow category in which the event occurs, resulting in an RPS value of 0 (Wilks 1995).

RPS is calculated in much the same way as the Brier score (Brier 1950; Wilks 1995) (mean square error); however, the RPS allows multiple observation categories and cumulative forecast probabilities to be considered at once. In addition, the RPS is said to be ''sensitive to distance'' because it increasingly penalizes forecasts that assign probability to streamflow categories further from the observation (Wilks 1995). In contrast, the Brier score focuses only on one category, lumping the probability for all other categories while examining one (Hartmann et al. 2002b). RPS is useful to decision makers interested in overall forecast quality rather than how the forecasts perform in a particular flow category.

The quality of forecasts is difficult to assess based on the RPS alone (Wilks 1995); therefore, the ESP forecasts were compared to a reference forecast. Due to a lack of other available probabilistic streamflow forecasts, climatology forecasts were generated from the historical observations to serve as a reference forecast (the climatological periods were equal to the evaluation periods in Table 1). The relative skills of the ESP forecasts were evaluated against the climatology forecasts through the use of the ranked probability skill score (RPSS):

$$\mathrm{RPSS} = \frac{\overline{\mathrm{RPS}}_f - \overline{\mathrm{RPS}}_{\mathrm{cl}}}{0 - \overline{\mathrm{RPS}}_{\mathrm{cl}}} \times 100\%, \quad (5)$$

where $\overline{\mathrm{RPS}}_f$ is the average RPS of the forecasts for a particular forecast period, and $\overline{\mathrm{RPS}}_{\mathrm{cl}}$ is the average RPS of the climatology forecasts for the same period (Wilks 1995). A positive RPSS indicates that the forecast of interest more closely predicted the observation than the climatology did, which is defined as ''improvement over climatology.'' A perfect RPSS score is 100%. A negative RPSS indicates that the ESP forecasts performed worse than climatology.

Discrimination and reliability (Murphy and Winkler 1992, 1987; Murphy et al. 1989; Wilks 1995) were used to assess the prediction capabilities of the forecasts in specific categories. The streamflow volume categories examined in this part of the study were the lowest 30%, middle 40%, and highest 30% of the historical distributions and are referred to as low-, middle-, and high-flow categories. The same five forecast probability categories used for RPS were used to represent the magnitude of the probability given to each of the three flow categories.

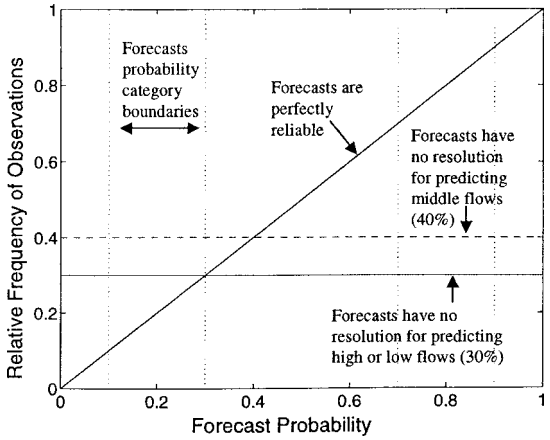

FIG. 3. Example reliability diagram describing the behavior of forecasts that fall in particular regions of the diagram. The light vertical lines demarcate forecast probability categories.

Reliability summarizes the information contained in the conditional distribution [$p(o \mid f)$] and describes how often an observation occurred given a particular forecast. Ideally:

$$p(o = 1 \mid f) = f \quad (6)$$

(Murphy and Winkler 1987). That is, for a set of forecasts where a forecast probability value f was given to a particular observation o, the forecasts are considered perfectly reliable if the relative frequency of the observation equals the forecast probability (Murphy and Winkler 1992). For example, given all the times in which high flows were forecasted with a 50% probability, the forecast system would be considered perfectly reliable if the actual flows turned out to be high in 50% of the cases.

The reliability diagram is used to display forecast reliability (Fig. 3). The conditional distribution [$(p(o \mid f)$] of a set of perfectly reliable forecasts will fall along the 1:1 line on the diagram. Forecasts that fall to the left of the perfect reliability line are underforecasting or not assigning enough probability to the subsequent observation. Those that fall to the right of the line are overforecasting. Forecasts that fall on the no-resolution line are unable to identify occasions when the event is more or less likely than the overall climatology (Wilks 1995). Conditional distributions of forecasts lacking resolution plot along the horizontal line associated with their climatology value.

As forecasts become sharper, or more refined, the forecast probability becomes more narrowly distributed and is more frequently assigned to the extreme nonexceedance categories (i.e., 0%–10% and >90%–100%) (Murphy et al. 1987). Thus, the sample sizes within the middle probability categories become smaller with sharper forecasts. A relative frequency diagram displays

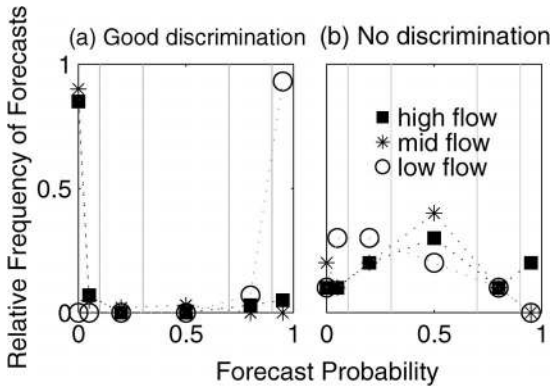

FIG. 4. Examples of discrimination diagrams for forecasts issued prior to low-flow observations: (a) forecasts properly discriminated for low flows (low flows are forecasted with 90%–100% probability, and middle and high flows are forecasted with 0% probability, at close to 100% frequency), and (b) forecasts that have no discrimination for low flows (all flow categories tend to be predicted with similar probabilities).

forecast resolution and also allows the user to determine which reliability results may be most valid based on the sample size within the probability category (bin). Statistics calculated from a small number of forecasts are more susceptible to being dominated by sampling variations and make assessing forecast quality difficult (Wilks 1995). In addition, with smaller sample sizes, it is more likely that some bins have no data because there are not enough forecasts to represent all combinations of forecast probability and flow categories, resulting in erratic-looking diagrams.

The likelihood that a particular forecast would have been issued prior to a specific observation is expressed in the conditional distribution of the forecasts given the observed category [$p(f \mid o)$] (Wilks 1995). If the value of $p(f \mid o)$ for a particular observation category is similar to that for a different observation, the forecasts are not discriminatory for that observation. On the other hand, when $p(f \mid o)$ equals zero for all possible observations except one, the forecast procedure is perfectly discriminatory for forecasts of that observation (Murphy and Winkler 1987).

The discrimination diagram displays the conditional probability distributions [$p(f \mid o)$] of each possible flow category as a function of forecast probability (Fig. 4). Note that each diagram includes only forecasts issued prior to a specific observation. If the forecasts are discriminatory, then the probability distribution functions of the forecasted flow categories will not overlap to a great degree on the discrimination diagram (Murphy et al. 1989). Ideally, a forecast issued prior to an observation of a low flow should say that there is 100% chance of having a low flow and 0% chance of having high or middle flows. A set of forecasts that consistently provide such strong and accurate statements will produce a discrimination diagram similar to that given in Fig. 4a. If there is little discrimination, then there will be considerable overlapping of the probability distributions (Murphy et al. 1989). A case where the sample of forecasts is unable to consistently assign the largest probability to the occurrence of low flows versus the other two is illustrated in Fig. 4b. Users of forecasts from such a system could have no confidence in the predictions.

A discrimination diagram is produced for occurrences of observations in each flow category; therefore, forecasts that were issued prior to observations that occurred in the lowest 30% (low flows) are plotted on a separate discrimination diagram than forecasts that were issued prior to observations that occurred in the middle 40% (midflows), etc. The number of forecasts represented on each plot is dependent upon the number of historical observations in the respective flow category.

Discrimination and reliability provide the user with comprehensive forecast evaluations and allow performance in all streamflow categories to be examined individually. In addition, these forecast quality measures examine the actual probability value within each category in contrast to ''hit'' and ''miss'' scores that convert the probability to an implied 100% probability for the category of interest (Hartmann et al. 2002b). However, their sensitivity to small sample sizes is an acknowledged limitation.

3. Results

Because of the large number of forecast locations, forecasts, and statistics included in this study, it is impossible to show all of the results from each analysis. Wherever possible, statistics for all locations are provided for select dates. When inclusion of all locations was impractical, results for four basins (Gila River, Verde River, Animas River, and East River) were provided. These locations are representative of both the upper and lower basins and are of interest to forecast users for a variety of reasons, including recreation, water supply, and power generation. In addition, the East River has the largest sample size of all the locations (49 years), and the Verde River emphasizes forecast system issues. Results for all locations and issue dates studied can be found in Franz et al. (2003).

a. Traditional statistical analysis

PBias and correlation coefficients are provided for the median trace forecasts for three dates: the first and last forecasts of the season, and a midseason forecast that coincides with the typical occurrence of significant snowmelt within the respective basins (Figs. 5 and 6). Overall, both the upper and lower basin median forecasts showed a trend toward improved performance as the season progressed. In addition, the forecasts for the upper basin locations performed better than those in the lower basin. For most locations and lead times, the median underforecasted the seasonal volumes; the bias was

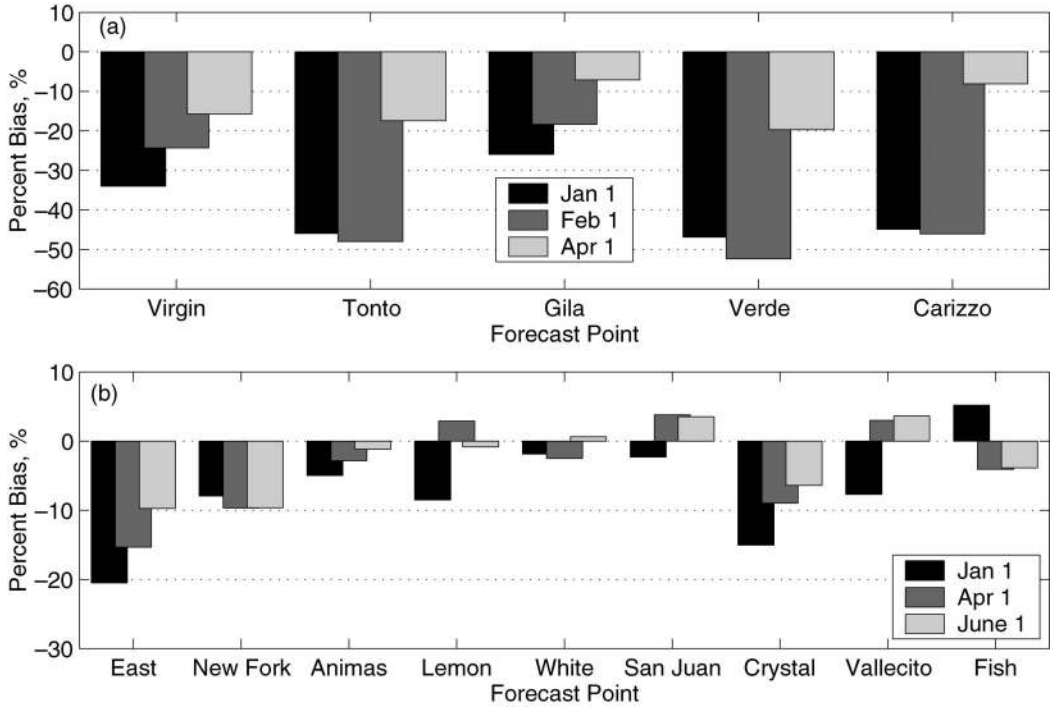

FIG. 5. Percent bias for the median volume forecast for (a) the lower basin forecast points and (b) the upper basin forecast points. Note scale difference between the two plots.

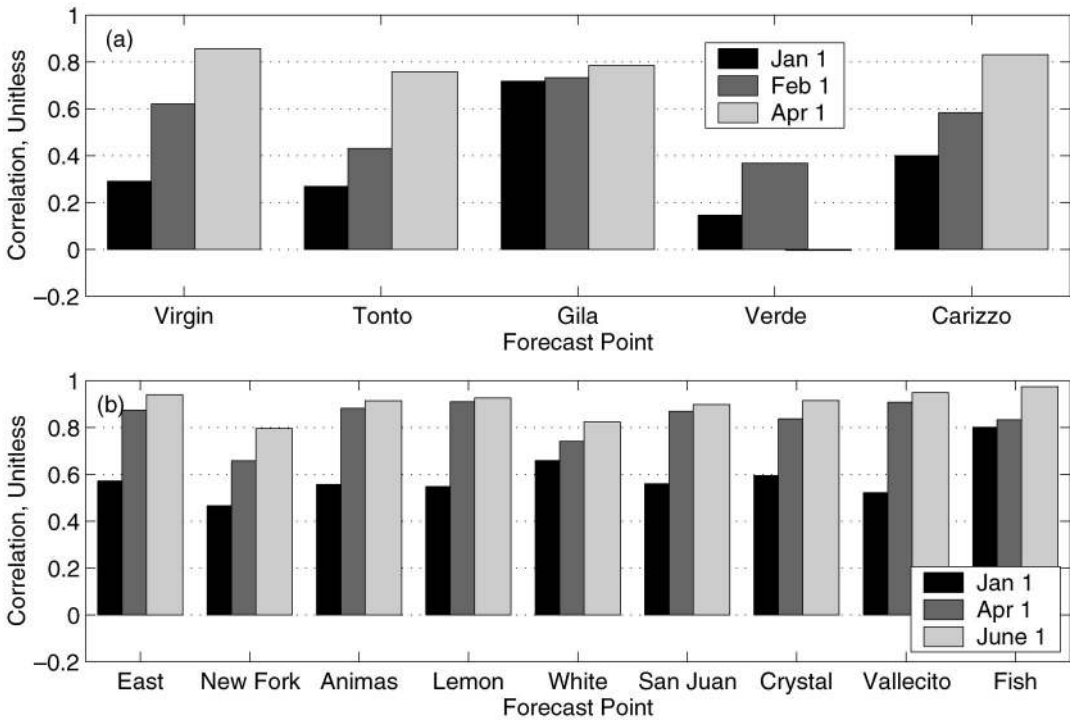

FIG. 6. Correlation coefficient for the median volume forecast for (a) the lower basin forecast points and (b) the upper basin forecast points.

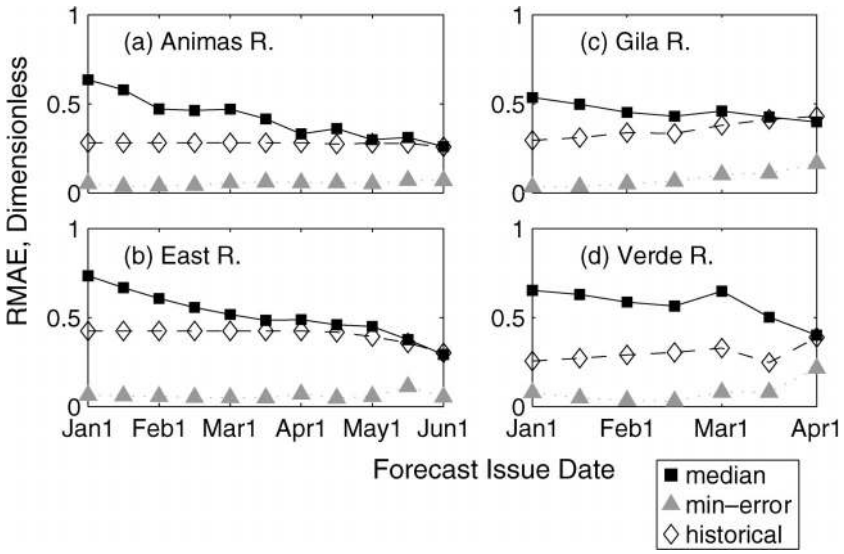

FIG. 7. Average mean absolute error divided by the standard deviation of the observations (RMAE) for the median trace forecasts, minimum-error (min-error) traces, and historical traces.

highest for Tonto Creek, Carizzo Creek, and the Verde River. Correlation between the median and observation increased as lead times decreased for all forecast points except for the Verde River, which showed very low values throughout. In general, correlation values were higher in the upper basin than in the lower basin.

Median forecasts for both basins show a trend toward improved performance (lower RMAE) into the forecast season (see Fig. 7 for example basins). The upper basin median forecasts had higher errors than the lower basin in the beginning of the season, but showed similar accuracy during the 1 March–1 April period. The RMAE of the minimum-error traces increased with shorter lead times for the Gila and Verde rivers, but were lower than the historical and median traces throughout the season. The RMAE of the historical traces, which represent the quality of the calibration, were relatively constant for the upper basin locations and decreased for the lower basin during the March–April simulations.

b. Probabilistic verification

Overall ESP forecast performance, analyzed using the RPSS, was better than the climatology forecasts for all basins except the Virgin and the Verde rivers at the start of the forecast season (Fig. 8). The upper basin forecasts performed better than the lower basin forecasts and on average improved as the season progressed. In the lower basin, only the Virgin River and Tonto Creek showed marked improvement in the RPSS from the first to the last forecast. The Verde River forecasts never showed improvement over climatology at any time.

In general, the reliability diagrams for the Animas River forecasts are typical of those for other upper basin points and will be used to illustrate the results for this section (Fig. 9). Similarly, the Gila River forecasts are typical of those for other lower basin locations (Fig. 10). The forecasts in both the upper and lower basin display fair reliability early in the forecast season, particularly for forecast probabilities less than 70% (Figs. 9 and 10a,b,c). By 15 March in the upper basin and 1 February in the lower basin, the forecasts display near perfect reliability for low flows and a tendency to overforecast middle flows and underforecast high flows (Figs. 9 and 10d,e,f). In general, reliability of forecasts issued after these dates decrease (Figs. 9 and 10g,h,i); however, forecasts that assign less than 10% or greater than 90% nonexceedance probability continue to show good reliability.

Plots of the relative frequency of the forecasts are embedded within the reliability diagrams. As indicated by the increased bin sizes in the 0%–10% and >90%–100% probability categories for 1 June and 1 April forecasts, late season forecasts tend to give extreme probability values most frequently and therefore display high resolution and confidence. In contrast, early in the season (1 January), the forecasts almost exclusively assign probability in the middle to low forecast probability categories. The problems that arise when sample sizes within the bins are small are illustrated in Fig. 9i; the data look scattered because the three center bins contain a small sample of forecasts, some of which obviously performed poorly with respect to reliability.

Early season ESP forecasts issued prior to high flows for the Animas and Gila rivers show little discrimination between the likelihoods for middle and high flows but some degree of discrimination between the likelihoods for high and low flows, which was typical of many forecast points (Figs. 11d,g). Forecasts issued midseason and later seldom assign probabilities greater than

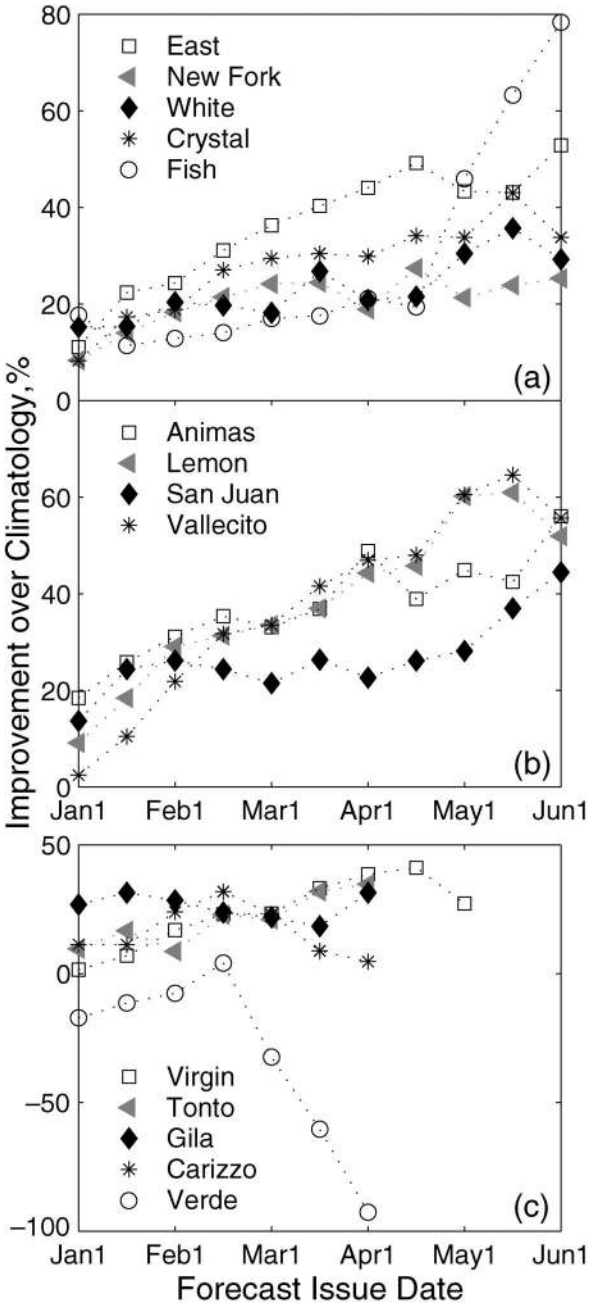

FIG. 8. RPSS for (a) and (b) the upper and (c) the lower basin forecast points. RPSS is interpreted at the percent improvement over climatology forecasts and a value greater than zero is desired.

10% to the occurrence of low flows (Figs. 11b,e,h). The forecasts do not show perfect discrimination for high flows at any time, because probability >30% is occasionally given to flows in the middle category throughout the season. Discrimination of the Verde River forecasts for high flows was significantly poorer than all other locations, especially for late-season forecasts (Figs. 11k,l). There was a general trend for all lower basin forecasts to display a decrease in discrimination for high flows after 1 March, as illustrated by the Gila River (Figs. 11h,i) (about 25% of the time, the forecasts predict middle flows with 90%–100% probability). Discrimination results for forecasts issued prior to flows in the middle 40% category are on average worse than those for the other two categories. Results for this category are not shown but can be found in Franz et al. (2003).

Forecasts issued prior to low-flow observations showed some discrimination between the likelihoods for low flows versus high flows early in the season; however, there is little discrimination between the likelihoods for middle and low flows in most basins (Fig. 12). The later upper basin forecasts improve and, by 1 April, never assign probabilities higher than 30% to high flows (Figs. 12b,e). By 1 June, the forecasts are infrequently assigning high probabilities to middle flows (Figs. 12c,f). There is no discrimination between the likelihoods for the middle and low flows for most lower basin forecasts until late in the season, and there are still problems as late as 1 April (Fig. 12i). The Verde River forecasts perform relatively well for discrimination of low flows (Figs. 12k,l) in contrast to some of the poorer verification results seen earlier for this location.

4. Discussion

For the basins studied, the upper basin forecasts performed better than the lower basin forecasts, even at the longer lead times. Because the upper basin hydrology is less variable, this result is not unexpected. Shafer and Huddleston (1984) found that the highest errors of the historical regression forecasts occurred for forecast points in Arizona (in the lower Colorado basin) and concluded that forecast accuracy potential is highly dependent on the variability of the streamflow.

In general, statistics for the probabilistic ESP forecasts improved as the season progressed and forecast lead times became shorter. Day et al. (1992) stated that, at the beginning of the season, the future meteorology is the main source of uncertainty. As the season progresses, the relative importance of the initial states and meteorological inputs changes. With shorter model runs, the initial conditions dominate, because there is less opportunity (time) for the meteorological inputs to overcome the influence of the initial state values. Because the actual meteorology experienced by the basin becomes ''stored'' in the model states as snowpack, baseflow, and soil storages, it is expected that the forecasts would become more accurate later in the season when the initial states are better known. However, forecasts for some lower basin locations tended to show a reduction in accuracy for late season forecasts. Because the majority of the snowpack in these locations has melted by this time, the impact of snowpack on streamflow is minimal. Without the large snowmelt runoff dominating the projected hydrograph and dampening the effects of individual precipitation events, the value of the

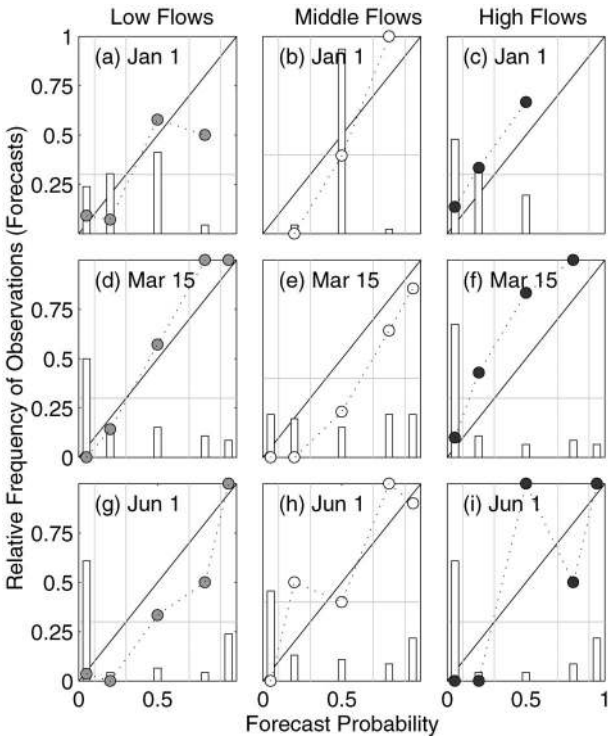

FIG. 9. Reliability diagrams for the (a)–(c) 1 Jan, (d)–(f) 15 Mar, and (g)–(i) 1 Jun forecasts issued for the Animas River. Reliability data are plotted as circles, and the y axis plots the relative frequency of observations. Diagrams depict forecast probability assigned to low flows (lowest 30%), middle flows (middle 40%), and high flows (highest 30%). The relative frequency of the forecasts is displayed by the bar graph.

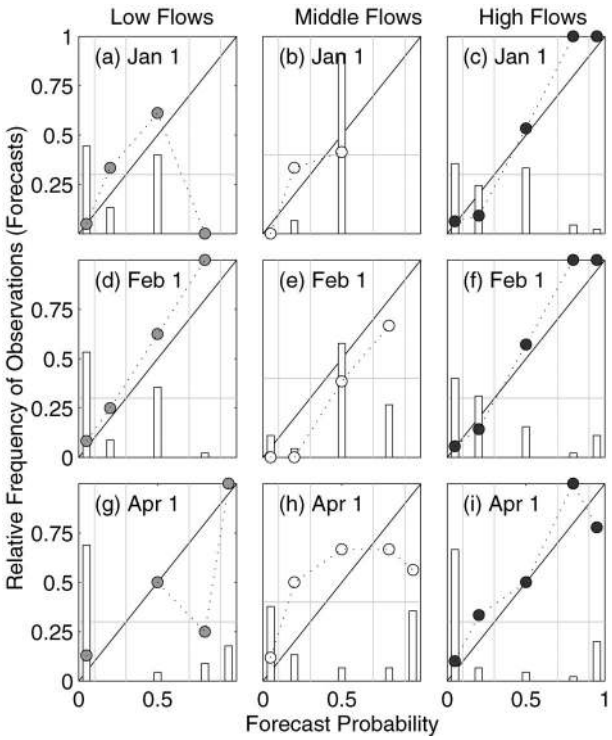

FIG. 10. As in Fig. 9, except reliability diagrams for the (a)–(c) 1 Jan, (d)–(f) 1 Feb, and (g)–(i) 1 Apr forecasts issued for the Gila River.

ensemble members may become more disparate, producing ensembles that reflect meteorological variability rather than the initial conditions. Further investigation is needed to better understand the relative importance of the snowpack and initial states on forecast accuracy.

The median forecasts showed large biases in predicted seasonal volumes for several forecast points, particularly those in the lower basin. The performance of the median with respect to PBias, correlation, and RMAE improved with decreased lead time. Analysis of the median and minimum-error traces revealed that there was at least one other trace that performed better than the median, which alone is not surprising. However, it was surprising that the minimum-error trace occurred consistently closer in value to the observed volume than the historical trace. This result indicates that there are forecast system errors (model, calibration, data, initial states) such that meteorology from a totally unrelated year produces a trace that matches the observation more closely than the trace produced using meteorology from the forecast year itself. Information available from comparing the historical trace to the best trace would be useful to forecasters and modelers for investigating modeling improvements and the effects of system errors on forecast accuracy.

The minimum-error trace is not a viable operational forecast because it is identified by comparison to the observation and cannot be determined prior to the occurrence of the observation. However, it is useful for revealing forecast system performance issues; for example, if the minimum-error trace performed the best because of system biases rather than the input meteorology being similar to the forecast year, a basis for forecast or system adjustments can be developed. If the minimum-error trace is found to consistently occur within the same percentile (e.g., the 60th percentile trace generally has the lowest absolute error), models or parameters could be altered to reflect this information. Additionally, the streamflow volume indicated by the mean percentile of the minimum-error traces could be used as an expected-value alternative to a biased median trace.

Overall, the probabilistic ESP forecasts provided more accurate forecast information than could be obtained from climatology, as illustrated by the RPSS statistics. Forecast reliability peaks at about 15 March for all basins and then appears to become worse as the season progresses, particularly in the middle forecast probability categories. As mentioned, empty or low sample sizes can cause irregular-looking diagrams. Even though there was a ''large'' sample set for the Animas River, compared to what is generally available for verification of operational forecasts, the effect of the low

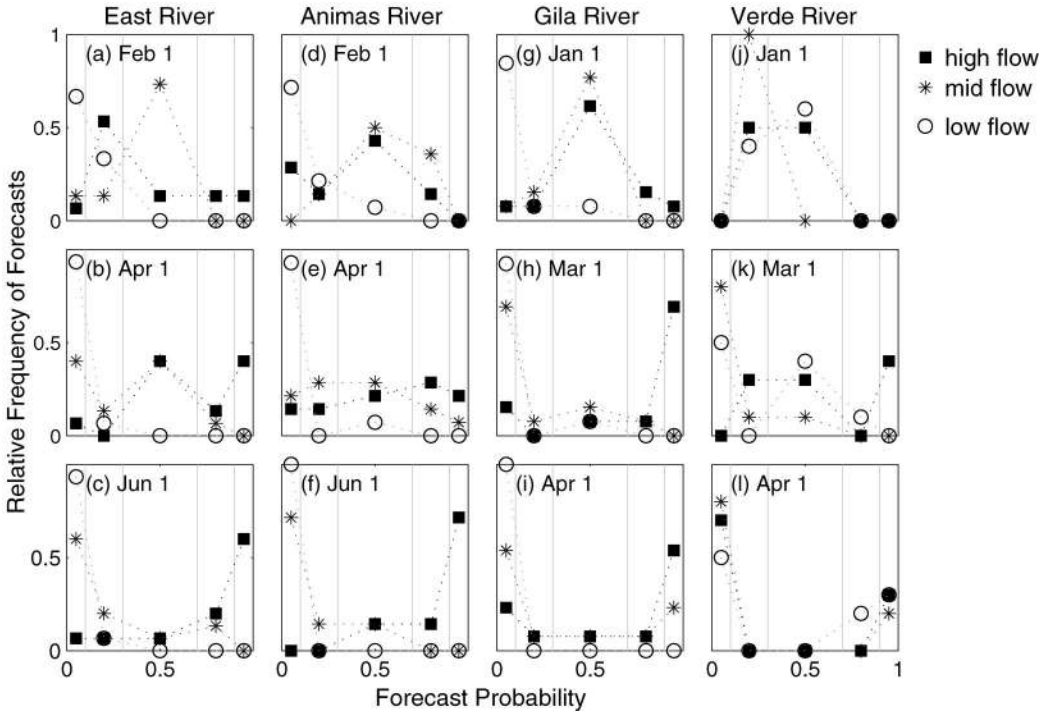

FIG. 11. Discrimination diagrams for forecasts issued prior to high-flow observations (highest 30% of streamflow distribution) for issue dates shown. The number of forecasts for each basin are: East, 15; Animas, 14; Gila, 13; and Verde, 10.

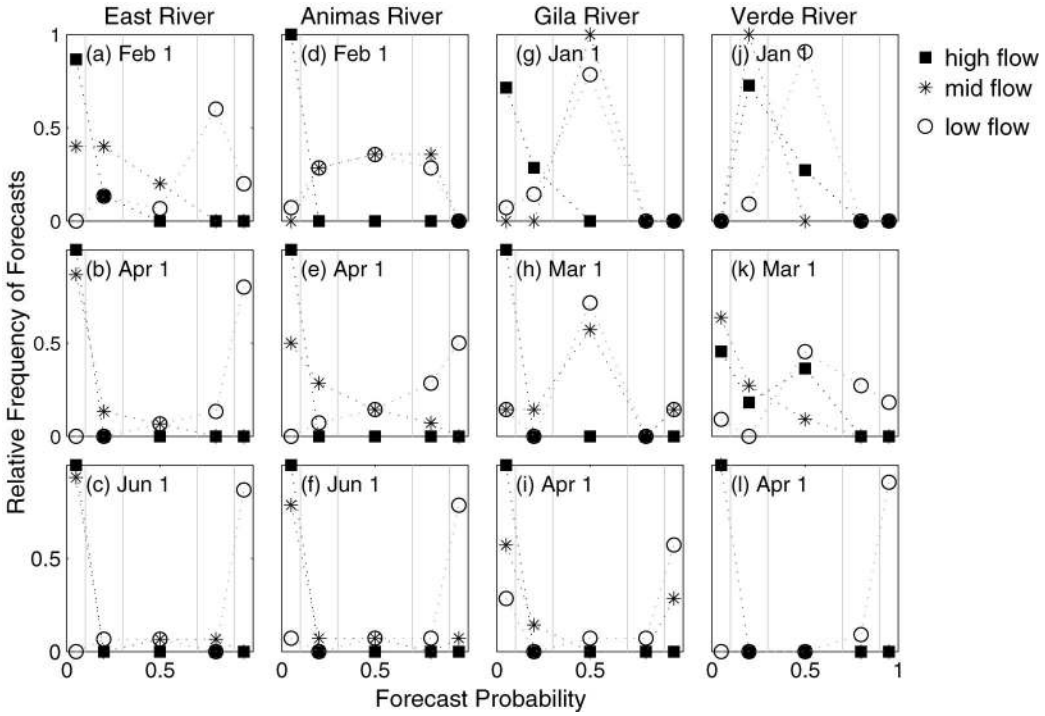

FIG. 12. Discrimination diagrams for forecasts issued prior to low-flow observations (lowest 30% of streamflow distribution) for issue dates shown. The number of forecasts for each basin are: East, 15; Animas, 14; Gila, 14; and Verde, 11.

sample size in probability categories between 10% and 90% is apparent. Because the forecast resolution increases for late season forecasts, fewer forecasts were made with middle range probabilities. As a result, it is difficult to determine whether the poor reliability in the middle probability categories is a result of problems with the forecast system or a poor sample. However, from a practical perspective, the lack of middle range probabilities is nonproblematic because the extreme probabilities have high reliability (i.e., strong probability statements with high reliability are preferable to weak probability statements with high reliability).

Early season ESP hindcasts were not able to fully discriminate among the likelihoods for high, middle, or low flows, thus giving no indication which type of flow was most likely. The forecasts do, however, clearly indicate a tendency to give lower likelihood to the extreme opposite flows for some forecasts (i.e., low flows were given low probability for cases when the subsequent observation was high). This information may give forecast users an early indication of what type of flow conditions *not* to expect. For example, during a several year period of drought, forecasts that tend to properly predict that high flows will not occur at 7- to 11-month lead times would give water resource managers an early indication that a large runoff should not be anticipated. Although the flow intervals used in this study may not be precise enough for some applications, the distribution-oriented measures are flexible enough that the bins can be adjusted as necessary. Later in the season, discrimination improves, providing more accurate predictions for users (such as flood managers) that require short lead times. The ESP hindcasts were limited in their ability to give useful information for the prediction of middle flows; however, this flow range is likely of least concern to forecast users because it represents ''normal'' conditions, where the consequences of uninformative forecasts are least problematic.

While creating a deterministic forecast simplified the statistical analysis and allowed the use of traditional statistics with which many people are familiar, these statistics are deficient for fully analyzing ESP forecasts. The traditional statistical methods applied to the median, minimum-error, and historical volume traces only evaluate whether the forecast is right or wrong (Wilks 1995). Distribution-oriented measures provided a method for verifying the probabilistic ESP hindcasts, which are considered to be never completely right or wrong. In addition, it was shown that the median volume trace did not most closely predict the observation; therefore, using only this value gave a suboptimal forecast. More important, considering only a single trace ignored useful predictive information in the entire ensemble distribution as illustrated by the skill in the probabilistic hindcasts.

While this paper attempts to advance discussion about evaluation of probabilistic hydrologic outlooks, a comprehensive discussion about forecast performance should also consider confidence limits on the estimated forecast quality measures. This aspect of verification was not addressed here and represents an important next step in the verification process. In addition, the use of climate forecasts, El Niño–Southern Oscillation states, or other climate information for generating trace weighting schemes comprise important opportunities for advancing ESP forecasting research. Although beyond the scope of this work, initial investigations indicated that trace weighting based on the ENSO state improved the RPS in both the upper and lower basins (Franz et al. 2003).

5. Conclusions and summary

Probabilistic water supply forecast capabilities, based on conceptual models and an ensemble approach, have evolved in the NWS to the extent that forecast generation is operationally feasible. ESP forecasts require more data processing and modeling than the regression-based products currently issued as the official forecasts, but theoretically offer advantages through their more sophisticated incorporation of current basin states and meteorological uncertainty. However, lack of verification precludes routine issuance and application of ESP products with any quantitative basis for confidence. Using hindcasting to simulate operational forecast generation as closely as possible, we evaluated ESP forecasts for nine headwater locations in the upper Colorado River basin and five in the lower basin, all unaffected by streamflow regulation. Probabilistic forecasts have fundamentally different character than single-value deterministic forecasts, yet need to be compared to traditional forecasts in judging which products warrant continued generation or are better for different applications.

We evaluated the ESP hindcasts using a mix of verification criteria, including traditional statistics and distribution-oriented measures. Evaluations based on traditional statistics require selection of a single value to represent the entire forecasted distribution of potential streamflow volumes. The median forecast volume is an intuitive choice for representing the entire distribution, but our analyses confirm that such forecasts are biased and suboptimal compared to forecasts derived from other distribution percentiles. Further, examination of specific ensemble traces (e.g., minimum-error and historical traces) can provide insight about the limitations of the forecast system and process, including proper model identification and parameterization, and the respective roles of initial conditions and meteorological uncertainty in affecting basin response.

However, single-value forecasts necessarily ignore important information embodied in the entire distribution produced by use of ESP techniques, and from that perspective we recommend that they not be considered the standard ESP forecast product. Instead, we recommend that forecasting agencies issue probabilistic forecast products that describe the entire forecasted distri-

bution, or at least several portions of the distribution that have meaning for practical applications. Evaluation of such forecasts requires techniques not typically applied to hydrologic forecasts, such as the RPSS and distribution-oriented measures. The RPSS is most useful for decision makers concerned with forecast performance across the full range of possible conditions, rather than performance focused only on specific conditions (e.g., high or low flows). Distribution-oriented measures (e.g., reliability and discrimination) provide the most comprehensive evaluation of forecast characteristics, but are most affected by small sample sizes. RPSS, reliability, and discrimination are practical for real-time computation within the NWSRFS and ESPVS framework, and we recommend that frequently updated verification statistics be issued with any operational probabilistic forecast products. From a user's perspective, a good option would be an interactive Web site that allows users to evaluate hindcasts and forecasts that cover the periods and lead times relevant to their situation, using the specific forecast performance measures that reflect their sensitivity to different forecast qualities.

Our ESP forecast evaluations provide insights about the forecast system and performance of interest to forecasters and water resource decision makers. Clearly, forecasts for the Verde River should not be relied upon for decision making until fundamental issues of data quality or model identification and parameterization can be resolved. In general, the headwater locations showed different forecast performance behavior across the upper and lower Colorado basins, but common behavior within the basins. Overall, forecasts are better for locations in the upper basin, and forecasts issued 15 March and later are generally the best. However, most locations in the lower basin did show forecast skill, compared to the use of climatology forecasts based on historical streamflow volumes, even for the earliest forecast issue dates.

The ESP hindcasts showed good reliability for most locations and forecast issue dates, with the caveat that additional work is required to develop confidence limits for the reliability statistics. With the same caveat, the hindcasts show that discrimination is excellent for late season forecasts. Additionally, as indicated by the discrimination diagrams, forecasts for all locations were able to indicate that extreme opposite conditions were *less* likely to occur even in forecasts issued 1 January, representing important information for water resource applications unavailable in extant products. Further improvements in the probabilistic forecasts may be possible by adjusting forecast ensembles to reflect climate forecasts or persistent forcing, or combining ensemble and regression-based forecasts.

Based on our hindcast evaluations, we recommend that water managers begin to consider probabilistic forecasts in their operations. Water managers should also consider how to exploit nontraditional information, such as embodied in discrimination diagrams, in their operations; they should also begin to distinguish among the confidence levels required for each of their myriad decisions (e.g., emergency preparedness versus reservoir releases). Because probabilistic forecasts, and their verification criteria, are different in character than traditional forecast products, we recommend education efforts focused on the proper interpretation, evaluation, and application of new products.

Acknowledgments. Primary financial support from NOAA's Office of Global Programs through the Climate Assessment for the Southwest Project Grant NA86GP0061 is gratefully acknowledged. Partial support was also provided by the NSF Science and Technology Center for Sustainability of Semi-Arid Hydrology and Riparian Areas (Grant EAR-9876800), and the National Weather Service (Grant 40-AA-NW-217447). Special thanks are due to E. Welles, H. Herr, D. Brandon, and C. McCarthy of the National Weather Service for providing their technical support, data, and resources.

REFERENCES

Anderson, E. A., 1973: National Weather Service River Forecast System-Snow Accumulation and Ablation Model. NOAA Tech. Memo. NWS Hydro-17, U.S. National Weather Service. [Available from Office of Hydrologic Development, NOAA/NWS, 1325 East–West Highway, Silver Spring, MD 20910.]

Brandon, D., 1998: Forecasting streamflow in the Colorado and Great Basins using 'El Niño' indices as predictors in a statistical water supply forecast system. *Proc. Flood Plain Management Association Spring Conference: Winter '97–'98: Year of the Great El-Nino?,* San Diego, CA, Flood Plain Management Association.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.,* **78,** 1–3.

Bureau of Reclamation, 1998: Colorado River System consumptive uses and losses report: 1986–1990. U.S. Department of the Interior Rep., 40 pp. [Available online at http://www.usbr.gov/uc/library/envdocs/reports/crs/reports/1986.pdf.]

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

——, R. L. Ferrel, and R. A. McGuire, 1973: A generalized streamflow simulation system: Conceptual modeling for digital computers. Report of the Joint Federal–State River Forecast Center, Department of Water Resources, State of California and the National Weather Service, 204 pp.

Croley, T. E., 2000: *Using Meteorology Probability Forecasts in Operational Hydrology.* American Society of Civil Engineers Press, 206 pp.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plann. Manage.,* **111,** 157–170.

——, L. E. Brazil, C. S. McCarthy, and D. P. Laurine, 1992: Verification of the National Weather Service extended streamflow prediction procedure. *Proc. AWRA 28th Annual Conf. and Symp.,* Reno, NV, American Water Resource Association, 163–172.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Evaluation of National Weather Service ensemble streamflow predictions for the Colorado River Basin. Tech. Rep. HWR 03-010, Dept. of Hydrology and Water Resources, The University of Arizona, Tucson, AZ, 177 pp.

Hartmann, H. C., 2001: Stakeholder driven research in a hydrocli-

matic context. Ph.D. dissertation, The University of Arizona, Tucson, AZ, 256 pp.
——, R. Bales, and S. Sorooshian, 1999: Weather, climate and forecasting for the southwest U.S. Rep. Series CL2-99, Institute for the Study of Planet Earth, The University of Arizona, Tucson, AZ, 172 pp.
——, ——, and ——, 2002a: Weather, climate, and hydrologic forecasting for the U.S. Southwest: A survey. *Climate Res.*, **21,** 239–258.
——, T. C. Pagano, R. Bales, and S. Sorooshian, 2002b: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.*, **83,** 683–698.
Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115,** 1330–1338.
——, and ——, 1992: Diagnostic verification of probability forecasts. *Int. J. Forecasting*, **7,** 435–455.
——, B. G. Brown, and Y. Chen, 1989: Diagnostic verification of temperature forecasts. *Wea. Forecasting*, **4,** 485–501.
Palmer, P. L., 1988: The SCS snow survey water supply forecasting program: Current operations and future directions. *Proc. Western Snow Conf.*, Kalispell, MT, Western Snow Conference, 43–51.
Riverside Technology, Inc., 1999: National Weather Service Extended Streamflow Prediction Verification System (ESPVS). Draft Users Manual, U.S. National Weather Service.
Shafer, B. A., and J. M. Huddleston, 1984: Analysis of seasonal volume streamflow forecast errors in the western United States. *Proc. A Critical Assessment of Forecasting in Water Quality Goals in Western Water Resources Management*, Bethesda, MD, American Water Resources Association, 117–126.
Sheppard, P. R., A. C. Comrie, G. D. Packin, K. Angersbach, and M. K. Hughes, 1999: The climate of the Southwest. CLIMAS Rep. Series CL1-99, Institute for the Study of Planet Earth, The University of Arizona, Tucson, AZ, 39 pp.
Wilks, D. S., 1995: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, Academic Press, 233–283.
——, 2000: Diagnostic verification of the Climate Prediction Center long-lead outlooks, 1995–98. *J. Climate*, **13,** 2389–2403.