Hydrol. Earth Syst. Sci. Discuss., 3, 1987–2035, 2006
www.hydrol-earth-syst-sci-discuss.net/3/1987/2006/

Hydrology and Earth System Sciences Discussions

Papers published in *Hydrology and Earth System Sciences Discussions* are under open-access review for the journal *Hydrology and Earth System Sciences*

A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction

D.-J. Seo, H. D. Herr, and J. C. Schaake

Hydrology Laboratory, Office of Hydrologic Development, National Weather Service, National Oceanic and Atmospheric Administration, 1325 East-West Highway, Silver Spring, MD 20910, USA

Received: 9 February 2006 – Accepted: 13 July 2006 – Published: 1 August 2006

Correspondence to: D.-J. Seo (dongjun.seo@noaa.gov)

Abstract

In addition to the uncertainty in future boundary conditions of precipitation and temperature (i.e. the meteorological uncertainty), parametric and structural uncertainties in the hydrologic models and uncertainty in the model initial conditions (i.e. the hydrologic uncertainties) constitute a major source of error in hydrologic prediction. As such, accurate accounting of both meteorological and hydrologic uncertainties is critical to producing reliable probabilistic hydrologic prediction. In this paper, we describe and evaluate a statistical procedure that accounts for hydrologic uncertainty in short-range (1 to 5 days ahead) ensemble streamflow prediction (ESP). Referred to as the ESP post-processor, the procedure operates on ensemble traces of model-predicted streamflow that reflect only the meteorological uncertainty and produces post-processed ensemble traces that reflect both the meteorological and hydrologic uncertainties. A combination of probability matching and regression, the procedure is simple, parsimonious and robust. For a critical evaluation of the procedure, independent validation is carried out for five basins of the Juniata River in Pennsylvania, USA, under a very stringent setting. The results indicate that the post-processor is fully capable of producing ensemble traces that are unbiased in the mean and in the probabilistic sense. Due primarily to the uncertainties in the cumulative probability distributions (CDF) of observed and simulated flows, however, the unbiasedness may be compromised to a varying degree in real world situations. It is also shown, however, that the uncertainties in the CDF's do not significantly diminish the value of post-processed ensemble traces for decision making, and that probabilistic prediction based on post-processed ensemble traces significantly improves the value of single-value prediction at all ranges of flow.

1 Introduction

To account for uncertainties in the future meteorological boundary conditions of Quantitative Precipitation Forecast (QPF) and Quantitative Temperature Forecast (QTF), the

U.S. National Weather Service (NWS) has been using the ensemble streamflow prediction (ESP) methodology for long-range streamflow prediction since the early 1990's (Day, 1985; Smith and Day, 1992; Schaake and Larson, 1998; McEnery et al., 2005). Recently, the practice of ESP has been expanded to short-range prediction in support of the Advanced Hydrologic Prediction Service (AHPS, McEnery et al., 2005). In its original form, ESP accounts for only the meteorological uncertainty and does not consider the parametric (Beven and Binley, 1992; Kuczera and Parent, 1998; Schaake et al., 2001; Vrugt et al., 2003; Duan et al., 2006) and structural (Refsgaard and Knudsen, 1996; Butts et al., 2004) uncertainties in the hydrologic models or the uncertainties in the initial and boundary conditions of the hydrologic models (Sperflage and Georgakakos, 1996; Refsgaard, 1997; Seo et al., 2003; Slater and Clark, 2006). As such, if significant hydrologic uncertainties exist, the ESP products are likely to be biased in the probabilistic sense.

While the hydrologic uncertainties have long been recognized as major sources of error in hydrologic prediction (Binley et al., 1991; Beven and Freer, 2001; Carpenter and Georgakakos, 2004; McMichael et al., 2006), it is only recently that rigorous and systematic efforts have begun to address them comprehensively in the ensemble prediction framework (Frantz et al., 2005). The basic components necessary in the ensemble prediction framework to reduce and to account for hydrologic uncertainties include parameter estimation (Duan et al., 1993, 2006; Gupta et al., 1998), characterization, assessment and modeling of parametric uncertainty (Beven and Binley, 1992; Kuczera and Parent, 1998; Schaake et al., 2001; Vrugt et al., 2003; Duan et al., 2006), state updating or data assimilation (DA) (Sperflage and Georgakakos, 1996; Refsgaard, 1997; Seo et al., 2003; Slater and Clark, 2006), ensemble DA (Zupanski, 2005), multimodel ensemble prediction (Georgakakos et al., 2004; Raftery et al., 2005) and post-processing (Perica et al., 1999). While tremendous progress has been made in recent years to produce, mostly in research settings, a number of promising state-of-the-art techniques and methodologies for these basic components, much additional work is needed to test and to evaluate their operational worthiness and to exploit

fully the complementarity between the automatic algorithms and human forecasters. Also, most of the existing operational forecast systems such as the National Weather Service River Forecast System (NWS 2000) currently lack the hardware and software infrastructure necessary to implement many of the basic components of an ensemble prediction system. For these reasons, the transition to full-blown ensemble prediction is expected to occur incrementally over time. It is also expected that, even if all basic components of hydrologic ensemble prediction are put into operation, there will always remain a level of hydrologic uncertainty of mixed origin that is not fully represented due to less than complete source-by-source accounting of uncertainties. As such, it is expected that statistical procedures such as the one described in this work will always be necessary to account for the residual hydrologic uncertainty and hence to render the probabilistic products as reliable and skillful as possible. It must also be pointed out that, if all major sources of hydrologic uncertainties are properly addressed by the basic components in the ensemble prediction system, the residual uncertainty should largely be statistically structureless. As such, the end state envisioned for the post-processor within the ensemble prediction framework is a very simple statistical procedure, ideally nothing more than a white- (but generally heteroskedastic) noise generator, rather than one of increasing complexity and sophistication. The purpose of this work is then to develop a simple, parsimonious and computationally inexpensive statistical procedure, referred to herein as the ESP post-processor, that accounts for hydrologic uncertainties in short-range (1 to 5 days ahead) ESP in order to improve reliability and, to a lesser extent, resolution (Wilks, 1995) of the short-range ESP products. The post-processor is intended for operational use nationally as a basic component in the ensemble prediction system under widely varying conditions of data availability, goodness of the hydrologic models used, and hydroclimatology of the basins. As such, parsimony and robustness are particularly important considerations. The main contributions of this paper to the hydrologic ensemble prediction literature include: a conceptual description of the hydrologic ensemble prediction framework that allows flexible decomposition and accounting of total hydrologic uncertainty for an evolving forecast system such as

NWSRFS, development of a very simple and parsimonious statistical post-processor that is well-suited for the above framework and inexpensive to update in an evolving forecast system, and a critical independent validation of post-processed streamflow ensembles using commonly used verification measures that can be communicated directly to the ensemble and probabilistic prediction community.

Post-processing of streamflow ensembles from hydrologic models has been used for bias correction in seasonal or long-range ESP (Carpenter and Georgakakos, 2001; Wood et al., 2002), but is only beginning to gain attention in short-range ESP. To the best of the authors' knowledge, Perica et al. (1999), a predecessor to the work described herein, represents the first attempt at post-processing short-range streamflow ensembles. Recently, Hashino et al. 2002) evaluated three bias correction methods for long-range ESP for diagnostic assessment, the event-based correction (EBC, Smith et al., 1992), the regression-type, and the quantile mapping (or probability matching, PM) methods. Their results showed (Hashino et al., 2002) that all bias correction methods improve skill scores, mostly by reducing both the conditional and unconditional biases, the EBC-corrected forecasts tend to have the lowest sharpness and discrimination (Bradley et al., 2004) over all flow quantiles, the PM-corrected tends to give the highest sharpness and discrimination, and the regression type methods seem to be in between the above two. The ESP post-processor developed in this work may be considered a combination of the regression-type and PM methods in that it performs probability matching conditional on recently observed streamflow.

As noted above, parsimony and robustness are very important for operational viability of the procedure. As such, we make conscientious effort to keep the regression model as simple as possible. Because the proposed procedure is only one of many pieces in the hydrologic ensemble prediction system under development at NWS, it is necessary to ascertain its important, but limited, role in end-to-end hydrologic ensemble prediction. For that, we present below an overview of the ESP process from the end-to-end uncertainty-accounting perspective and describe the simplification made to treat (for now) all hydrologic uncertainties as an aggregate, in the absence of individual

hydrologic uncertainty processors in the current short-range ESP operations in NWS.

2 Overview of ESP and aggregation of hydrologic uncertainty

The complete ESP process that accounts for all major sources of uncertainty (see Fig. 1) amounts to numerically evaluating the probability distribution of the forecast flow at some future times given the observed flow up to and including the current time via the total probability law (Krzysztofowicz, 1999):

$$f_1(q_f|q_o) = \int f_2(q_f|q_o, s_f) f_3(s_f|q_o) ds_f \tag{1}$$

where f_i's denote probability density functions (PDF), and q_f, q_o and s_f denote the predictand, i.e. the streamflow at some future times, the observed flow up to and including the current time, and the model-predicted streamflow at the future times, respectively. In writing PDF's, we denote only the experimental values of the random variables for notational brevity in this paper. For example, $f_1(q_f|q_o)$ in Eq. (1) abbreviates $f_{1_{Q_f|Q_o}}(q_f|q_o)$, where Q_f and Q_o denote the random variables, and q_f and q_o denote the respective experimental values of the random variables. Throughout this paper, we denote the experimental values of the random variables and the random variables themselves by lower- and uppercase letters, respectively. The PDF's in the integrand in Eq. (1) may be rewritten as:

$$f_3(s_f|q_o) = \iiint f_4(s_f|b_f, i, p, q_o) f_5(b_f|i, p, q_o) f_6(p|i, q_o) f_7(i|q_o) db_f di dp \tag{2}$$

$$f_2(q_f|q_o, s_f) = \left| \frac{\partial \varepsilon(q_f, q_o, s_f)}{\partial q_f} \right| f_8(\varepsilon(q_f, q_o, s_f)|q_o, s_f) \tag{3}$$

where b_f denotes the boundary conditions of precipitation and temperature at the future times, i denotes the model initial conditions, p denotes the model parameters,

and $\varepsilon(\cdot)$ denotes the error in predicting q_f given q_o and s_f, w, expressed as a function of q_f, q_o and s_f, i.e., $w=\varepsilon(q_f, q_o, s_f)$. The error function $\varepsilon(\cdot)$ is obtained by inverting the statistical prediction model for q_f, $\eta(\cdot)$, expressed as a function of q_o, s_f and w, i.e., $q_f=\eta(q_o, s_f, w)$ (see Sect. 3 for an example in the normal space). The conditional PDF's $f_5(b_f|i, p, q_o)$, $f_6(p|i, q_o)$ and $f_7(i|q_o)$ in Eq. (2) correspond in ESP to the meteorological input pre-processor, the parametric uncertainty processor, and the initial condition uncertainty processor (see Fig. 1) that generate ensemble traces of B_f, P and I, respectively. In Eq. (2), because B_f is independent of I, P and Q_o, $f_5(b_f|i, p, q_o)$ may be reduced to $f_5(b_f)$ and, because model parameters are independent of I and Q_o, $f_6(p|i, q_o)$ may be reduced to $f_6(p)$. Also, if there is no updating of model states based on observed streamflow, $f_7(i|q_o)$ and $f_4(s_f|b_f, i, p, q_o)$ in Eq. (2) may be reduced to $f_7(i)$ and $f_4(s_f|b_f, i, p)$, respectively. The ensemble traces of B_f, P and I thus obtained may then be input to the hydrologic models, represented by $f_4(s_f|b_f, i, p, q_o)$ in Eq. (2), to generate ensemble traces of S_f that reflect the meteorological uncertainty, the parametric uncertainty and the uncertainty in the model initial conditions. The conditional PDF $f_8(w|q_o, s_f)$ in Eq. (3) represents the hydrologic error model and generates ensemble traces of W that reflect the uncertainty in predicting q_f given q_o and s_f. Finally, the statistical prediction model, $q_f=\eta(q_o, s_f, w)$, inputs the ensemble traces of W and generates ensemble traces of the predictand q_f, which now reflect both the meteorological and hydrologic uncertainties.

The motivation for decomposing total hydrologic uncertainty into constituent uncertainties as described above is at least three-fold. The first is to utilize the existing and emerging DA, including state updating, capabilities (Sperflage and Georgakakos, 1996; Refsgaard, 1997; Seo et al., 2003; Zupanski, 2005; Slater and Clark, 2006) so that hydrologic uncertainty in initial conditions may be reduced beyond what purely statistical techniques may bear. The second is to assess relative importance among the major sources of hydrologic error contributing to the total hydrologic uncertainty. Such assessment is necessary to prioritize improvements in different components of the forecast system. The third is to whiten the residual hydrologic uncertainty W as

much as possible so that statistical modeling of the post processor may be made as simple and parsimonious, and hence as less data-intensive, as possible. The last consideration is of particular importance under climate change conditions. In operational implementation of the above framework, it is likely that the number of ensemble members for initial conditions or model parameters may have to be kept relatively small to keep computational burden in check, and that the uncertainties not explicitly captured by the above ensembles may have be to accounted for in the residual uncertainty W by the post processor. Such a practice, however, is not a departure from the above framework, but a practical compromise given available human and computational resources, data availability and understanding of the uncertainties involved. It is also possible that, for now, explicit accounting of parametric uncertainty may be computationally too expensive to be practical for distributed models (see e.g. Liu et al., 2005). In such a case, one may deal only with initial condition uncertainty explicitly and account for parametric uncertainty in the residual hydrologic uncertainty W via a statistical post processor within the above framework.

Development of the hydrologic uncertainty processors described above requires probabilistic description of the errors from individual sources, P, I and W. In practice, such individual modeling of the errors is very difficult because, very often, very little is known about them a priori (Kuczera and Parent, 1998; Butts et al., 2004; Carpenter and Georgakakos, 2004; Pappenberger et al., 2005; McMichael et al., 2006). In the absence of individual hydrologic uncertainty processors in the current short-range ESP operations in NWS, in this work we aggregate the parametric uncertainty and the uncertainty in the initial conditions with W, and model only a single hydrologic uncertainty processor (see also Krzysztofowicz, 1999; Krzysztofowicz and Maranzano, 2004). Such "error aggregation" amounts to placing all sources of hydrologic uncertainty in W by assuming perfectly known P and I, and approximating the added uncertainties due to P and I in W. As noted in Sect. 1, it is expected that the above approximation will ease over time following incremental implementation of the basic components in the ensemble prediction system (see Fig. 1). Once W is modeled, the

post-processor generates ensemble traces of forecast streamflow, Q_f, from the model prediction, s_f, the observed flow, q_o, and the aggregate hydrologic error, w, via the statistical prediction model, $q_f = \eta(q_o, s_f, w)$, which we describe in Sect. 3 in some detail. Because the scope of this work is limited to dealing with only the hydrologic uncertainty, we assume in the development below that no meteorological uncertainties exist; i.e., the future boundary conditions of precipitation and temperature are known clairvoyantly (i.e. $f_5(b_f|i, p, q_o)$ in Eq. (2) is an impulse function). It is therefore to be understood that we make no distinctions between model simulation and model prediction unless stated otherwise in this paper.

The above simplified approach of error aggregation in modeling of total hydrologic uncertainty is adopted earlier in Perica et al. (1999), Krzysztofowicz (1999) and Krzysztofowicz and Maranzano, 2004) also. Beyond the notable difference that Krzysztofowicz and Maranzano (2004) is quasi-analytical and models river stage whereas the post-processor is numerical and models river flow, an important distinction between the two is that the former is Bayesian, and hence models the prior and the likelihood function to obtain the posterior via Bayes'rule, whereas latter is not and amounts to modeling the posterior directly. Theoretically speaking, the Bayesian approach is expected to do better than, or at least as well as, the post processor's approach. Due to the added complexity of deriving the posterior through Bayes' rule, however, the Bayesian approach heavily parameterizes the prior and the likelihood function (Krzysztofowicz and Maranzano, 2004), the parameters of which are then assumed, implicitly, perfectly known. The post processor, on the other hand, is extremely parsimonious as shown below and hence may be minimally subject to parametric uncertainty (i.e. not of the hydrologic model but of the statistical model). In this regard, comparative evaluation between the two approaches via objective independent validation is critically needed in order to develop a better sense of balance among complexity of the statistical model, parsimony of the model parameters, and data requirements necessary for cost-effective implementation of hydrologic uncertainty processors in the hydrologic ensemble forecast system depicted in Fig. 1. Such a comparative evaluation, however, is a major undertaking and well beyond the scope of this work.

3 Statistical prediction model

The statistical prediction model, $q_f = \eta(q_o, s_f, w)$, is based on the following recursive linear regression in the normal space:

$$Z^o_{k+1} = (1 - b)Z^o_k + bZ^s_{k+1} + \mathrm{E}_{k+1} \quad (4)$$

where Z^o_k and Z^o_{k+1} denote the normalized observed flows at time steps k and k+1, respectively, Z^s_{k+1} denotes the normalized model-predicted flow at time step k+1, E_{k+1} denotes the random error representing the aggregate hydrologic uncertainty at time step k+1 in the normal space, and b denotes the weight given to the normalized model prediction ($0 \le b \le 1$). The choice of the above autoregressive-1 model with a single exogenous variable, or ARX(1,1) (Box and Jenkins, 1976), is based on the previous work (Perica et al., 1999) and evaluation of different candidate models (mostly of the ARX class), via parameter estimation and independent validation similar to those described in Sect. 4. The choice of ARX(1,1), which results from the instantaneous stochastic dependence (Maranzano and Krzysztofowicz, 2006) and the a posteriori Markov of order one dependence (Maranzano and Krzysztofowicz, 2006) between the modeled and observed streamflow processes, is also supported by the Bayesian approach (Maranzano and Krzysztofowicz, 2006) based on stochastic modeling of the river stage processes. We also note that the ARX(1,1) model has been in use for years as the noise model in real-time flood forecasting applications, particularly in the form of a Kalman filter (Todini, 2005). There are, however, additional considerations that went into the choice as explained below. It is expected that DA for state updating (see e.g. Seo et al., 2003) will greatly reduce the need for autoregressive statistical modeling, and that the parametric uncertainty processor will further reduce serial correlation of forecast error at both small and large time lags. Given these prospects, it is very much desirable to develop and implement (assuming of course that the performance is satisfactory) as

simple and parsimonious as possible a statistical post processor so that its parameters may be updated very easily as the forecast system evolves.

With Eq. (4), the ensemble traces of the predictand Q_f may be generated recursively from $q_f = \eta(q_o, s_f, w)$ via the following sequence of operations: 1) normal quantile-transform q_k^o and q_{k+1}^s to obtain z_k^o and z_{k+1}^s, respectively, where q_k^o and q_{k+1}^s denote the observed and model-predicted flows at time steps k and k+1, respectively, 2) sample a normal random deviate of E_{k+1} from $N(m_E, \sigma_E^2)$, where m_E and σ_E^2 denote the mean and variance of E_{k+1}, respectively, 3) pass the normal random deviate through Eq. (4) to obtain a realization of z_{k+1}^o, 4) normal quantile inverse-transform z_{k+1}^o to obtain a realization of 1 day-ahead prediction of q_{k+1}^o, 5) treating the 1 day-ahead prediction of q_{k+1}^o as observed, repeat Steps 1 through 5 out to 5 days into the future, 5) repeat Steps 2 through 5 as many times as the number of ensemble traces desired. The above process amounts to conditional simulation (Deutsch and Journel, 1992; Seo et al., 2000) of Z^o, making use of the fact that realizations of Z^o are simulated outcomes of observed streamflow. The normal quantile transform (NQT) in Step 1 maps flow to standard normal deviate by matching the empirical cumulative probability of the flow with the standard normal cumulative probability:

$$z_k^o = \mathrm{nqt}_o(q_k^o) \tag{5}$$

$$z_k^s = \mathrm{nqt}_s(q_k^s) \tag{6}$$

where $\mathrm{nqt}_o(\cdot)$ and $\mathrm{nqt}_s(\cdot)$ denote the empirical NQT functions for observed and model-simulated flows, respectively. The procedure thus combines PM (or quantile mapping in Hashino et al., 2002) with linear regression. If b=1 and E_{k+1}=0 (i.e. the hydrologic model prediction is perfect in the normal space), the procedure is reduced to PM. If b=0 (i.e. the hydrologic model prediction has no skill), the procedure is reduced to the (purely statistical) autoregressive-1 model (AR(1)) (Bras and Rodriguez-Iturbe, 1985) in the normal space with variable transformation based on NQT.

Because we are using empirical, rather than theoretically fitted, cumulative probability distribution functions (CDF), the normal random deviates from Step 3 may not be

inverse-transformed if they lie outside of the historically observed range. To extend the empirical CDF's (ECDF) beyond the historical maxima, we use the following hyperbolic approximation for the uppermost-tail of the distribution (Deutsch and Journel, 1992):

$$q = \left[\frac{\lambda}{1 - \Pr[Q < q]}\right]^{\frac{1}{\omega}} \tag{7}$$

In the above, Pr[·] denotes the probability of occurrence of the event bracketed, $\Pr[Q<q]$ is given by $\Pr[Z<z]$ where z denotes the random deviate from Eq. (4) of the standard normal variable Z exceeding the historical maximum, λ is given by $\lambda = q_{\max}^{\omega}(1 - \Pr[Q<q_{\max}])$ where $q_{\max}$ denotes the historical maximum flow (m^3/s), and ω is a parameter that controls the fatness of the tail (the smaller ω, the fatter the tail). Sensitivity analysis indicates that the performance of the post-processor is moderately sensitive to ω, and that ω=3.25 is a good choice for all basins studied in this work (see Table 1).

By rewriting Eq. (4) in terms of E_{k+1} and minimizing its variance with respect to b, we obtain the following least-squares solution for b in the normal space:

$$b = \{1 - \rho_o(|1|) - \rho_c(|1|) + \rho_c(|0|)\}/\{2(1 - \rho_c(|1|))\} \tag{8}$$

In the above, $\rho_o(|l|)$ and $\rho_c(|l|)$ denote the serial correlation of Z^o and the cross correlation between Z^o and Z^s at lag l, respectively. Hence, under the least squares criterion in the normal space, the coefficient b is expressed as a function of the autocorrelation of Z^o and the cross-correlation between Z^o and Z^s. Thus, the effects of auto- and cross-correlations in the probabilistic sense in the observed and modeled streamflow processes are captured in the post processor by the coefficient b through the above relationship. The least-squares solution of Eq. (8), however, is specific only to the one-step transition shown in Eq. (4). For the coefficient b to be lead time-invariant so that Eq. (4) is a theoretically valid recursive model, additional assumptions are necessary on the correlation structures of Z^o and Z^s (the details are not central to the development of this paper and are not given here). While it may make an interesting exercise to

hypothesis-test such assumptions, statistical acceptance of the hypotheses in the normal space may not necessarily mean that the performance of the post processor would be satisfactory. For example, even if all statistics in Eq. (8) are perfectly estimated, optimality (in the minimum error variance sense) in the normal space of the estimate of b as obtained from Eq. (8) does not necessarily imply satisfactory performance of the post processor in the original space. Our experience strongly indicates that an optimal (in the minimum error variance sense) estimate of b in the normal space produces good performance in the original space in the unconditional mean sense, but rather poor performance in the conditional mean sense, particularly for all-important large streamflows (this tendency, however, is expected to be less pronounced for river stage, owing to smaller skewness in distribution than streamflow). As such, the approach taken in this work is to perform nonlinear optimization of the parameter b numerically in the original space under a multi-objective criterion as described in Sect. 4. In this way, one may ensure satisfactory performance of the post processor not only in the unconditional mean sense but also in the conditional mean sense.

Because the correlation structure depends greatly on the magnitude of the streamflow, so does the coefficient b (in the normal as well as in the original space). As such, we stratify the parameter b (and hence m_E and σ_E^2) according to the magnitude of the model-simulated flow as follows:

$$b = \begin{cases} b_{\text{low}} & \text{if } q_{k+1}^s \geq q_c \\ b_{\text{high}} & \text{otherwise} \end{cases} \tag{9}$$

where the cutoff flow q_c (m^3/s) is chosen to be the median of the historically observed flow based on sensitivity analysis. It is noted here that we tried other, seemingly more sophisticated, models for the dependence of b on q_{k+1}^s, but could not realize noticeable, if at all, improvement in independent validation. The above conditioning of b on the flow regime is a practical but empirical attempt to reflect the observation that disparate stochastic processes are at work in low- and high-flow situations in modeling of hydrologic uncertainty. Further research is needed for flow regime-dependent modeling of residual hydrologic error that draws from understanding of the underlying

physical processes. To account for the seasonality of streamflow (see Fig. 2), the statistical parameters for W and the NQT functions in Eqs. (5) and (6) should preferably be stratified at the monthly or seasonal scale. Sensitivity analysis, however, indicates that such stratification does not yield reliable CDF's from a 20-yr record of daily streamflow data. As such, we choose a semi-annual stratification to increase the sample size and estimate the parameters for wet (December through May) and dry (June through November) seasons. Additional research is needed to identify and account for possible nonstationarities due to factors other than seasonality, such as changes in the rating curve and climate change, from analysis of residuals.

In place of Eq. (4), one may consider other formulations for the statistical prediction model, the bivariate AR(1) model (Bras and Rodriguez-Iturbe, 1985; Seo et al., 2000) being an obvious candidate. Our experience, however, indicates that the use of a multivariate time series model with NQT is rather tricky because, whereas the model parameters estimated in the normal space may be optimal with respect to the normal quantile-transformed flow (i.e. in the minimum error variance sense), they are generally very poor estimates with respect to large flows in the original space.

4 Parameter estimation and validation

The post-processor described above has only one free parameter, the coefficient b in Eq. (4), as stratified according to Eq. (9). To estimate, we minimize the following objective function:

$$\text{Minimize } J = \alpha \frac{1}{n} \sum_{i=1}^{n} [q_k^o - q_k^p]^2 + \frac{1}{n} \sum_{i=1}^{n} [(q_k^o)^{\text{ord}} - (q_k^p)^{\text{ord}}]^2 \tag{10}$$

where q_k^o denotes the observed flow at the k-th time step, q_k^p denotes the conditional mean of the observed flow at the k-th time step given the observed and model-simulated flows at the (k-1)st and k-th time steps, respectively,

i.e. $q_k^p = E[Q_k^o | Q_{k-1}^o = q_{k-1}^o, Q_k^s = q_k^s]$ (see Eq. 10), n denotes the total number of data points, α denotes the weight given to the first term, and $(\cdot)^{\text{ord}}$ denotes that the variable superscripted is sorted in the ascending order of magnitude. The first term in Eq. (10) is the mean square error of the ensemble mean prediction from the post-processor. The second term in Eq. (10) is the mean square deviation from the 45° line in the quantile-quantile plot between the observed and the ensemble mean-predicted flows. The purpose of the second term is to render the marginal PDF of the predictand close to that of the observed (cf PM). With the second term in the objective function, the role of the first term is primarily that of maximizing the cross correlation between the predictand and the observed flow. Eq. (10) thus forces the bivariate distribution of the post-processed ensemble mean and the observed flows to tighten symmetrically along the 45° line. In this work, the weight α in Eq. (9) is specified empirically based on sensitivity analysis. Our experience suggests that $\alpha \approx 0.1$ is a reasonable choice for all cases studied in this work. Because α balances the two, often competing, objectives of minimizing unconditional error variance and reducing conditional bias, some trial and error may usually be necessary. To that end, further research is needed to determine the weight objectively and to provide the user with objective guidance for its specification. The predictand, q_k^p, in Eq. (10) is evaluated by:

$$E[Q_k^o | Q_{k-1}^o = q_{k-1}^o, Q_k^s = q_k^s] = \int_{-\infty}^{\infty} \text{nqt}_o^{-1}(z_k^o) f(z_k^o | z_{k-1}^o, z_k^s) dz_k^o \quad (11)$$

where $\text{nqt}_o^{-1}(\cdot)$ denotes the normal quantile inverse-transformation of observed flow, and the conditional PDF of Z_k^o, $f(\cdot|\cdot)$, is given by $N((1-b)\cdot z_{k-1}^o + b\cdot z_k^s + m_E, \sigma_E^2)$. Because we are using empirical functions for $\text{nqt}_o(\cdot)$, closed-form evaluation of the theoretical ensemble statistics, such as the right-hand side of Eq. (10), is not possible. In this work, we evaluate them quasi-analytically via piecewise approximation of $\text{nqt}_o^{-1}(\cdot)$. Being able to evaluate the theoretical ensemble statistics a priori is an extremely important feature of the post-processor developed in this work as it allows optimal, in the ensemble mean sense, estimation of the parameter b (and hence m_E and σ_E^2), without having

to generate a large number of ensemble traces to obtain reliable sample statistics. To test empirically the lead time-invariance of b put forth in Sect. 3, we also carried out nonlinear optimization of b for 2- to 5-step transitions based on the recursion in Eq. (4) and validated the post processor results against observed streamflow both dependently and independently using 5 different estimates of b as optimized using 1- to 5-step transitions via recursion. The results indicate that the dependence of b on lead time is rather small, and that the difference in the performance of the post processor is only minor among the different estimates of b. In the following, all results presented are based on estimates of b from 1 step-ahead transition, regardless of lead time.

The basins used in this work for evaluation of the post-processor are SPKP1, HUNP1, SAXP1, MPLP1 and NPTP1 of the Juniata River in the NWS Middle Atlantic River Forecast Center's (MARFC) service area (see Fig. 3 and Table 1). They are chosen from all of the Juniata River basins mainly for the availability of lengthy (47 years, covering 1951 through 1997) observations of daily streamflow so that independent validation and sensitivity analysis may be carried out under various scenarios. The hydrologic models used are the Antecedent Precipitation Index (API, Anderson, 1993) and SNOW-17 (Anderson 1973) models for the rainfall-runoff and snowmelt processes, respectively. The hydrologic models were calibrated at MARFC using 11 or 13 years of data (David Zanzalari, personal communication). The model calibration is generally of very high quality, particularly for high flows, with the Nash-Sutcliffe and cross-correlation coefficients ranging from 0.79 to 0.87 and from 0.89 to 0.94, respectively, for daily simulation in the validation periods (see also Fig. 6). The flow at RTDP1 is regulated by the Raystown Dam, which also influences the basins downstream. This regulation, which is not modeled in this work, presents a significant source of additional uncertainty in model simulations at MPLP1 and NPTP1 (see Fig. 3), and hence offers a particularly challenging test for the post-processor.

To evaluate the post-processor, we carried out independent validation as follows: 1) divide the 47-yr record of the observed and simulated daily streamflow into two, 2)

estimate the post-processor parameters (including the NQT functions in Eqs. 5 and 6) using the first half of the data, 3) run the post-processor through the second half at lead times of 1 through 5 days, and 4) repeat the above steps with the two halves of the data interchanged. In an operational setting, estimation of the post-processor parameters should be a part or an extension of hydrologic model calibration. Given that 11- to 13-yr data were used to calibrate the hydrologic models for these basins, it is of particular interest to estimate the post-processor parameters using data sets of comparable length (e.g., by dividing the 47-yr data into four periods). It is noted here that we carried out such an experiment as a part of a larger sensitivity study. The results indicate that the ECDF's (and subsequently the NQT functions) obtained from 11-yr records are subject to rather large sampling uncertainties, which significantly compromises the post-processor's performance (these points will be made clearer in the next section). As such, we focus on parameter estimation and validation based on 23-yr records, the results of which are summarized below.

5 Results

5.1 Deterministic results

In the deterministic sense, the primary purpose of the post-processor is to remove or reduce systematic biases in the model-predicted flow. As such, a viable post-processor should produce a prediction that is in general better than or at least as good as the model prediction in the deterministic sense. To verify this, we evaluate the quality of the ensemble mean prediction from the post-processor as obtained from Eq. (10). Figures 4 and 5 show the ratio of the sum of the observed flow to that of the post-processor (PP)-predicted flow (denoted as RATIO) and the percent reduction in the root mean square error (RMSE) by the PP-predicted flow over the model-predicted, respectively. All results shown in Figs. 4 and 5 are from the parameter estimation periods (see Sect. 4). Hence, they represent the upper bound of the post-processor's

performance under the assumption that the parameters are known very accurately. The closer the RATIO is to unity in Fig. 4 and the larger the reduction in RMSE is in Fig. 5, the better the performance is. A negative reduction in RMSE, however, does not necessarily mean that the performance of the post-processor is poor because certain model biases may only be corrected at the expense of an increased RMSE. Note in the figures that the post-processor renders RATIO close to unity at all lead times and significantly reduces RMSE. That the post-processor produces RATIO's that are consistently, albeit slightly, less than unity (see Fig. 4) is an indication that it has a small bias of its own. The source of this bias is not clear, and further investigation is needed for its identification and correction. That the reduction in RMSE levels off rather quickly (see Fig. 5) indicates that the inclusion of observed flow in the post-processor adds significant skill (i.e. beyond that owing to bias removal) to 1 day-ahead prediction only. This is in agreement with state updating or DA results for basins of comparable size (see e.g. Refsgaard, 1997; Seo et al., 2003) and indicates that, for fast-responding runoff processes in these basins, the memory of the initial conditions lasts a day or so. Similarly, examination of the cross correlation between the observed and the model-predicted flows and between the observed and the PP-predicted flows (not shown) indicates that the post-processor may provide significant additional predictive skill for the first day.

Figure 6 shows an example of the scatter and quantile-quantile plots of daily flow between the observed and the model-predicted flows (upper panels) and between the observed and the PP-predicted flows (lower panels). All results shown therein are for 1 day-ahead prediction from a parameter estimation period. The scatter plots are shown in both linear (right panels) and log (middle panels) scales so that the performance in the low-flow ranges may also be scrutinized. A desirable post-processor would produce ensemble mean prediction that is conditionally unbiased at all ranges of flow and has the same marginal probability distribution as the observed flow. Such a processor would place the scatter and quantile-quantile plots on the 45° line in the lower panels of Fig. 6. Though difficult to see in the linear plot (upper-left panel), it is evident in

the log-log plot (upper-middle panel) that the model prediction is significantly biased in the low-flow ranges. Note that the post-processor removes this conditional bias all but completely and tightens the scatter considerably (lower-left and -middle panels).

Figures 7 and 8 are the same as Figs. 4 and 5, respectively, but for the two validation periods (see Sect. 4). It is important to note that all validation results presented in this paper represent the post-processor's performance under very stringent conditions where, once the parameters are estimated from a daily streamflow record of about 20 years, the post-processor is run continuously for about 20 years thereafter, without ever updating the parameters. With parameter updating in the real world, it is likely that the performance would be significantly better than seen in these two figures. Compared to the results from the parameter estimation periods (Figs. 4 and 5), significant deterioration in the post-processor's performance is seen in the validation periods. Figure 7 shows that, though the RATIO of the PP-predicted flow does spread around the line of unity, the volume bias in the PP-predicted flow can exceed 10 percent in the worst case (HUNP1). Figure 8 indicates that, while the pattern of reduction in RMSE in the validation periods is similar to that in the parameter estimation (see Fig. 5), the magnitude of reduction is significantly smaller (due in large part to the increased bias). In Fig. 7, by far the largest departure in RATIO from the line of unity is at HUNP1. Given that the deterioration of the performance in the validation periods must come from the uncertainty in the post-processor parameters, it is suspected that the ECDF's used in NQT (see Eqs. 5 and 6) are the primary source of that uncertainty. Figure 9 shows the ECDF's of observed and model-simulated flows for the wet season (December through May) at HUNP1. In the figure, there are two ECDF's each for observed and simulated flows corresponding to the two parameter estimation periods (see Sect. 4). Also shown in the lower-right corner of the figure is a magnification of the upper-tail portion of the ECDF's. If the two ECDF's differ significantly, it is likely that the normal quantile back-transformation in Eq. (10) results in significant errors. Examination of the ECDF's of observed and simulated flows for all basins (not shown) indicates that HUNP1, which has the largest error in RATIO of the PP-predicted flow (see Fig. 7), has

the largest difference in the ECDF's between the two parameter estimation periods (see Fig. 9). SAXP1, on the other hand, has the smallest difference (not shown), and yields the smallest error in RATIO of the PP-predicted flow (see Fig. 7). The above observation suggests that the magnitude of long-term variability in the ECDF's is a very good indicator of the parametric uncertainty in the post-processor, and hence of the post-processor's performance.

5.2 Probabilistic results

In the probabilistic sense, the primary purpose of the post-processor is to render the predicted probability unbiased (i.e. assuming that the meteorological input is unbiased). As such, by far the most important performance measure is reliability, which measures unbiasedness of the predicted probability against the observed frequency of the prediction (Wilks, 1995; Jolliffe and Stephenson, 2003). Figures 10 through 12 show the reliability diagrams (Wilks, 1995) for NPTP1 at the thresholds of 50th-, 85th- and 97.5th-percentile flows, or 19, 53 and 132 (m^3/s), respectively, in the parameter estimation periods. All reliability diagrams shown in this paper are based on 99 ensemble traces. In each figure, the thick solid and dashed lines represent the reliability diagrams of 1 and 5 day-ahead probabilistic predictions, respectively. Note that there are two lines each for each forecast lead time corresponding to the two parameter estimation periods (see Sect. 4). Each point in the reliability diagram represents the frequency of observed flow exceeding the threshold (i.e. the observed frequency on the y-axis) given the probability that observed flow will exceed the threshold as predicted by the post-processed ensemble traces (i.e. the predicted probability on the x-axis). A perfectly reliable probabilistic prediction would place the points exactly on the 45° line, shown as the thin solid line along the diagonal. In each figure, the thin dotted horizontal and vertical lines, referred to as the "no-resolution" lines (Wilks, 1995), correspond to the climatological probability that the observed flow exceeds the threshold. Resolution measures the sensitivity of the observed frequency of the prediction to the predicted probability, or that of the predicted probability to the observed frequency of the predic-

tion. Note that, on the horizontal line of no resolution, different predicted probabilities lead to the same observed probability and, on the vertical line of no resolution, the same predicted probabilities lead to different observed probabilities. The thin dotted line between the 45° and the horizontal no-resolution line represents the "no-skill" line. Points on this line do not contribute to the Brier score whereas those below it contribute negatively to the score (Wilks, 1995). At high thresholds, reliability diagrams are subject to significant sampling errors due to small sample size. Such diagrams are characterized by jagged or saw tooth-like lines (cf. Fig. 12) and must be interpreted with caution.

Figures 10 through 12 indicate that, if the parameters are known accurately, the post-processed ensemble traces are extremely reliable at all thresholds and at all lead times. Also shown in the lower-right corner of the reliability diagram are the histograms of the predicted probability, where the solid and dotted lines denote the day-1 and -5 predictions, respectively. A "U"-shaped histogram indicates high sharpness in the probabilistic prediction (Wilks, 1995). That is, the prediction is able to "stick its neck out" and differ from climatology (Jolliffe and Stephenson 2002). A flat histogram, on the other hand, indicates little sharpness in that the prediction differs little from climatology. Note in Fig. 10 that the histogram of the 1 day-ahead PP-prediction at the median flow is characterized by a deeper "U" than that of the 5-day ahead prediction, an indication that the 1 day-head prediction is significantly sharper at this threshold. The histograms at higher thresholds (see Figs. 11 and 12), on the other hand, do not differ very much between 1 and 5 day-ahead predictions. The above observations indicate that the improvement in skill of the PP-prediction comes largely from low-flow ranges, in agreement with the observations made from the scatter-plots in Fig. 6.

Figures 13 through 15 are the same as Figs. 10 through 12 but for the two validation periods. Note that, in general, the reliability diagrams depart significantly from the 45° line. At the thresholds of 50th- 85th- and 97.5th-percentile flows, the reliability diagrams have maximum errors (in probability) of over 0.3, under 0.1 and over 0.2, respectively. These error bounds represent the largest among the five basins and hence represent

the worst case results. The reliability diagrams for other basins (not shown) are qualitatively similar, and the magnitude of departure varies significantly from basin to basin (smaller for SPKP1, SAXP1 and MPLP1, and approximately the same for HUNP1). As with the departure of RATIO from the line of unity in Fig. 7, the departure of the reliability diagrams from the 45° line is attributable primarily to the uncertainty in the ECDF's of observed and simulated flows. If the ECDF's from the two parameter estimation periods do not differ very much, the departure is significantly smaller. As seen in Figs. 13 through 15, the magnitude of the error in the reliability diagram depends also on the threshold, which is attributable again to the differences in the ECDF's between the two periods. Note in Fig. 9 that, at the exceedance probability levels of 0.5 and 0.975, the two ECDF's of observed flow differ significantly, resulting in large departures in the reliability diagrams at these thresholds. At the exceedance probability level of 0.85, on the other hand, the ECDF's are close to each other, resulting in only a small departure. Thus, as in the deterministic case (see Subsect. 5.1), the magnitude of long-term variability in the ECDF's of observed flow at various ranges is seen as a good indicator of the accuracy of the PP-predicted probability at the corresponding thresholds. Figures 13 through 15 indicate that, even with a 20-yr record (but with no parameter updating), the ECDF of observed flow is subject to significant uncertainties due to natural variability of streamflow, and that these uncertainties may result in significant errors in the PP-predicted probability.

5.3 Value of probabilistic prediction

While the magnitude of the errors in the predicted probability may be assessed through the reliability diagrams as described above, it is difficult to gauge the impact of such errors on the value of probabilistic prediction for decision making. For that purpose, we examine below the Relative Operating Characteristic (ROC, Jolliffe and Stephenson, 2003) of the post-processed ensemble traces. The ROC plots the hit rate (HT) against the false alarm rate (FAR) at various levels of predicted probability. The ROC curve of a perfect probabilistic prediction connects the points, (0,0), (0,1) and (1,1),

on the (HT,FAR) plane, and that of a worthless prediction connects (0,0) and (1,1). It is important to point out that the ROC measures only of discrimination ability of the set of forecasts, i.e. the degree to which the forecasts differ for a specific observation (Bradley et al., 2004), and does not rely on the forecasts being well calibrated (i.e., reliable) (Glahn, 2004). As such, the ROC curve measures performance only in discrimination, and not in reliability, in the post-processed ensemble traces.

Figures 16 through 18 show the ROC curves for the probabilistic prediction based on the post-processed ensemble traces at HUNP1 at the thresholds of 50th-, 85th- and 97.5th-percentile flows, respectively. The results for other basins are qualitatively similar and are not shown. In each figure, the thick and thin solid lines denote the ROC curves of the post-processed ensemble traces in the validation and parameter estimation periods, respectively. The area enclosed by the ROC curve and the diagonal line is referred to as the ROC area, which is considered a proxy for the economic value of the prediction (Zhu et al., 2002). Also shown in the figure are the (HT,FAR) positions of the deterministic, or single-value, model prediction (denoted by squares) and the ensemble mean prediction from the post-processor (denoted by triangles) in the validation (denoted by solid markers) and in the parameter estimation (denoted by empty markers) periods. For deterministic prediction, the ROC area is defined by the triangle connecting (0,0), (1,1) and the (HT,FAR) positions of the prediction. In calculating the hit and false alarm rates in these figures, the hit, miss, false-alarm and correct-rejection counts were conditioned on the event that the daily observed flow exceeds the two-thirds of the threshold flow. In other words, we excluded the paired data points in the observed and predicted flow time series if the observed flow is below the two-thirds of the threshold flow. This conditioning pulls the ROC curves toward the diagonal line, and hence amplifies the differences in the ROC's among different predictions. Figures 16 through 18 may be summarized as follows.

Though the uncertainties in the ECDF's may introduce significant errors in the PP-predicted probabilities at the low and high thresholds of 50th- and 97.5th-percentile flows, their impact on the discrimination ability and value of probabilistic prediction for

decision making, as measured by the ROC, is only modest at all thresholds. Compared to the deterministic predictions, the probabilistic prediction based on the post-processed ensemble traces significantly increases the value of prediction at all thresholds. In the deterministic sense, the ensemble mean prediction from the post-processor is significantly better than the (raw) model prediction at the low threshold of median flow. At the medium and high thresholds of 85th- and 97.5th-percentile flows, the ensemble mean prediction is comparable to the model prediction. Lastly, it is worth noting that the ROC at the high threshold of 97.5th-percentile flow (Fig. 18) is very similar, both qualitatively and quantitatively, to the ROC of the multi-model ensemble at the flood-flow threshold (see Fig. 15 of Georgakakos et al., 2003) from the Distributed Model Intercomparison Project (DMIP, Smith et al., 2004). The similarity between the two hints that post-processed ensemble traces from a single "high-quality" model may be of comparable value for decision making to a multi-model ensemble of varying quality.

6 Summary, conclusions and future research recommendations

A statistical post-processing procedure has been developed to account for hydrologic uncertainty in short-range (1- to 5-days ahead) ensemble streamflow prediction (ESP). The procedure, referred to as the ESP post-processor, combines probability matching and recursive linear regression to generate ensemble traces of daily streamflow that reflect both the meteorological and hydrologic uncertainties from those that reflect only the meteorological uncertainty. The hydrologic uncertainties considered are parametric and structural uncertainties in the hydrologic models and the uncertainty in the model initial conditions. The procedure has only one free parameter that needs to be estimated via optimization. All other parameters, including the empirical cumulative distribution functions (ECDF) of observed and simulated flows, are estimated directly from the data. For a critical evaluation of the post-processor, independent validation was carried out under very stringent conditions for five basins of the Juniata River in Pennsylvania, USA.

The results indicate that the post-processor is fully capable of producing ensemble traces that are unbiased in the mean and in the probabilistic sense. Due primarily to the uncertainties in the cumulative probability distributions functions (CDF) of observed and simulated flows, however, the unbiasedness may be compromised to a varying degree in real world situations. It is shown that, with the parameters estimated from a 20-yr record, the post-processor generally improves the accuracy of the model prediction in the mean sense, the improvement being most significant for low flows. The worst case results based on the parameters estimated from a 20-yr record indicate that the ensemble mean prediction from the post-processor may be in error by up to 10 percent, and that the post-processor-predicted probabilities may be in error by up to 30, 10 and 20 percent for median, 85th- and 97.5th-percentile flows, respectively. It is also shown, however, that the uncertainties in the CDF's do not significantly diminish the value of post-processed ensemble traces for decision making, and that probabilistic prediction based on post-processed ensemble traces significantly improves the value of deterministic model prediction at all ranges of flow. In real world situations where the parameters may be updated as necessary (e.g. annually), it is expected that the performance of the post processor would be significantly better than the independent validation results seen in this paper.

The evaluation of probabilistic prediction carried out in this work is limited to dealing with hydrologic uncertainty only. To assess the integrated effects of combined uncertainty, it is necessary to extend the scope of evaluation to include both meteorological and hydrologic uncertainties. In this work, the uncertainty in the initial conditions of the hydrologic models is reduced implicitly by the inclusion of observed flow in the statistical prediction model (see Eq. 4). Being purely statistical, such an approach is of limited effectiveness. State updating or DA that explicitly addresses uncertainty in the model initial conditions (see e.g. Refsgaard, 1998; Sperflage and Georgakakos, 1996; Seo et al., 2003) should be incorporated to reduce the aggregate hydrologic uncertainty and to simplify statistical modeling. Finally, the significant sensitivity seen in this work of the post-processor's performance to long-term variability in the probability distribution

of observed flow suggests that the distributional parameters used in probability matching need to be updated frequently to minimize the effects of sampling errors due to natural variability of streamflow, and that further work is needed to assess such effects under climate change.

Acknowledgements. This work is supported by the Advanced Hydrologic Prediction Service (AHPS) Program of the NOAA/National Weather Service (NWS). This support is gratefully acknowledged. The authors would like to thank N. Pryor and D. Zanzalari of the NWS Middle-Atlantic River Forecast Center (MARFC), State College, PA, for providing the flow data and information pertaining to model calibration.

References

Anderson, E. A.: National Weather Service River Forecast System - Snow Accumulation and Ablation Model, NOAA Technical Memorandum NWS Hydro-17, U.S. Dept. of Commerce, Silver Spring, MD, 217pp, 1973.

Anderson, E. A.: A continuous incremental antecedent precipitation index (API) model for use with the National Weather Service River Forecast System, Proceedings of ASCE Int. Symp. On Eng. Hydrol., San Francisco, CA, 25 July, 1993.

Beven, K. J. and Binley, A. M.: The future of distributed models: Model calibration and uncertainty prediction, Hydrol. Processes, 6, 29–44, 1992.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11–29, 2001.

Binley, A. M., Beven, K. J., Calver, A., and Watts, L. G.: Changing responses in hydrology: Assessing the uncertainty in physically based model predictions, Water Resour. Res., 27(6), 1253–1261, 1991.

Box, G. and Jenkins, G.: Time Series Analysis: Forecasting and Control. Holden Day, San Francisco, USA, 1976.

Bradley, A. A., Schwartz, S. S., and T. Hashino, T.: Distributions-oriented verification of ensemble streamflow predictions, J. Hydrometeor., 5, 532–545, 2004.

Bras, R. L. and Rodriguez-Iturbe, I.: Random Functions and Hydrology, Addison-Wesley, 559pp, 1985.
Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modeling uncertainty for streamflow simulation, J. Hydrol., 298, 242–266, 2004.
Carpenter, T. M. and Georgakakos, K. P.: Assessment of Folsom lake response to historical and potential future climate scenarios: 1. Forecasting, J. Hydrol, 249(1–4), 148–175, 2001.
Carpenter, T. M. and Georgakakos, K. P.: Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model, J. Hydrol, 298, 202–221, 2004.
Day, G. N.: Extended streamflow forecasting using NWSRFS, J. Water Resour. Plann. Manage., 111(2), 157–170, 1985.
Deutsch, C. V. and Journel, A. G.: GSLIB Geostatistical Software Library and User's Guide, Oxford University Press, 340pp, 1992.
Duan, Q., Gupta, H., Sorooshian, S., Rousseau, A., and Turcotte, R. (Ed.): Advances in Calibration of Watershed Models, Water Science and Application Series 6, American Geophysical Union, Washington, D.C., p. 345, 2003.
Duan, Q., Schaake, J., Andreassian, V., Franks, S., Gupta, H. V., Gusev, Y. M., Habets, F., Hall, A., Hays, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O. N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., and Wood, E. F.: Model Parameter Estimation Experiment (MOPEX): Overview and summary of the second and third workshop results, J. Hydrol, 320(1–2), 3–17, 2006.
Frantz, K., Ajami, N., Schaake, J., and Buizza, R.: Hydrologic Ensemble Prediction Experiment focuses on reliable forecasts, Section News Hydrology, Eos, 86(25), 23, 2005.
Georgakakos, K. P., Seo, D.-J., Gupta, H., Schaake, J., and Butts, M. B.: Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, J. Hydrol., 298(1–4), 222–241, 2004.
Glahn, B.: Discussion of verification concepts in forecast verification: A practitioner's guide in atmospheric science, Weather and Forecasting, 19(4), 769–775, 2004.
Gupta, H. V., Sorooshian, S., and Yapo, P.O.: Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, Water Resour. Res., 24(4), 751-763, 1998.
Hashino, T., Bradley, A. A., and Schwartz, S. S.: Verification of probabilistic streamflow fore-

casts, IIHR Report No. 427, IIHR-Hydroscience & Engineering and Dept. of Civil and Environ. Eng., The Univ. of Iowa, Iowa City, IA, 125pp, 2002.
Jolliffe, I. T. and Stephenson, D. B. (Eds.): Forecast verification: a practitioner's guide in atmospheric science. J. Wiley, 240pp, 2003.
Krzysztofowicz, R.: Bayesian theory of probabilistic forecasting via deterministic hydrologic model, Water Resour. Res., 35(8), 2739–2750, 1999.
Krzysztofowicz, R. and Maranzano, C. J.: Bayesian system for probabilistic stage transition forecasting, J. Hydrol., 299, 15–44, 2004.
Kuczera, G. and Parent, E.: Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, J. Hydrol., 211, 69–85, 1998.
Liu, Z., Martina, M. L. V., and Todini, E.: Flood forecasting using a fully distributed model: application of the TOPKAPI model to the Upper Xixian Catchment, Hydrol. Earth Syst. Sci., 9(4), 347–364, 2005.
Maranzano, C. J. and Krzysztofowicz, R.: Identification of likelihood and prior dependence structures for hydrologic uncertainty processor, J. Hydrol., 290(1–2), 1–21), 2006.
McEnery, J., Ingram, J., Duan, Q., Adams, T., and Anderson, L.: NOAA's Advanced Hydrologic Prediction Service: Building pathways for better science in water forecasting, Bull. Amer. Meteorol. Soc., 86(3), 375–385, 2005.
McMichael, C. E., Hope, A. S., and Loaiciga, H. A.: Distributed hydrological modeling in California semi-arid shrublands: MIKE SHE model calibration and uncertainty estimation, J. Hydrol., 317, 307–324, 2006.
National Weather Service: The National Weather Service River Forecast System (NWSRFS) user's manual, National Weather Service, Office of Hydrologic Development, Silver Spring, MD, 2000.
Pappenberger, F., Beven, K., Horritt, M., and Blazkova, S.: Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations, J. Hydrol., 302, 46–69, 2005.
Perica, S., Schaake, J. C., and Seo, D.-J.: Accounting for hydrologic model errors in ensemble streamflow prediction, Preprints, 14th Conf. on Hydrol., Dallas, TX, 10–15 January, 1999.
Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to calibrate forecast ensembles, Mon. Wea. Rev., 133, 1155–1174, 2005.
Refsgaard, J. C.: Validation and intercomparison of different updating procedures for real-time forecasting, Nord. Hydrol., 28, 65–84, 1998.

Refsgaard, J. C. and Knudsen, J.: Operational validation and intercomparison of different types of hydrological models, Water Resour. Res., 32(7), 2189–2202, 1996.

Schaake, J. and Larson, L.: Ensemble streamflow prediction (ESP): Progress and research needs, in Preprints, Special Symposium on Hydrology, J19–J24, Am. Meteorol. Soc., Boston, MA, 1998.

Seo, D.-J., Perica, S., and Schaake, J.: Simulation precipitation fields from Probabilistic Quantitative Precipitation Forecast, J. Hydrol., 239, 203–229, 2000.

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

Slater, A. G. and Clark, M. P.: Snow Data Assimilation via an Ensemble Kalman Filter, J. Hydrometeorol., 7(3), 478–493, 2006.

Smith, J. A., Day, G. N., and Kane, M. D.: Nonparametric framework for long-range streamflow forecasting, J. Water Resour. Plann. Manage., 118(1), 82–91, 1992.

Smith, M. B., Seo, D.-J., Koren, V. I., Reed, S. M., Zhang, Z., Duan, Q., Moreda, F., and Cong, S.: The distributed model intercomparison project (DMIP): motivation and experiment design, J. Hydrol., 298(1–4), 4–26, 2004

Sperflage, J. A. and Georgakakos, K. P.: Implementation and testing of the HFS operational as a part of the National Weather Service River Forecast System (NWSRFS), HRC Tech. Rep. 1, Hydrologic Research Center, San Diego, CA, 213pp, 1996.

Todini, E.: Rainfall-runoff Models for Real-time Forecasting, Encyclopedia of Hydrological Sciences Vol. 3, Part 11: Rainfall-runoff Modeling, Section 123, 1869–1896, John Wiley and Sons, 2005.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S.: A Shuffled Complex Evolution Metropolis algorithm for optimization of uncertainty assessment of hydrological model parameters, Water Resour. Res., 39(8), 1201, doi:10.1029/2002WR00164, 2003.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press, San Diego, CA, 465pp, 1995.

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P.: Long range experimental hydrologic forecasting for the eastern U.S., J. Geophy. Res. Atmos., 107(D20), 4429, doi:10.1029/2001JD000659, 2002.

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K.: The economic value of ensemble-based weather forecasts. Bull. Am. Meteorol. Soc., 83, 73–83, 2002.

2015

Zupanski, M.: Maximum likelihood ensemble filter: Theoretical aspects, Mon. Wea. Rev, 133, 1710–1726, 2005.

2016

Table 1. Juniata River basins used in the study.

Basin ID	Gauging Station	USGS ID	Drainage Area (km^2)	Mean Elevation (m)
SPKP1	Spruce Creek	01558000	754	471
HUNP1	Huntingdon	01559000	2113	447
SAXP1	Saxton	01562000	1958	463
MPLP1	Mapleton Depot	01563500	5258	353
NPTP1	Newport	01567000	8687	258

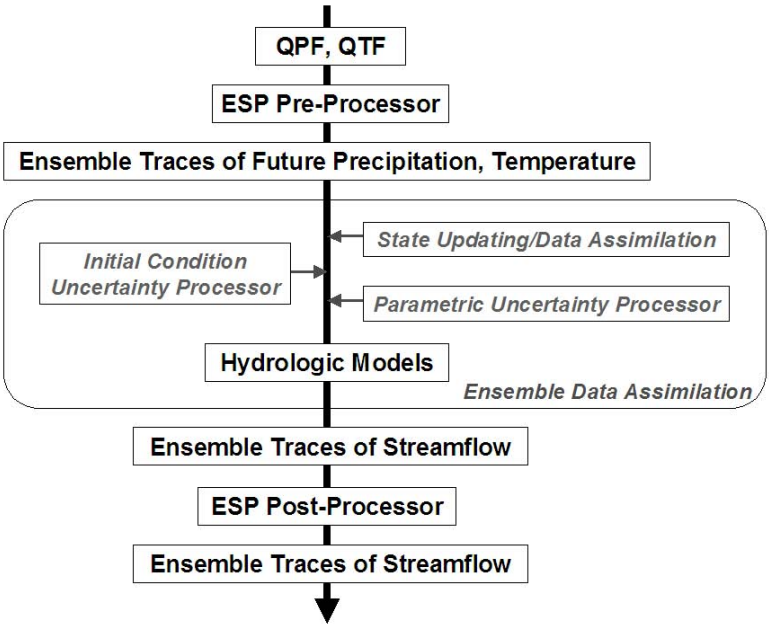

Fig. 1. Schematic of the ensemble streamflow prediction (ESP) process.

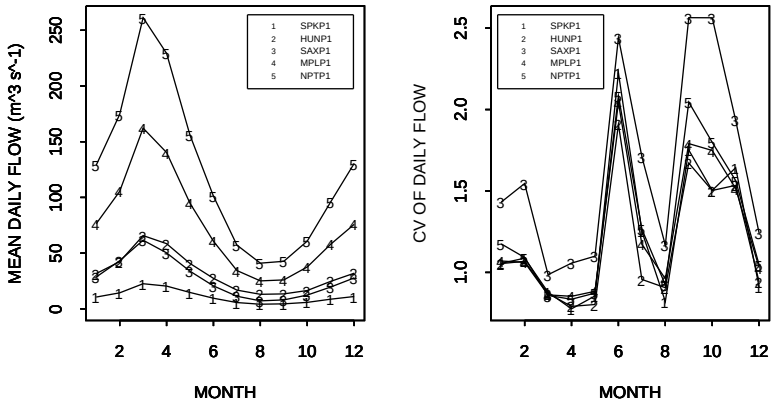

Fig. 2. Monthly mean (left panel) and coefficient of variation (CV, right panel) of daily flow at the five study basins (see Table 1) of the Juniata River, PA, USA.

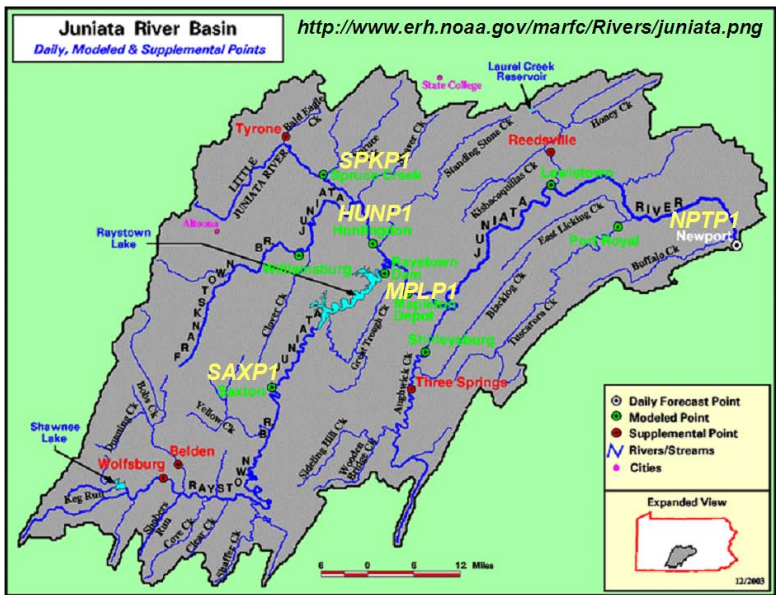

Fig. 3. Map of the Juniata River basin, PA, USA. The state of Pennsylvania is depicted in the lower-right corner.

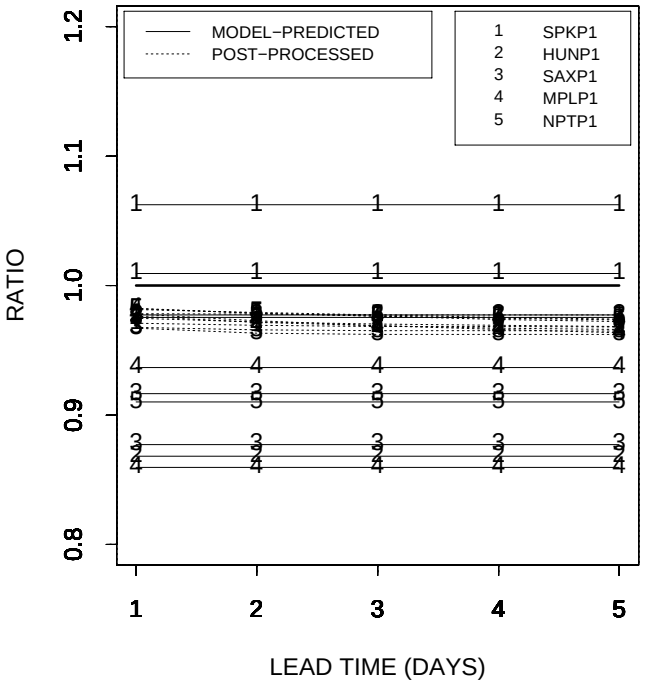

Fig. 4. Ratio of the sum of the observed flow to that of the model-predicted (in solid line) or the post-processed (in dotted line) in the parameter estimation period versus lead time (days).

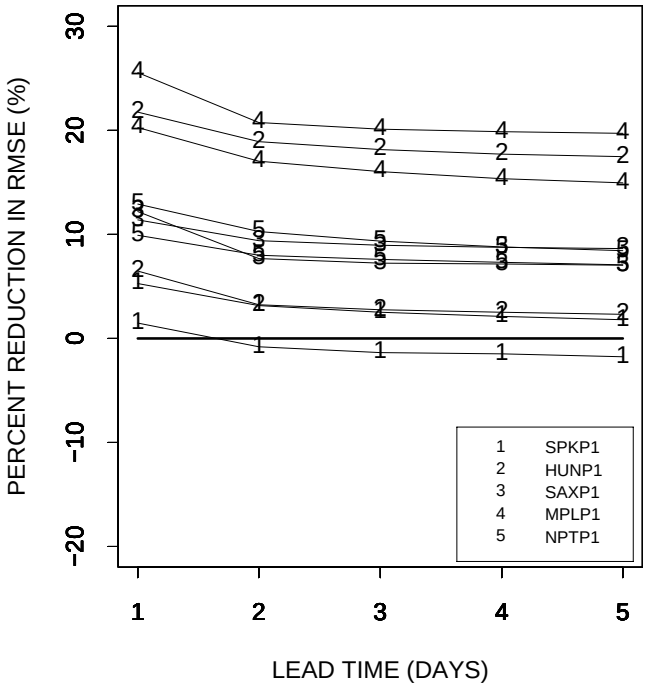

Fig. 5. Percent reduction in root mean square error (RMSE) by the post-processed flow over the model-predicted in the parameter estimation periods versus lead time (days).

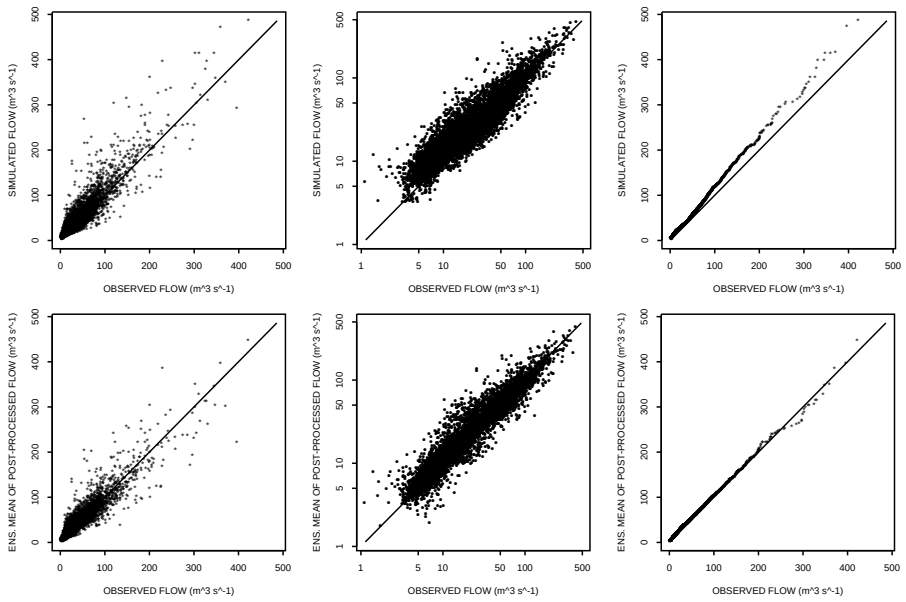

Fig. 6. Scatter-plots in linear scale (left panels) and in log scale (middle panels), and quantile-quantile plots (right panels) of daily flow between the observed and the model-simulated flows (upper panels) and between the observed and the post-processed flows (lower panels) at HUNP1 for a parameter estimation period. The lead time is 1 day.

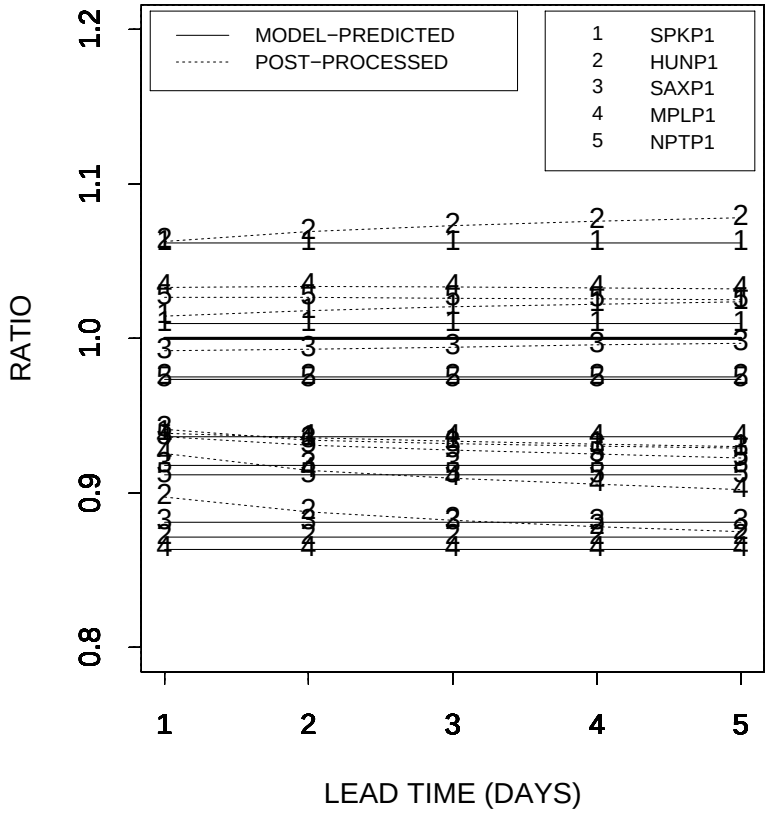

Fig. 7. Same as Fig. 4, but for the validation periods.

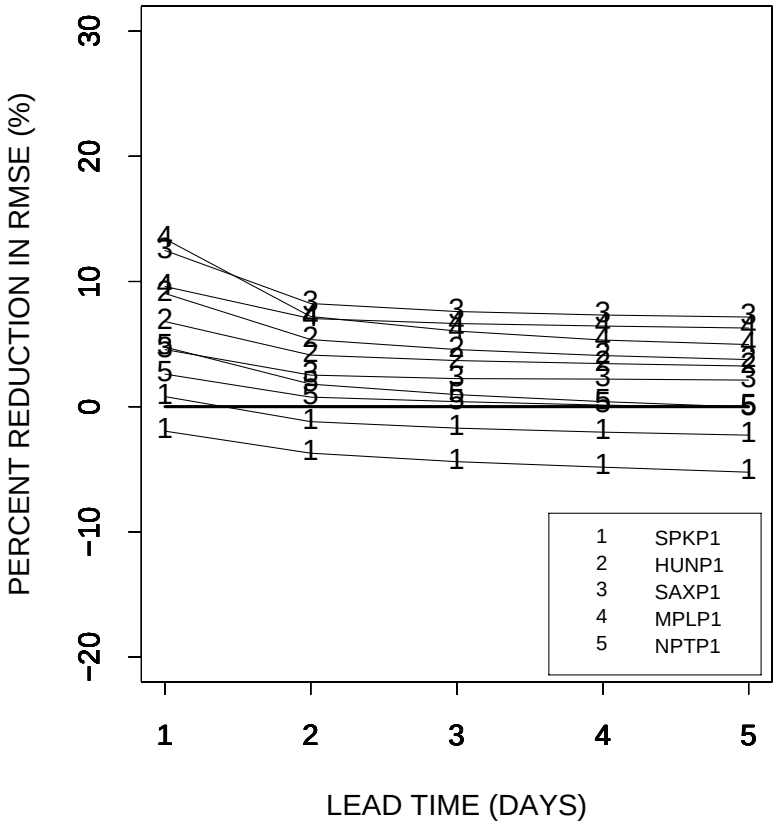

Fig. 8. Same as Fig. 5, but for the validation periods.

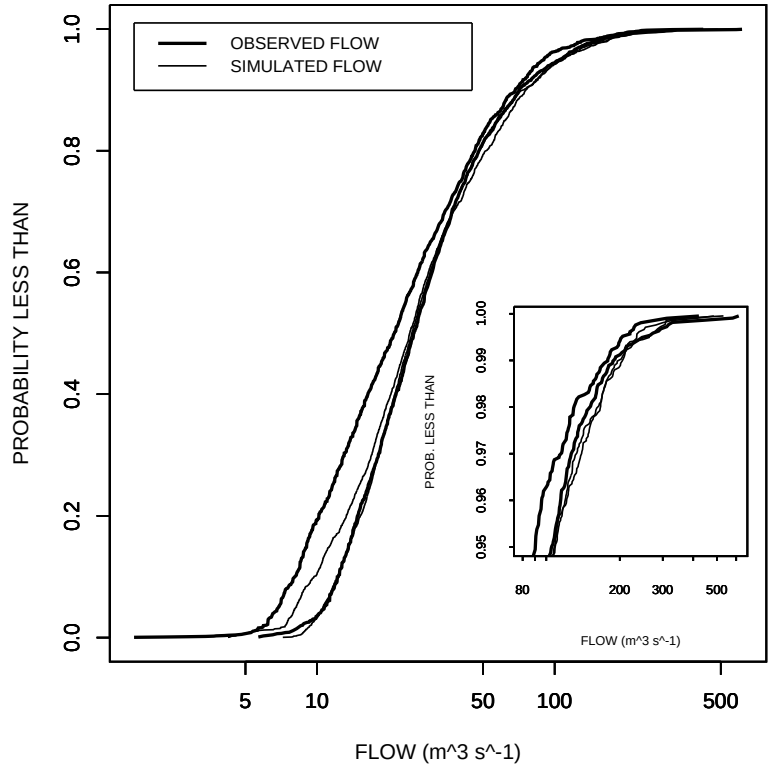

Fig. 9. Empirical cumulative distribution functions (ECDF) of observed (thick solid line) and simulated (thin solid line) flows at HUNP1. The upper-tail portion is magnified in the lower-right corner.

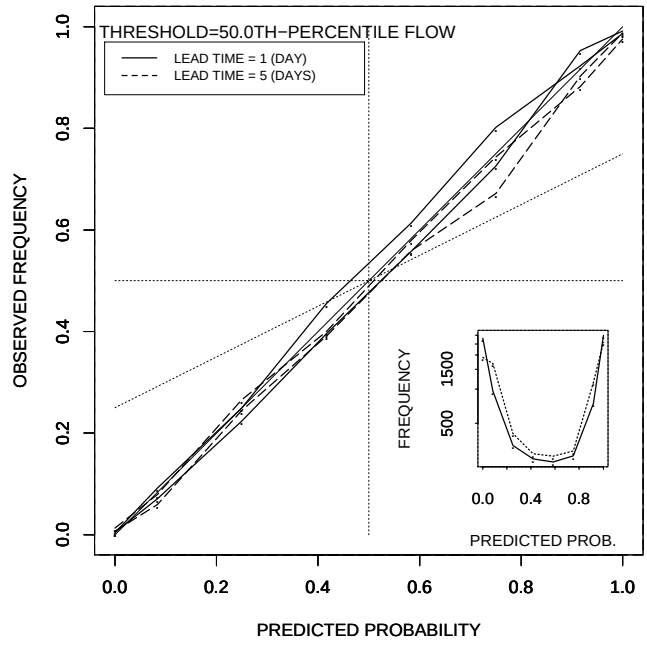

Fig. 10. Reliability diagrams for probabilistic prediction at NPTP1 based on post-processed ensemble traces in the parameter estimation periods. The diagrams for 1 and 5 day-ahead predictions are in solid and dashed lines, respectively. The two lines for each lead time correspond to the two parameter estimation periods. The threshold is the median flow. The solid and dotted lines in the lower-right corner are the histograms of the predicted probability for lead times of 1 and 5 days, respectively.

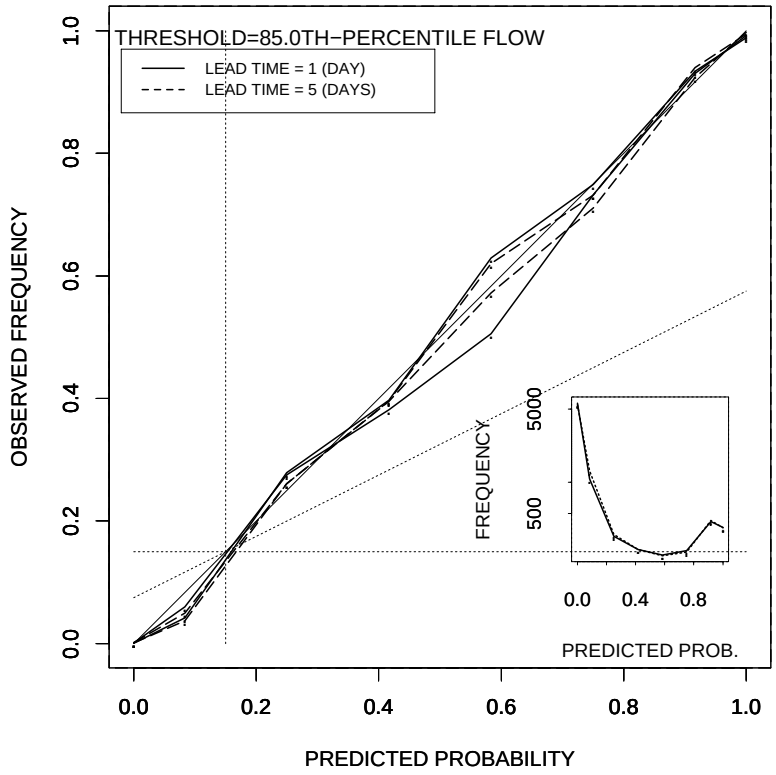

Fig. 11. Same as Fig. 10, but for the threshold of 85th-percentile flow.

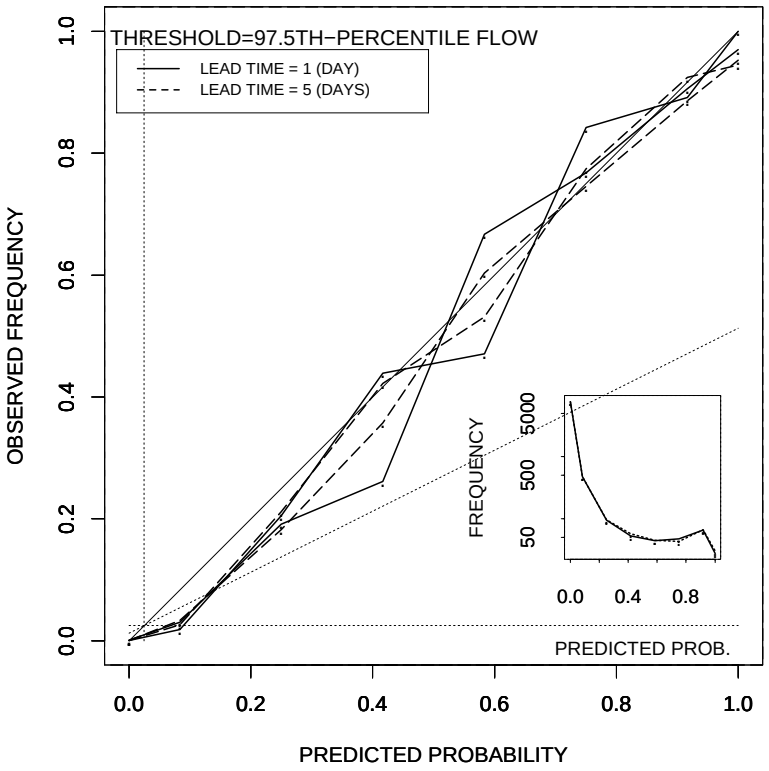

Fig. 12. Same as Fig. 11, but for the threshold of 97.5th-percentile flow.

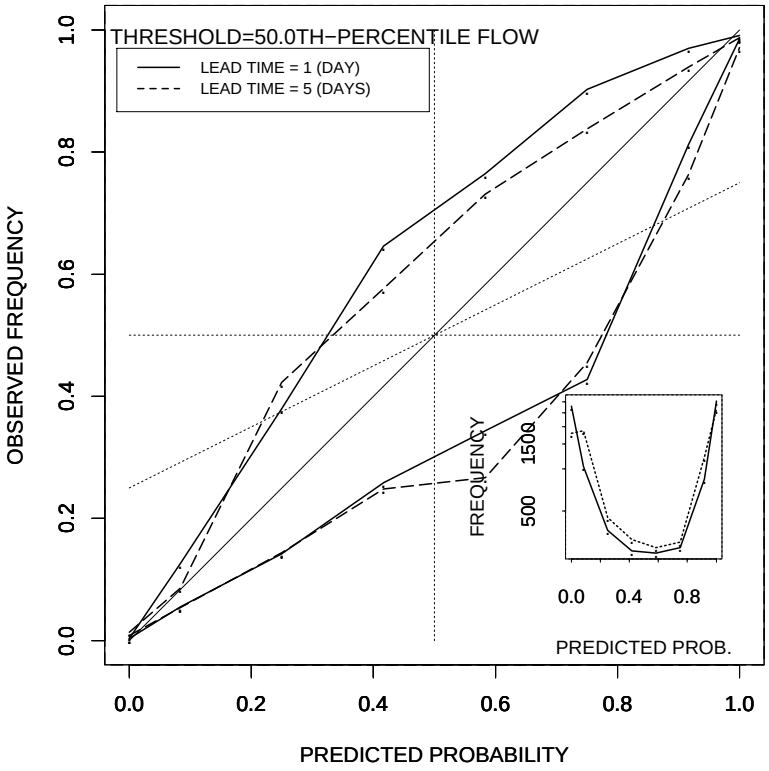

Fig. 13. Same as Fig. 10, but in the validation periods.

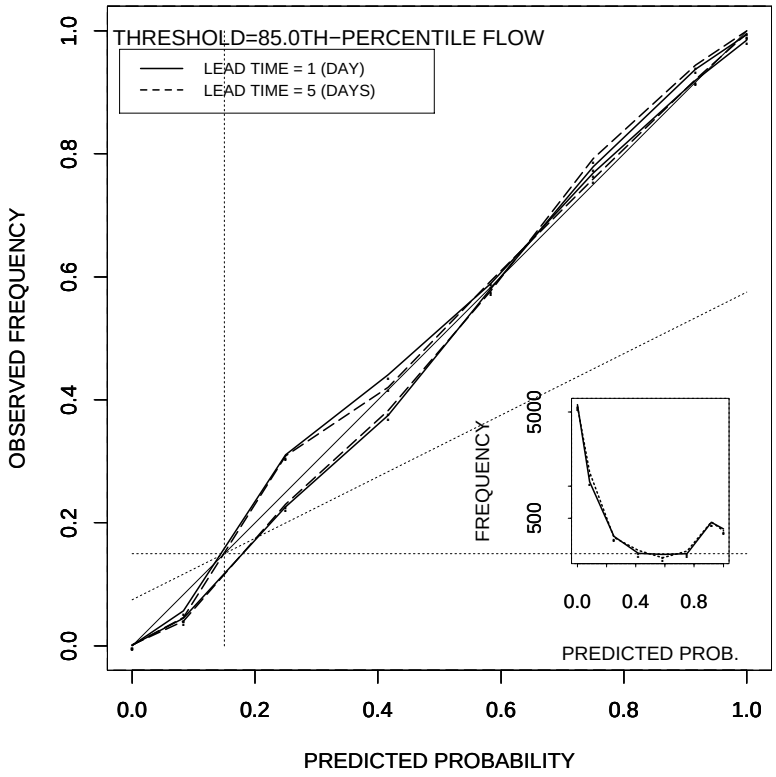

Fig. 14. Same as Fig. 13, but for the threshold of 85th-percentile flow.

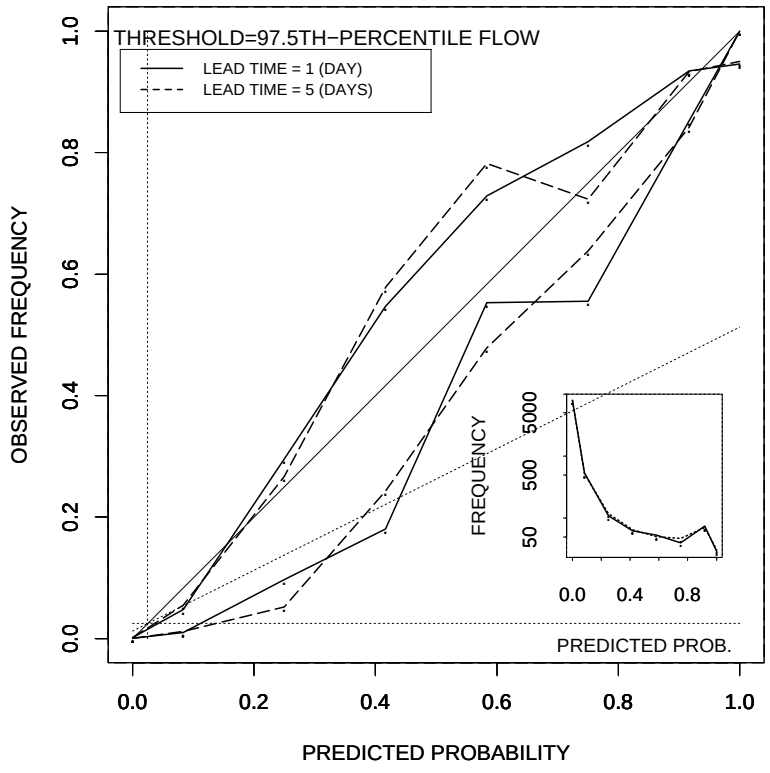

Fig. 15. Same as Fig. 13, but for the threshold of 97.5th-percentile flow.

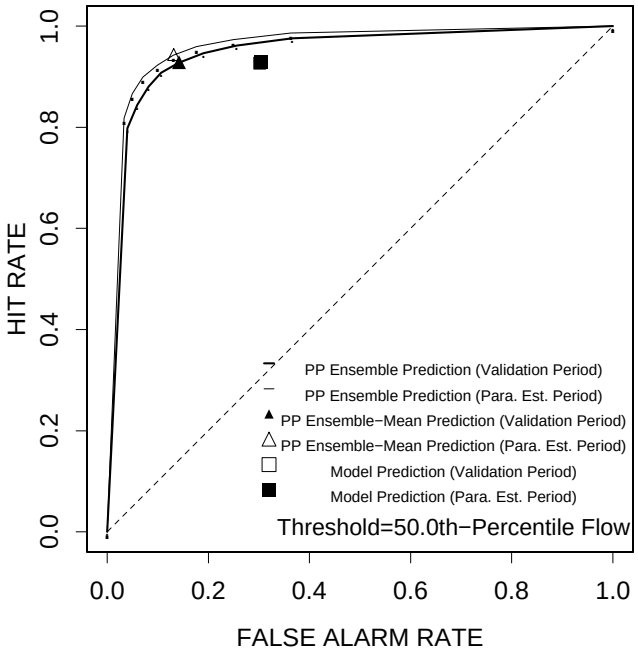

Fig. 16. Relative Operating Characteristic (ROC) curves of the post-processed ensemble traces at HUNP1 in the parameter estimation (thin solid line) and validation (thick solid line) periods. The threshold is the median flow. The lead time is 1 day. The squares and triangles denote the (hit rate, false alarm rate) positions of the (raw) model prediction and ensemble mean prediction from the post-processor, respectively. The solid and empty markers denote the positions in the validation and parameter estimation periods, respectively.

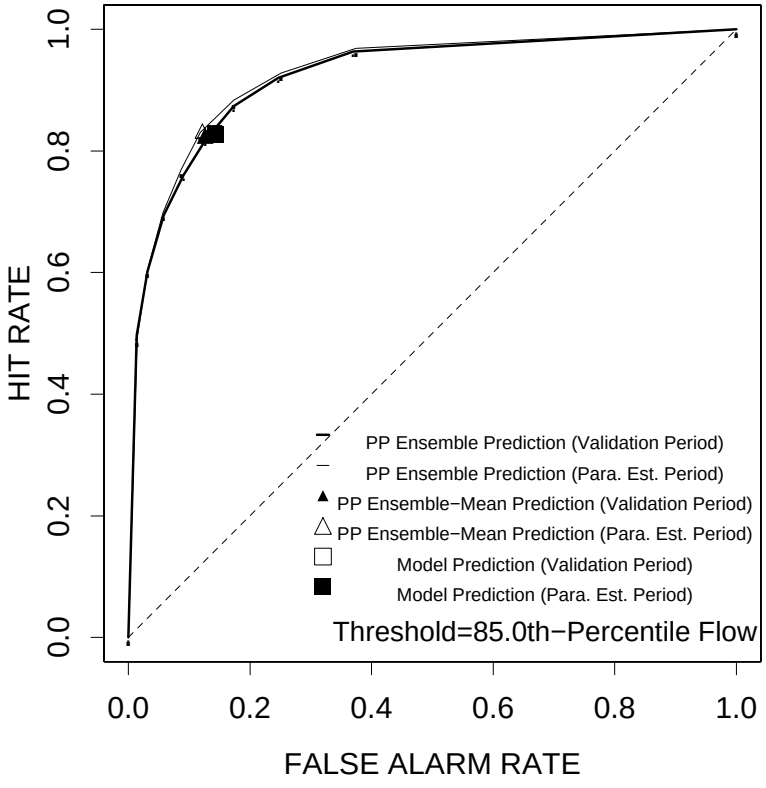

Fig. 17. Same as Fig. 16, but for the threshold of 85.0th-percentile flow.

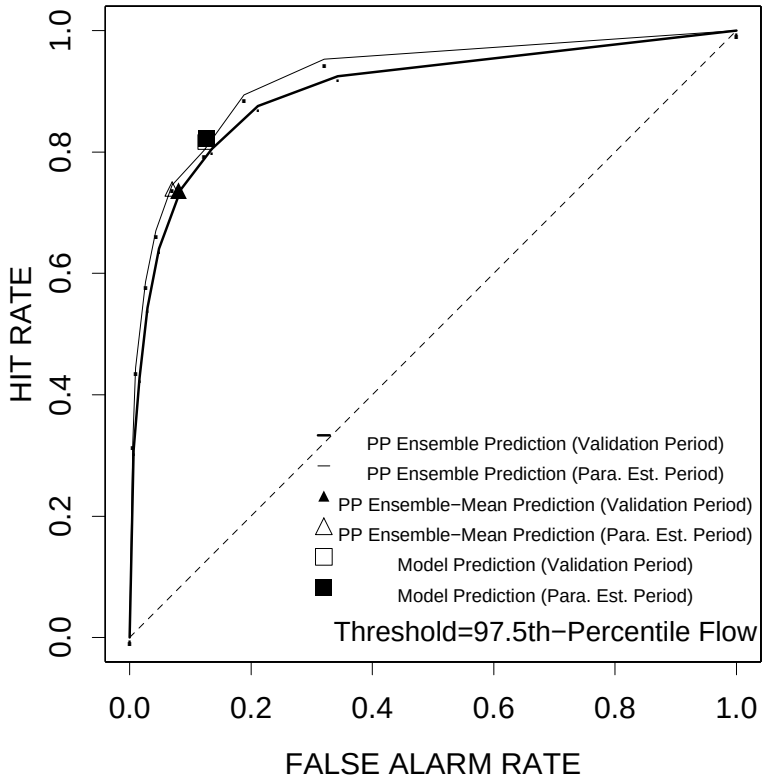

Fig. 18. Same as Fig. 16, but for the threshold of 97.5th-percentile flow.