A Nonparametric Postprocessor for Bias Correction of Hydrometeorological and Hydrologic Ensemble Forecasts

JAMES D. BROWN AND DONG-JUN SEO

NOAA/National Weather Service, Office of Hydrologic Development, Silver Spring, Maryland, and University Corporation for Atmospheric Research, Boulder, Colorado

(Manuscript received 25 June 2009, in final form 10 December 2009)

ABSTRACT

This paper describes a technique for quantifying and removing biases from ensemble forecasts of hydrometeorological and hydrologic variables. The technique makes no a priori assumptions about the distributional form of the variables, which is often unknown or difficult to model parametrically. The aim is to estimate the conditional cumulative distribution function (ccdf) of the observed variable given a (possibly biased) real-time ensemble forecast. This ccdf represents the ''true'' probability distribution of the forecast variable, subject to sampling uncertainties. In the absence of a known distributional form, the ccdf should be estimated nonparametrically. It is noted that the probability of exceeding a threshold of the observed variable, such as flood stage, is equivalent to the expectation of an indicator variable defined for that threshold. The ccdf is then modeled through a linear combination of the indicator variables of the forecast ensemble members. The technique is based on Bayesian optimal linear estimation of indicator variables and is analogous to indicator cokriging (ICK) in geostatistics. By developing linear estimators for the conditional expectation of the observed variable at many thresholds, ICK provides a discrete approximation of the full ccdf. Since ICK minimizes the conditional error variance of the indicator variable at each threshold, it effectively minimizes the continuous ranked probability score (CRPS) when infinitely many thresholds are employed. The technique is used to bias-correct precipitation ensemble forecasts from the NCEP Global Ensemble Forecast System (GEFS) and streamflow ensemble forecasts from the National Weather Service (NWS) River Forecast Centers (RFCs). Split-sample validation results are presented for several attributes of ensemble forecast quality, including reliability and discrimination. In general, the forecast biases were substantially reduced following ICK. Overall, the technique shows significant potential for bias-correcting ensemble forecasts whose distributional form is unknown or nonparametric.

1. Introduction

Forecasts of hydrometeorological and hydrologic variables often contain large uncertainties (Beven and Binley 1992; Anderson and Bates 2001; Gupta et al. 2005; NRC 2006; Ajami et al. 2007). Ensemble techniques are widely used in meteorology and, increasingly, in hydrology to quantify these uncertainties (Stensrud et al. 1999; Jolliffe and Stephenson 2003; Brown and Heuvelink 2005; Olsson and Lindström 2008). For example, the National Weather Service (NWS) River Forecast Centers (RFCs) produce ensemble forecasts of streamflow at a variety of lead times (Seo et al. 2006; Schaake et al. 2007). In one experimental operation, ensemble traces of precipitation and temperature are generated from single-valued forecasts using an ensemble preprocessor (Schaake et al. 2007). These traces are input into the Ensemble Streamflow Prediction (ESP) subsystem of the NWS River Forecast System (NWSRFS), from which ensemble traces of streamflow are output. To meet the varied needs of the RFCs and their customers for probabilistic water forecasts, the Experimental Ensemble Forecast System (XEFS) is currently being developed. Upon completion, the XEFS will support the quantification of uncertainty, its propagation through the forecast system, correction of forecast biases, verification of probabilistic forecasts, and the generation of a wide range of graphical and numerical outputs for scientific research and operational use (Demargne et al. 2009).

Whether they explicitly account for uncertainty or not, forecasts of environmental variables are subject to error. These errors may be correlated in space and time

Corresponding author address: James Brown, NOAA/NWS/OHD, 1325 East-West Highway, Silver Spring, MD 20910.
E-mail: james.d.brown@noaa.gov

DOI: 10.1175/2009JHM1188.1

and may be systematic. The skill of an ensemble forecasting system can depend largely on its systematic biases (Jolliffe and Stephenson 2003; Hashino et al. 2006; Wilczak et al. 2006). Forecast evaluation or ''verification'' is necessary to identify these biases and to establish the skill of the forecasting system under a range of observed and forecast conditions. To this end, verification studies are usually diagnostic in nature; they seek to quantify the biases present under a range of conditions (e.g., Bradley et al. 2004; Hersbach 2000; Georgakakos 2003; Franz et al. 2003; Murphy and Winkler 1987; Roulston and Smith 2002). Such studies can lead to improvements in forecasting models and methods. In operational forecasting, however, there is a need to estimate the quality of a *specific* forecast in real time and, if necessary, to correct for biases in that forecast. This is equivalent to estimating the probability distribution of the observed variable given the real-time ensemble forecast. The same problem is addressed in postprocessing, whereby a statistical relationship is developed between common attributes of the past forecasts and observations (e.g., precipitation amount and storm type) and used to bias-correct subsequent forecasts, conditional upon those attributes. For example, a common application of model output statistics (MOS) involves a linear regression between a single-valued forecast amount and an observed amount (Glahn and Lowry 1972). Gneiting et al. (2005) extend this approach to ensemble forecasts by including the ensemble spread alongside the ensemble mean in estimating the observed amount. Other examples of parametric postprocessors include logistic regression (Hamill et al. 2008; Wilks 2009) and Bayesian model averaging (Raftery et al. 2005; Sloughter et al. 2007).

Reliable estimation of the joint probability distribution of forecasts and observations, or attributes thereof, is central to all types of statistical postprocessing and verification (Jolliffe and Stephenson 2003; Murphy and Winkler 1987; Wilks 1995). In many cases, a parametric joint distribution is assumed. However, many hydrometeorological and hydrologic variables do not follow a parametric distribution. This is not surprising because the observed outcomes are an aggregate effect of different physical processes operating at different temporal and spatial scales. To circumvent this problem, the sample data may be transformed to follow a parametric distribution and a parametric assumption invoked (Gneiting et al. 2005; Schaake et al. 2007). For example, when postprocessing hydrometeorological and hydrologic ensembles, it is often assumed that the (transformed) forecasts and observations are joint normally distributed. However, distributional transforms, such as the normal quantile transform (NQT; Kelly and Krzysztofowicz 1997; Goovaerts 1997), are complicated by the need to model in transformed space and then back transform the estimated probabilities, where the optimality of the parameter estimates is no longer guaranteed. Furthermore, marginal transforms are not a sufficient condition to ensure that the joint probabilities follow a parametric joint distribution (Goovaerts 1997).

This paper describes a new approach for correcting biases in real-time ensemble forecasts of hydrometeorological and hydrologic variables. The technique makes no a priori assumptions about the distributional form of the variables and is calibrated with a large sample of historical ensemble forecasts and verifying observations. It attempts to estimate the probability distribution of the observed variable given the ensemble members of the real-time forecast as conditioning information. The technique is based on Bayesian optimal linear estimation of indicator variables (Schweppe 1973) and is analogous to simple cokriging of indicator variables (ICK) in geostatistics (Journel and Huijbregts 1978; Isaaks and Strivastava 1989; Deutsch and Journel 1992; Cressie 1993; Seo 1996). The paper is organized in three parts. The first part outlines the ICK approach for modeling the conditional distribution. The second part focuses on estimating the conditional distribution, where the aim is to develop a parsimonious estimate given the information content of the ensemble members. In the third part, the technique is applied to ensemble forecasts of precipitation from the National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS) and ensemble forecasts of streamflow from the NWS RFCs. Split-sample (independent) validation results are presented for several attributes of ensemble forecast quality, including reliability and discrimination.

2. Problem formulation and proposed solution

Let G denote a numeric variable of interest, such as temperature, amount of precipitation, or streamflow. Let X denote an observation of that variable at a single ''point'' in space and time, and let Y denote a corresponding forecast. Depending on the estimation problem, the scale or space–time support of X and Y can vary (e.g., a location or an area), and the support of X may differ from that of Y. For example, gauged precipitation, X, might be estimated from an ensemble forecast of mean areal precipitation, Y. For simplicity, it is assumed that X is an unbiased estimate of G, because these biases are either unknown (more commonly) or are known and, therefore, removable (see Jolliffe and Stephenson 2003; Katz and Murphy 1997; Wilks 1995). For brevity, the random variables X and Y are denoted by their experimental values only. Thus, the probability density functions of X and Y are denoted $f(x)$ and $f(y)$ and their

cumulative distribution functions (cdfs) are denoted $F(x)$ and $F(y)$, respectively. The joint cumulative distribution function of X and Y is denoted $F(x, y)$. For most practical applications, $F(y)$ is approximated numerically using a vector of m equally likely ensemble members, denoted $\mathbf{Y} = \{Z_1, \ldots, Z_m\}$. Here, it is assumed that the ensemble members are ranked by size, that is, $Z_{i-1} \le Z_i$, $i = 2, \ldots, m$. The problem then is to estimate the conditional cdf (ccdf) of X given Y for an operational forecast with ensemble member values, $\mathbf{y} = \{z_1, \ldots, z_m\}$:

$$F(x|z_1, \ldots, z_m) = \text{Prob}[X \le x | Z_1 = z_1, \ldots, Z_m = z_m]. \quad (1)$$

The probability that X is less than or equal to a threshold, c_a, is equivalent to the expectation of an indicator function

$$\text{Prob}[X \le c_a] = E[I(x; c_a)] \quad \text{where}$$
$$I(x; c_a) = \begin{cases} 1, & x \le c_a; \\ 0, & \text{otherwise}, \end{cases} \quad (2)$$

and $E[\cdot]$ is the expectation operator. The indicator variable $I(x; c_a)$ is a Bernoulli random variable with mean or "probability of success" $E[I(x; c_a)] = p$ and variance $\text{VAR}[I(x; c_a)] = p(1 - p)$. The joint probability that X is less than or equal to c_a and Y is less than or equal to a threshold, c_b, is

$$\text{Prob}[X \le c_a, Y \le c_b] = E[I(x; c_a) \times I(y; c_b)] \quad \text{where}$$
$$I(x; c_a) = \begin{cases} 1, & x \le c_a; \\ 0, & \text{otherwise} \end{cases} \quad \text{and}$$
$$I(y; c_b) = \begin{cases} 1, & y \le c_b; \\ 0, & \text{otherwise}. \end{cases} \quad (3)$$

The indicator variable $I(y; c_b)$ is a Bernoulli random variable with mean $E[I(y; c_b)] = q$ and variance $\text{VAR}[I(y; c_b)] = q(1 - q)$. In general, X and Y can assume an infinite number of possible outcomes, requiring an infinite number of indicator variables to fully define their joint distribution. While some applications require an estimate of the conditional probability of the observed variable at a specific threshold (e.g., c_a = flood stage), others require an estimate of the full ccdf. Even for a fixed value of $X = c_a$, the conditional probability in Eq. (1) comprises an infinite number of partitions of Y. However, a discrete approximation of the continuous ccdf is possible.

For an ensemble forecast comprising m members, the aim is to estimate $F(c_a|z_1, \ldots, z_m) = \text{Prob}[X \le c_a | Z_1 = z_1, \ldots, Z_m = z_m]$ for all possible values of c_a. In practice, an estimate is made at only some values of c_a. For a given value of c_a, the conditional probability is obtained from a finite number, υ, of the infinite number of possible indicator transforms of the ensemble members

$$E[I(x; c_a)|Z_1 = z_1, \ldots, Z_m = z_m] \approx E[I(x; c_a)|I(z_j; c_b) = i(z_j; c_b); \quad j = 1, \ldots, m; \quad b = 1, \ldots, \upsilon]$$
$$\text{and} \quad I(x; c_a) = \begin{cases} 1, & x \le c_a; \\ 0, & \text{otherwise}; \end{cases}$$
$$\text{and} \quad I(z_j; c_b) = \begin{cases} 1, & z_j \le c_b; \\ 0, & \text{otherwise}, \end{cases} \quad (4)$$

where $i(z_j; c_b)$ denotes the experimental value of $I(z_j; c_b)$. The conditional expectation in Eq. (4) may be estimated in several ways. Here, we use the Bayesian optimal linear estimator (optimal in the sense of minimum error variance; Schweppe 1973),

$$E[I(x; c_a)|I(z_j; c_b) = i(z_j; c_b); \quad j = 1, \ldots, m; \quad b = 1, \ldots, \upsilon] \approx E[I(x; c_a)] + \sum_{b=1}^{\upsilon} \sum_{j=1}^{m} \lambda_a(z_j; c_b) \times \{i(z_j; c_b) - E[I(z_j; c_b)]\}, \quad (5)$$

where each $\lambda_a(z_j; c_b)$ is a weight formed at the ath threshold of the observed variable, which is used to weigh the jth ensemble member at the bth threshold of the forecast variable (see section 3 on estimation). In Eq. (5), $E[I(x, c_a)]$ is the prior or "climatological" probability that X is less than or equal to c_a. The addition to this prior probability represents a conditional adjustment by the ensemble forecast. When averaging over an infinite number of conditional adjustments, the prior distribution is preserved, because the conditional adjustment has zero expectation, that is, $E[i(z_j; c_b) - E[I(z_j; c_b)]; j = 1, \ldots, m; b = 1, \ldots, \upsilon] = 0$. In other words, the unconditional probabilities are necessarily unbiased. Equation (5) is analogous to simple cokriging of indicator variables in geostatistics (Journel and Huijbregts 1978; Isaaks and Srivastava 1989; Cressie 1993), where $E[I(x, c_a)]$ is termed the "simple-kriging mean" and the $\lambda_a(z_j; c_b)$ are cokriging weights to estimate.

Kriging is closely related to other interpolators, such as regression splines (Dubrule 1983; Watson 1984) and kernel methods (Buja et al. 1989; Yandell 1993). Under certain conditions, these techniques are equivalent (Yandell 1993). For example, Borga and Vizzaccaro (1997) show that multiquadratic surface fitting with a conic spline is equivalent to kriging with a linear variogram model.

Elsewhere, Ali and Lall (1996) adopt a kernel estimator instead of indicator kriging to predict soil conductivity. Thus, other nonparametric techniques may be considered alongside ICK. However, in the current application, ICK employs a different objective function than the kernel approach. Specifically, ICK minimizes the Brier score (BS; Brier 1950) at each threshold, c_a, of the ccdf in Eq. (1). The BS is a well-known measure of probabilistic forecast quality (discussed later). To achieve such optimality with kernel smoothing, for example, an optimal bandwidth would be required for each forecast ensemble member, as well as the observation, at each threshold. The authors are unaware of any practical solution to such a high-dimensional estimation problem.

Unlike most parametric techniques, ICK comprises a separate linear estimator for each threshold of X. This allows the statistical dependence of X on Y to change with the forecast amount. Further, it allows for mixed distributions (such as precipitation) to be treated in the same way as continuous distributions. Indeed, the discontinuity in precipitation is simply another threshold at which to solve Eq. (5). However, it also implies a discrete approximation of the full ccdf in Eq. (1). This is formed at u thresholds of X, $\{c_1, \ldots, c_u\}$. A sufficiently large number of thresholds must be used to capture the variability in the observations and in the forecast ensemble members (see section 5). Because the estimates from Eq. (5) are probabilities, they must be greater than or equal to 0, less than or equal to 1, and nondecreasing as the threshold value increases. In practice, ICK may fail these conditions because the u optimal linear estimates are obtained separately, and their predictions are not constrained to be valid probabilities (see section 3).

In some cases, there may be prior knowledge about the biases in Y, or more knowledge about X than simply its unconditional, climatological, distribution. If this information is considered reliable and can be encoded into a set of auxiliary variables, $\mathbf{A} = \{A_1, \ldots, A_k\}$, then it may be included in the estimator. For example, precipitation forecasts are often subject to seasonal biases. Given a binary variable $A_1 = \{\text{warm season, cool season}\}$, identification of $F(x|z_1, \ldots, z_m; a_1)$ amounts to a separate cokriging for each of a_1 = warm season and a_1 = cool season.

3. Estimation of the conditional probability distribution

In practice, there is a finite sample of n historical forecasts and corresponding observations from which to estimate the statistical parameters in ICK. Together, the n pairs of forecasts and observations *are assumed to* represent independent samples of a stationary random process. In our application of ICK, this implies time-invariant marginal probabilities of the observed variable and the ranked ensemble members, as well as time-invariant joint probabilities.

There is one estimator for each indicator threshold of the observed variable, $\{c_1, \ldots, c_u\}$. Each estimator requires $mv + 1$ marginal probabilities: one for the prior probability of X at its ath indicator threshold, $E[I(x, c_a)]$, and one for the probability of Z_j at its bth indicator threshold, $E[I(z_j, c_b)]$, $j = 1, \ldots, m$; $b = 1, \ldots, v$. In the current application, all of the statistical parameters in ICK, including the marginal probabilities of each indicator variable, as well as the covariances and cross covariances between indicator variables, are estimated from sample data. This indicator covariance structure is assumed to be stationary, and the estimated structure is then used to bias-correct a real-time forecast. Unbiased estimates of the marginal probabilities can be obtained from a plotting position formula, such as the Weibull plotting position (O'Conner 2002). Sample estimates of the covariances and cross covariances are considered later.

Alongside the marginal probabilities, there are $mv \times 1$ regression coefficients to estimate for each indicator threshold of the observed variable. They are estimated by minimizing the conditional error variance of the conditional probability at the ath indicator threshold:

$$J = E\left[\left\{i(x; c_a) - E[I(x; c_a)|I(z_j; c_b) = i(z_j; c_b); \quad j = 1, \ldots, m; \quad b = 1, \ldots, v]\right\}^2 \Big| I(z_j; c_b) i = (z_j; c_b); \quad j = 1, \ldots, m; \quad b = 1, \ldots, v\right], \tag{6}$$

where $i(x; c_a)$ is the experimental value of the observed indicator variable at threshold c_a. Equation (6) is analogous to the BS at threshold c_a (Brier 1950). Since ICK estimates the conditional probability of the observed variable at each of the u (ranked) thresholds, it effectively minimizes the sum of the Brier scores across the u thresholds. Extension to infinitely many thresholds is, therefore, equivalent to minimizing the continuous ranked probability score (CRPS; Matheson and Winkler 1976; Hersbach 2000) on a threshold-by-threshold basis.

The optimal linear solution to Eq. (6) at the ath threshold of the observed variable, $\boldsymbol{\beta}_a$, is given by

$$\boldsymbol{\beta}_a = \mathbf{W}^{-1}\mathbf{b}_a$$

$$\begin{bmatrix} \boldsymbol{\theta}_a(c_1) \\ \vdots \\ \boldsymbol{\theta}_a(c_v) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\omega}(c_1, c_1) \ldots \boldsymbol{\omega}(c_1, c_v) \\ \vdots \quad \ddots \quad \vdots \\ \boldsymbol{\omega}(c_v, c_1) \cdots \boldsymbol{\omega}(c_v, c_v) \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\delta}_a(c_1) \\ \vdots \\ \boldsymbol{\delta}_a(c_v) \end{bmatrix}, \tag{7a}$$

where

$$\boldsymbol{\theta}_a(c_b) = \begin{bmatrix} \lambda_a(z_1; c_b) \\ \vdots \\ \lambda_a(z_m; c_b) \end{bmatrix}, \tag{7b}$$

$$\boldsymbol{\omega}(c_b, c_h) = \begin{bmatrix} \mathrm{Cov}[I(z_1; c_b), I(z_1; c_h)] \cdots \mathrm{Cov}[I(z_1; c_b), I(z_m; c_h)] \\ \vdots \quad \ddots \quad \vdots \\ \mathrm{Cov}[I(z_m; c_b), I(z_1; c_h)] \cdots \mathrm{Cov}[I(z_m; c_b), I(z_m; c_h)] \end{bmatrix}, \quad \text{and} \tag{7c}$$

$$\boldsymbol{\delta}_a(c_b) = \begin{bmatrix} \mathrm{Cov}[I(x; c_a), I(z_1; c_b)] \\ \vdots \\ \mathrm{Cov}[I(x; c_a), I(z_m; c_b)] \end{bmatrix} \quad b;\ h = 1, \ldots, v. \tag{7d}$$

The matrix $\mathbf{W}$ contains the covariances between each pair of indicator transforms of the m ensemble members at v thresholds. The vector $\mathbf{b}_a$ contains the cross covariances between the indicator transforms of the m ensemble members at v thresholds and the indicator transform of the observed variable at threshold c_a. In this work, the entries of $\mathbf{W}$ and $\mathbf{b}_a$ are sample covariances. In the absence of a plotting position formula for the joint probability, a normalizing constant of n (and not $n - 1$) is used to compute the sample covariances and cross covariances that populate $\mathbf{W}$ and $\mathbf{b}_a$, respectively. This is necessary to produce valid joint probabilities in the upper tails, because the marginal and joint probabilities are specified by sample data, rather than by a theoretical probability distribution.

An important aim of this work is to develop an estimate of Eq. (5) that is both computationally efficient and appropriately smooth, given the sampling uncertainties of the indicator statistics. In practice, the mv forecast indicator variables in $\mathbf{W}$ may be strongly intercorrelated (e.g., Eckel and Walters 1998). By eliminating these intercorrelations, an orthogonal decomposition of $\mathbf{W}$ should reduce the dimensionality of the estimation problem. This is analogous to cokriging with the first pp principal components of $\mathbf{W}$ such that $pp < mv$ (see Deutsch and Journel 1992). Here we adopt the singular value decomposition (SVD) of $\mathbf{W}$, which is numerically stable and is always defined (Golub and Van Loan 1996) as

$$\mathbf{W} = \mathbf{USV}^{\mathrm{T}}, \tag{8}$$

where $\mathbf{U}$ and $\mathbf{V}$ are $mv \times mv$ orthogonal matrices, whose columns form the eigenvectors of $\mathbf{WW}^{\mathrm{T}}$ and $\mathbf{W}^{\mathrm{T}}\mathbf{W}$, respectively. The matrix $\mathbf{S}$ is an $mv \times mv$ diagonal matrix, which contains the singular values of $\mathbf{W}$, that is, $\mathbf{S} = \mathrm{diag}(s_1, \ldots, s_{mv})$. By convention, the singular values are arranged in descending order from s_1 to s_{mv}. Since $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{S}$ is diagonal, they are easily inverted as follows:

$$\begin{aligned} \mathbf{W}^{-1} &= (\mathbf{USV}^{\mathrm{T}})^{-1} \\ &= (\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{S}^{-1}\mathbf{U}^{-1} \\ &= \mathbf{VS}^{-1}\mathbf{U}^{\mathrm{T}}. \end{aligned} \tag{9}$$

If $\mathbf{W}$ is singular or near singular, one or more of the diagonal entries in $\mathbf{S}$ will be zero or near zero, respectively. In practice, only those r column vectors of $\mathbf{U}$ and r row vectors of $\mathbf{V}$ corresponding to the nonzero singular values are calculated. The matrix $\mathbf{U}$ is then $mv \times r$, the matrix $\mathbf{S}$ is $r \times r$ diagonal, and the matrix $\mathbf{V}$ is $r \times mv$, where $r < mv$. To obtain a regularized solution [see Hansen (1987) for a discussion on regularization], the reciprocals of the smallest positive singular values in $\mathbf{S}$ may be substituted with zeros. The remaining pp positive singular values provide an efficient and well-posed solution to Eq. (5). The regularized form of ICK is analogous to indicator cokriging with measurement error (Saito and Goovaerts 2002). It acknowledges that both the observed and forecast

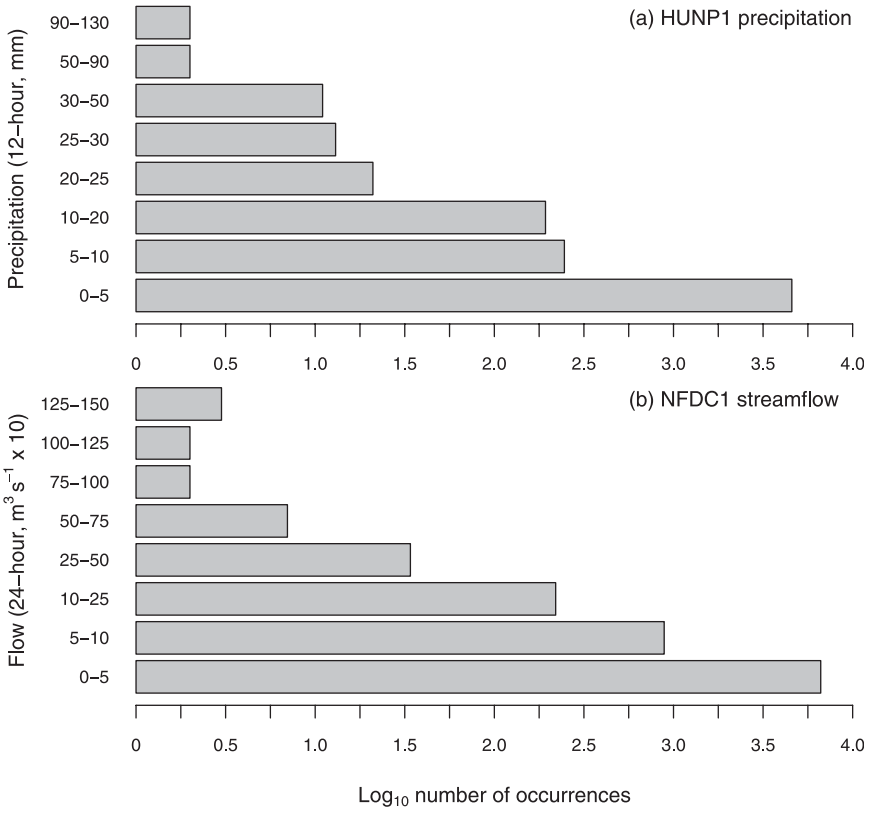

FIG. 1. Log frequency of observed events for the two ICK case studies.

data may contain noise. As such, not all details of the statistical model may be used in prediction, even under stationarity. Clearly, a decision is required about an ''appropriate'' degree of smoothing given the data uncertainties. In practice, this involves selecting a limited number of orthogonal covariates in ICK that lead to similarly good performance under dependent and independent validation (section 4). For a square symmetric matrix, **W**, the SVD is equivalent to diagonalization, or to solution of the eigenvalue problem, where the singular values of **W** are the positive square roots of the eigenvalues (Hansen 1987). The threshold below which singular values are zeroed (the so-called singularity threshold) can, therefore, be expressed as a proportion of the total variance retained in the solution.

As indicated earlier, the ICK technique is not explicitly constrained to produce valid probabilities. Furthermore, the probabilities are estimated at only a limited number of thresholds of the observed variable, yet the full ccdf is required for many practical applications, such as uncertainty propagation analysis. To provide a smooth approximation of the full ccdf, the conditional probabilities are approximated locally with a piecewise quadratic function of x, subject to the requirement of valid probabilities (He and Ng 1999). This ''smoothing spline'' is fitted to minimize the mean absolute errors of the estimates, which is a robust constrained smoother (He and Ng 1999).

4. Case studies and verification results

a. Experimental design

To evaluate the performance of the ICK technique, verification was conducted for forecasts of several hydrologic and hydrometeorological variables. The verification samples comprised paired forecasts and observations at multiple forecast lead times. The statistical parameters of ICK were estimated from the paired sample data. The estimated probabilities were then corrected and smoothed with a quadratic spline. The bias-corrected forecasts were compared to the ''raw'' forecasts using several attributes of forecast quality (discussed later). Two scenarios were considered, namely, 1) ''dependent validation,'' whereby all data were used to estimate the statistical parameters of ICK; and 2) ''independent validation,'' whereby a subsample was used for estimation and the remainder used for validation. For example, a 90/10 split comprised 10

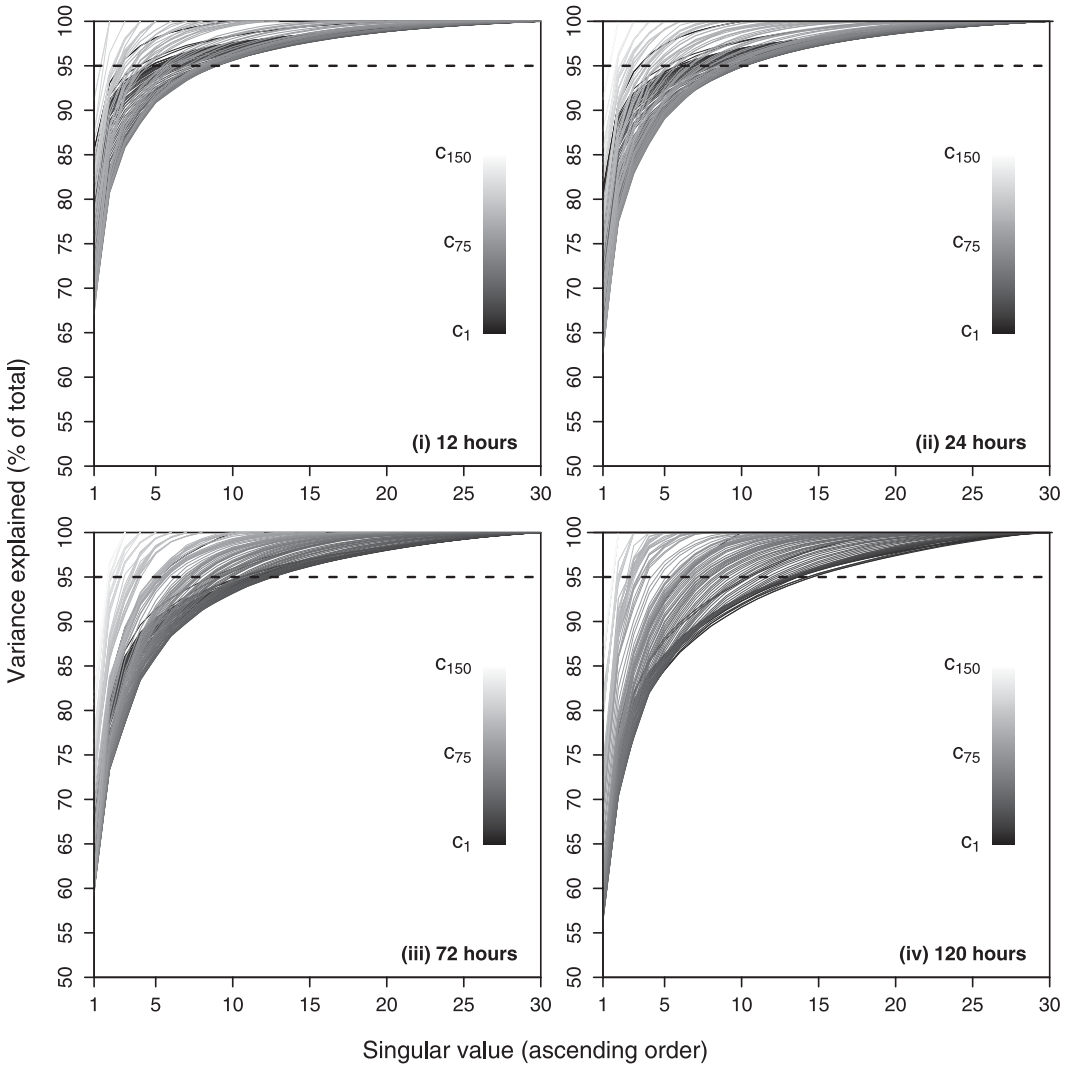

FIG. 2. Singular values by thresholds of X for GEFS precipitation forecasts.

equal periods, of which 9 were used for estimation and 1 for validation. This was repeated until each of the nine periods had been used for validation. While dependent validation is not representative of the actual forecasting process, it provides an upper bound for the performance of ICK given the sample data, as well as a reference point for quantifying sampling uncertainty and for determining an appropriate degree of regularization of the ICK problem (see section 3). In contrast, independent validation mimics the operational forecasting process, but it employs only a subset of the available data for estimation. Unless otherwise stated, the results are shown for forecasts issued under independent validation.

Key attributes of ensemble forecast quality include the reliability of the forecast probabilities and the ability of the forecasts to discriminate between different observed conditions (Jolliffe and Stephenson 2003). Forecast reliability was evaluated with the reliability diagram (Hsu and Murphy 1986), and discrimination was evaluated with the relative operating characteristic (ROC; Green and Swets 1966; Mason and Graham 2002) and, specifically, the area under the ROC curve (AUC; Fawcett 2006). The mean CRPS was also computed, as ICK explicitly minimizes the CRPS on a threshold-by-threshold basis. The CRPS is a popular verification metric and is ''strictly proper,'' meaning it cannot be hedged (Bröcker and Smith 2007). In addition, the average of the forecast ccdfs (i.e., the average probability at each value of $X = x$; Gneiting et al. 2007) was checked for unbiasedness against the observed climatology, as ICK is explicitly formulated to be unbiased in the unconditional sense.

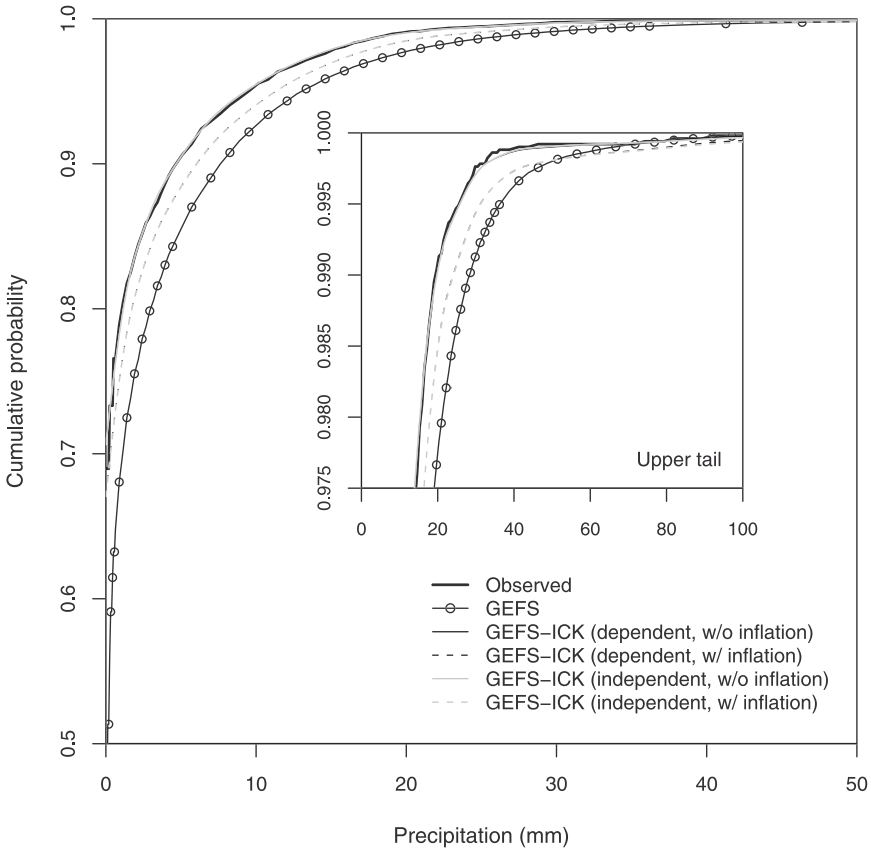

FIG. 3. Climatological cdfs for GEFS precipitation forecasts.

b. Precipitation forecasts from the NCEP GEFS

Operational forecasts of precipitation were obtained from the GEFS of the NCEP. The archive contains a continuous record of 12-hourly forecasts from 1 January 2000 to 15 August 2005, with lead times ranging from 12 to 120 h in 12-hourly increments. Each forecast comprises 10 ensemble members, and each member represents an equally likely prediction of the total precipitation within the 12-h period. More recent forecasts were available but not considered, as the data assimilation scheme used in the operational GEFS was upgraded on 16 August 2005. This led to significant changes in the predictive error and uncertainty of the precipitation forecasts. Other changes in the GEFS between 2000 and 2005, which included a mask rescaling on 23 April 2003, were not reflected in

TABLE 1. Observed and forecast PoP for GEFS.

			Dependent validation		Independent validation	
Lead time (h)	Observed	GEFS	ICK	ICK$^+$	ICK	ICK$^+$
12	0.31	0.48	0.30	0.33	0.29	0.33
24	0.31	0.49	0.30	0.33	0.30	0.33
36	0.31	0.49	0.30	0.33	0.29	0.33
48	0.31	0.49	0.30	0.33	0.30	0.33
60	0.31	0.49	0.30	0.33	0.29	0.33
72	0.31	0.50	0.30	0.33	0.30	0.33
84	0.31	0.50	0.30	0.33	0.29	0.33
96	0.31	0.53	0.30	0.33	0.30	0.33
108	0.31	0.55	0.30	0.33	0.30	0.33
120	0.31	0.55	0.30	0.32	0.30	0.32

TABLE 2. Mean CRPS and associated skill (%) of the GEFS precipitation forecasts.

Lead time (h)	GEFS	Dependent validation		Independent validation	
		ICK	ICK$^+$	ICK	ICK$^+$
12	1.24	0.81 (35)	0.88 (29)	0.84 (32)	0.89 (28)
24	1.27	0.86 (32)	0.93 (27)	0.89 (30)	0.97 (24)
36	1.24	0.89 (28)	0.97 (22)	0.91 (27)	0.99 (20)
48	1.19	0.91 (24)	0.99 (17)	0.94 (21)	1.02 (14)
60	1.17	0.97 (17)	1.04 (11)	0.97 (17)	1.07 (9)
72	1.22	1.02 (16)	1.12 (8)	1.02 (16)	1.14 (7)
84	1.27	1.07 (16)	1.17 (8)	1.09 (14)	1.17 (8)
96	1.35	1.12 (17)	1.19 (12)	1.14 (16)	1.24 (8)
108	1.42	1.17 (18)	1.27 (11)	1.19 (16)	1.27 (11)
120	1.47	1.19 (19)	1.3 (12)	1.22 (17)	1.32 (10)

the forecast error statistics in the 12–120-h lead periods and were, therefore, deemed unimportant.

The ICK was performed for several basins in the service area of the Middle Atlantic River Forecast Center (MARFC). Basin-averaged precipitation was used as the observed variable and was provided by MARFC (Larson 1976). The results were similar for each basin considered and are shown for a single basin, namely, Huntingdon, PA (RFC point HUNP1). Figure 1 shows the log frequency of different observed precipitation amounts over the GEFS forecast period. The ICK was performed separately for each forecast lead time. Independent validation was conducted with a 95/5 split sample, rotated 20 times. For computational efficiency, only two auxiliary thresholds were included alongside the main indicator threshold in each estimation. These corresponded to 0.8 and 1.4 times the main threshold. In practice, this was found to capture most of the additional variance explained by ICK when compared to indicator kriging with no auxiliary thresholds (Journel and Huijbregts 1978). The main thresholds were fixed at increasing quantiles of the climatological distribution, with more thresholds at light precipitation than heavy precipitation to cover the high probability densities there, and 150 thresholds in total. The thresholds were chosen by visually inspecting the bias-corrected probabilities for noise, and by checking

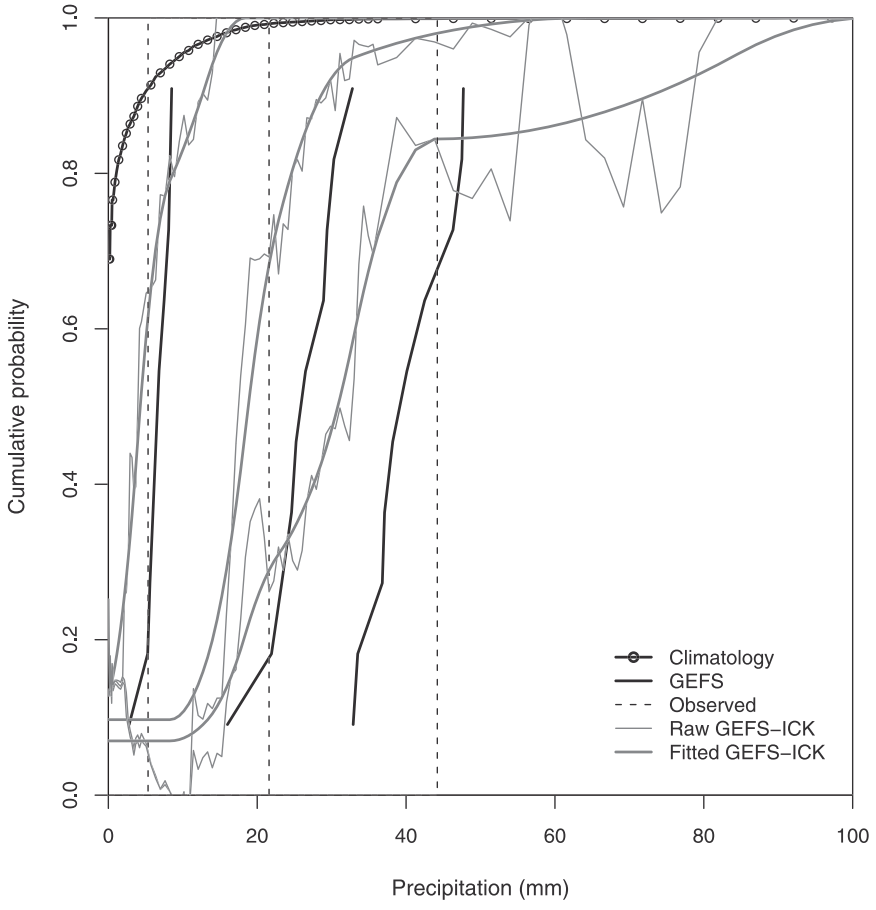

FIG. 4. Three example cdfs for GEFS precipitation forecasts.

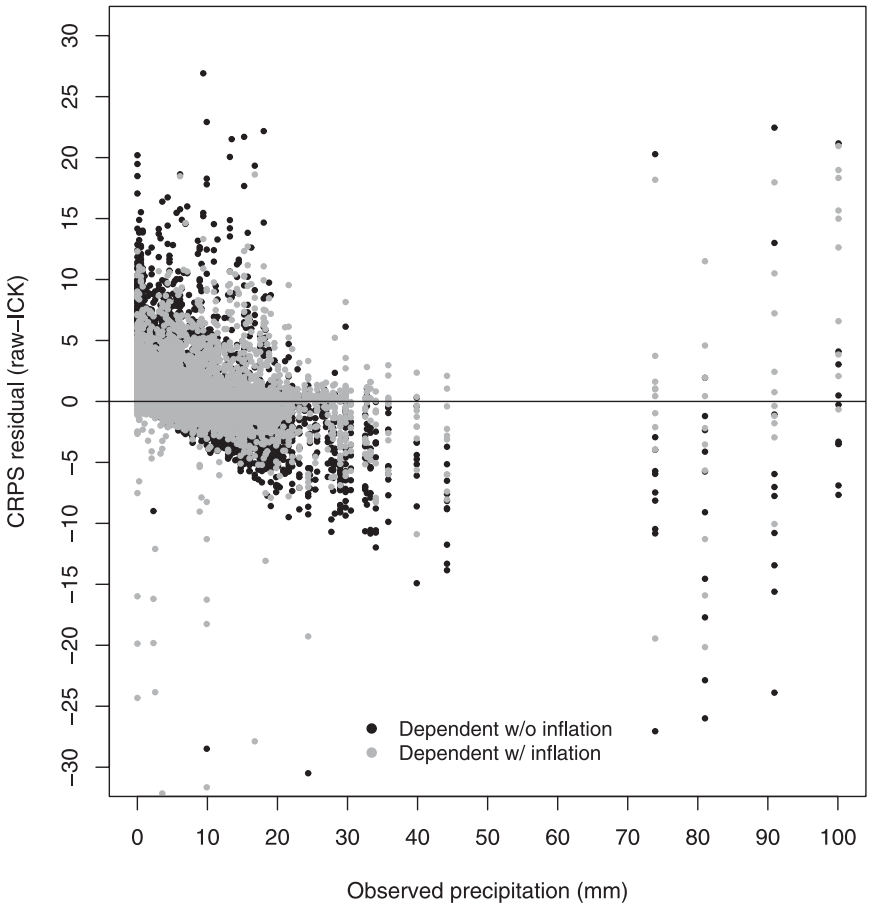

FIG. 5. CRPS residuals (raw-ICK) for precipitation forecasts at all lead times.

for stability of the mean CRPS and other verification statistics. The linear system was solved through SVD, with a singularity threshold of 95%. In other words, the smallest singular values, which together explained 5% or less of the total indicator variance, were suppressed. Figure 2 plots the contribution of each singular value to the total indicator variance explained at each threshold. The singular values are arranged in descending order of contribution. The results are shown for lead hours 12, 24, 72, and 120. As indicated in Fig. 2, most of the indicator variance is explained by only a few orthogonal components, and the value of adding further components declines exponentially (or even faster at high thresholds). In general, when measured by the total indicator variance explained, more information is captured by fewer orthogonal components at higher precipitation thresholds (lighter shades) and shorter lead times. In other words, the information content of the ensemble forecasts is spread across more members at lower precipitation thresholds and longer lead times. The singularity threshold was chosen by visually inspecting the bias-corrected probabilities for noise, and by checking for similar values of the mean CRPS and other verification statistics under dependent and independent validation. A constrained-quadratic spline was used to smooth the ICK estimates and to correct for any order-relation violations (see section 3).

Figure 3 shows the observed probabilities against the corresponding average probabilities of the raw and bias-corrected forecasts. For each indicator threshold of the observed variable, the forecast probabilities were averaged across all forecasts and lead times (see Gneiting et al. 2007). The resulting ''forecast climatology'' was compared to the observed climatology for unbiasedness. The y intercept in Fig. 3 corresponds to the probability of precipitation (PoP) and shows the ability of the ICK forecasts to capture the mixed behavior in precipitation occurrence versus amount (in an unconditional sense). Table 1 lists the climatological PoP and mean forecast PoP for the uncorrected and bias-corrected forecasts at each lead time. The scenario denoted ICK$^+$ is based on ICK with correlation inflation, which is explained later. By design, the unconditional forecast probabilities are

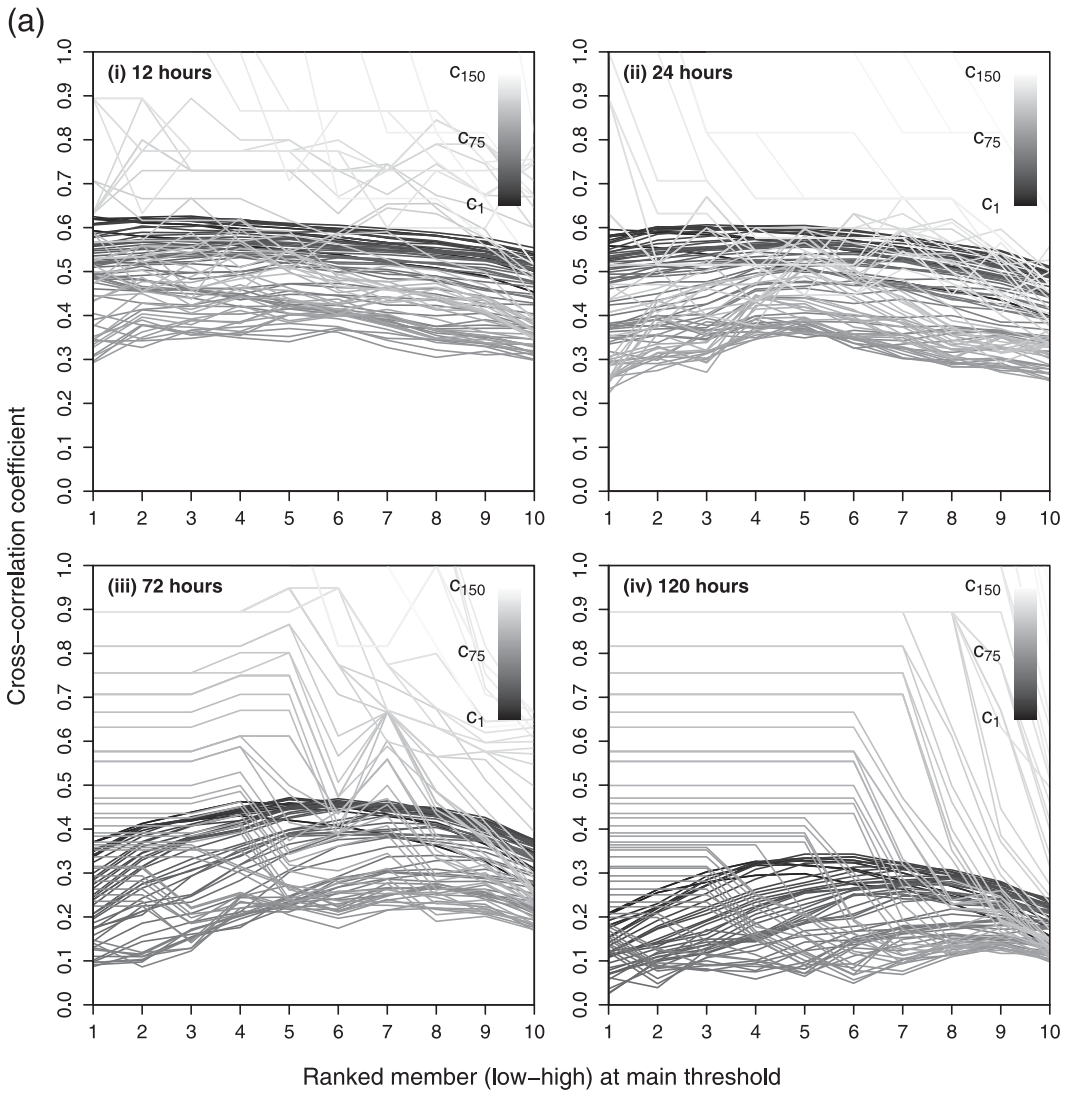

FIG. 6. (a) Cross-correlation coefficients by ranked ensemble member for GEFS. (b) Cross-correlation coefficients by precipitation amount for GEFS.

unbiased following ICK [see Eq. (5)], and this is confirmed in Fig. 3. However, correlation inflation no longer guarantees the conditional adjustment to be unbiased (Fig. 3), which is evidenced by an increase in the unconditional biases for ICK$^+$. Nevertheless, the ICK$^+$ climatology remains significantly less biased than the raw GEFS forecast climatology. This is also apparent in the PoP forecasts (Table 2), which are strongly biased in the raw GEFS, but are effectively unbiased following both ICK and ICK$^+$.

Examples of the ICK estimates and fitted ccdfs are shown in Fig. 4 for different precipitation amounts, together with the observed cdfs (Heaviside function) and the observed climatology. As indicated in Fig. 4, the raw probabilities from ICK are subject to increased sampling uncertainty with increasing precipitation amount, which reflects the smaller sample size at high precipitation thresholds. Table 2 shows the mean CRPS for the raw forecasts and ICK forecasts at each lead time, together with the continuous ranked probability skill score (CRPSS) of the ICK forecasts. Here, the CRPSS quantifies the percentage gain in mean CRPS ($\overline{\text{CRPS}}$) following bias correction, that is, $\text{CRPSS} = 100[1.0 - (\overline{\text{CRPS}}_{\text{ICK}}/\overline{\text{CRPS}}_{\text{GEFS}})]$. Figure 5 plots the difference in CRPS between the raw forecasts and ICK forecasts (raw-ICK) for each forecast produced under dependent validation. The differences are plotted against observed precipitation amount, which shows the conditional errors in the ICK forecasts. Actual differences between $\text{CRPS}_{\text{GEFS}}$ and CRPS_{ICK} were preferred over CRPSS, as many of

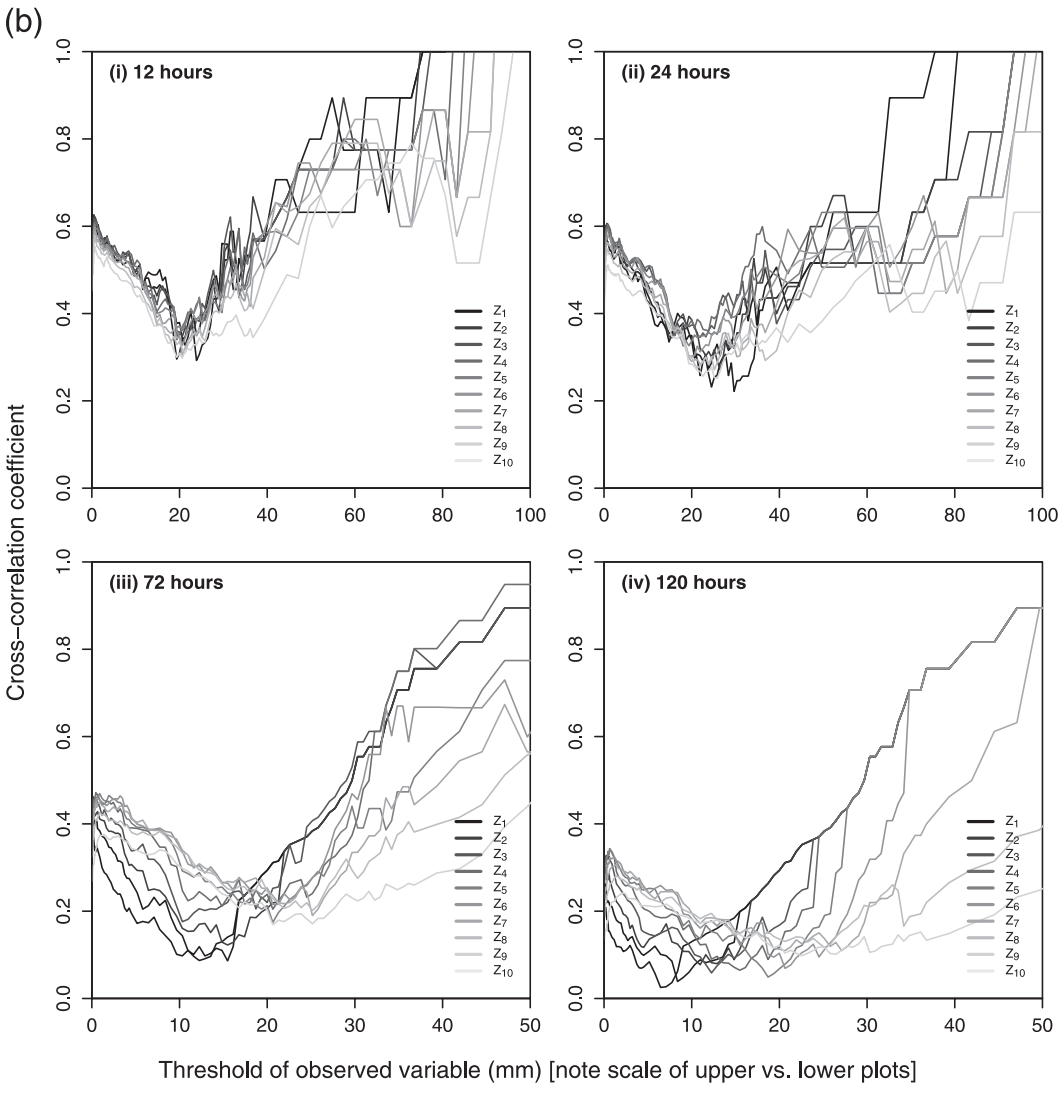

FIG. 6. (*Continued*)

the forecasts had zero CRPS, that is, undefined CRPSS. The results are shown for all lead times, so each observation is associated with 10 separate forecasts. A positive value indicates a gain in performance over the raw forecasts and a negative value indicates a loss of performance.

The CRPS provides a critical check on the performance of ICK, as the solution to ICK effectively minimizes the CRPS on a per quantile basis (see earlier). As shown in Table 1, the bias-corrected GEFS forecasts were significantly more skillful than the raw GEFS forecasts in terms of mean CRPS, with an overall improvement of 17%–32% following ICK. However, this improvement declined markedly with increasing observed precipitation due to a conditional bias in the ICK forecasts (Fig. 5). This bias stems from an ''attenuation effect,'' which is well known in least squares estimation (e.g., Fuller 1987; Draper and Smith 1998). In single-valued forecasting, it is manifest as a trade-off between the conditional bias and the mean square error of the estimate [see Ciach et al. (2000) for a precipitation example]. As noted by Fuller (1987), the attenuation effect is caused by predictors that are measured with substantial error (i.e., errors in variables). This results in an underestimation of the cross correlations between the predictors and response, and it is exaggerated by the least squares solution. However, this problem is not confined to least squares estimation; for example, similar behavior is observed in quantile regression (Koenker 2005). To manage this trade-off explicitly, the cross correlations can be

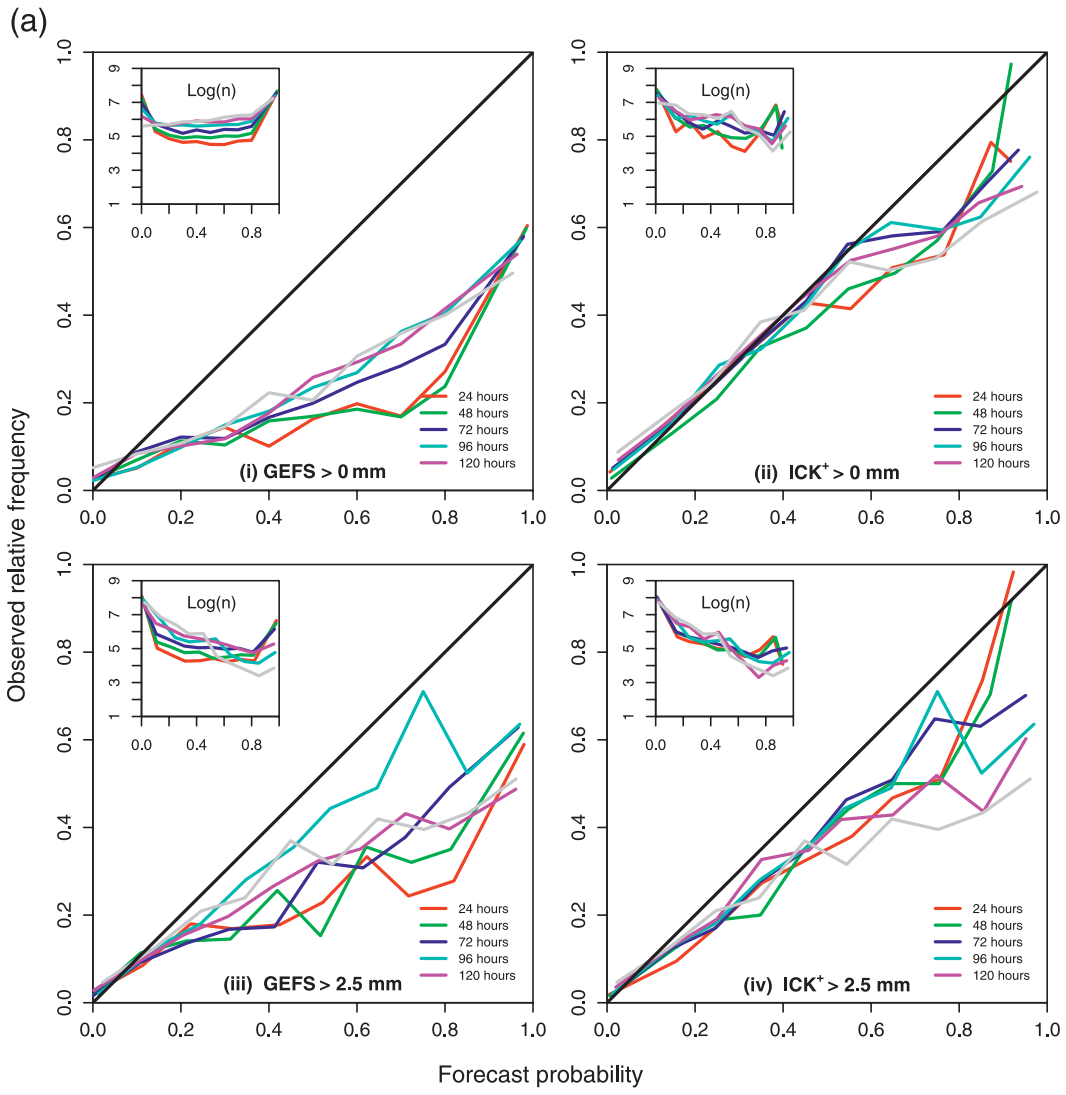

FIG. 7. (a) Reliability diagrams for raw GEFS and ICK$^+$ forecasts (PoP/light). (b) Reliability diagrams for raw GEFS and ICK forecasts (medium precipitation). (c) Spread-bias diagrams for raw GEFS and ICK forecasts.

optimized explicitly. However, this is difficult in practice, because ICK comprises several (u) estimators for which the indicator cross correlations must be optimized. Here, we propose a simpler approach, whereby the cross correlations are increased uniformly. This involves transforming the vector of cross covariances in Eq. (9) to cross correlations, inflating each correlation coefficient by a positive constant ρ^+ while capping the inflated correlation at 1, and back-transforming to covariances. The ρ^+ is chosen by trial and error. In this example, it was chosen through a visual inspection of the ccdfs and associated verification statistics, with particular attention to the CRPSS and bias in the conditional mean as ρ^+ increased. The ICK forecasts were regenerated with uniform correlation inflation, ρ^+, of 0.2 (denoted as ICK$^+$).

Figure 5 illustrates the importance of correlation inflation when bias-correcting large precipitation amounts. For example, only 8 of 40 forecasts whose paired observations exceeded 60 mm of precipitation in 12 h gained in skill following ICK, compared to 24 of 40 following ICK$^+$. However, these 40 forecasts comprised only 4 unique events, with forecasts of each event at 10 different lead times (i.e., 12–120 h). In practice, many more samples would be required to properly evaluate ICK for extreme precipitation. Increasing the sample size was not possible for the operational GEFS, and it is difficult in general without

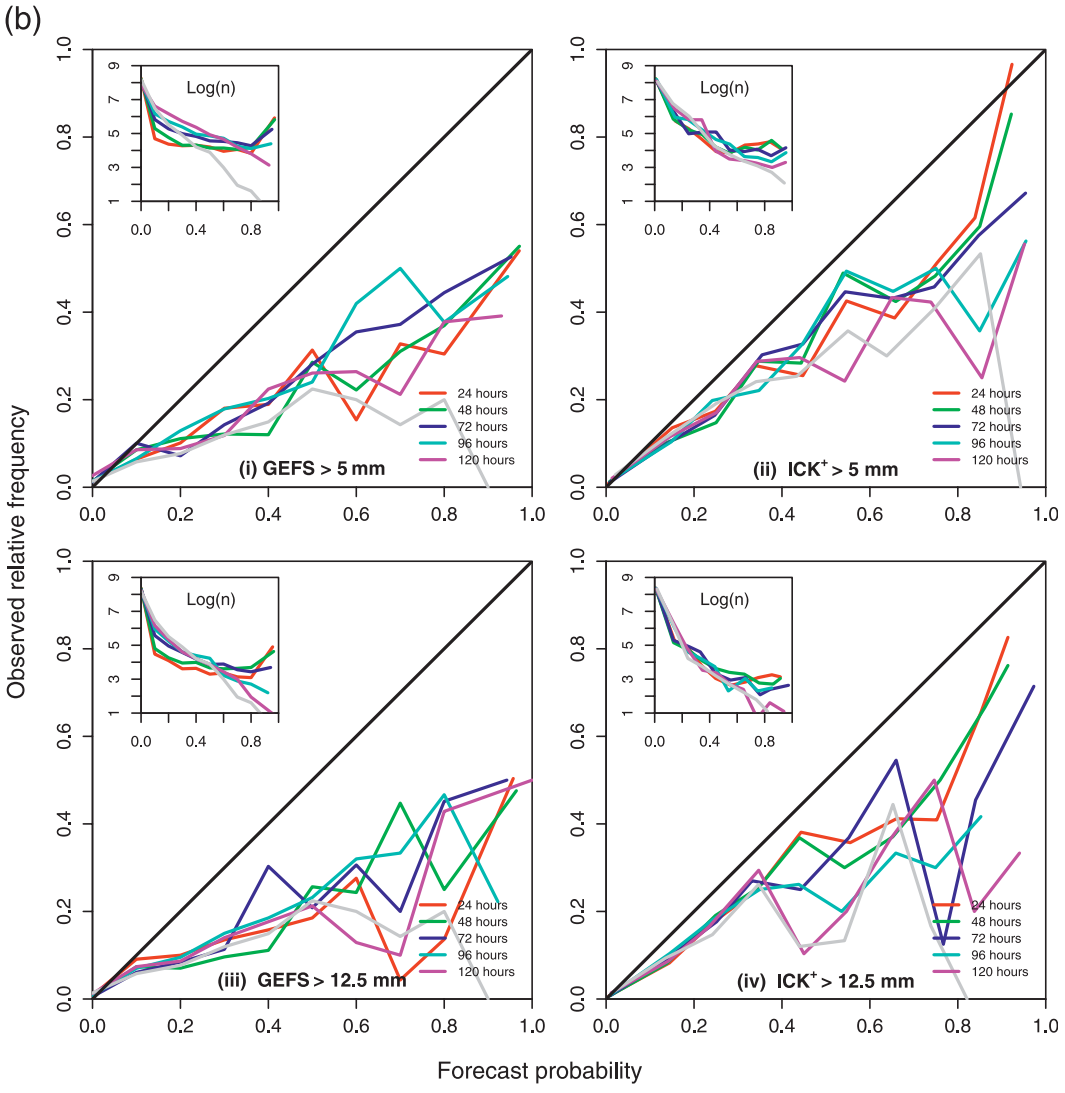

FIG. 7. (*Continued*)

long-term hindcasting for each model upgrade. In keeping with the trade-off between minimum CRPS and small conditional bias, the overall CRPSS fell by 4%–7% following correlation inflation (Table 2).

Figure 6a shows the cross correlations between the observed indicator variables and the indicator variables of the forecast ensemble members under dependent validation. The y axis shows the cross-correlation coefficient, and the x axis shows the ranked ensemble member at the main indicator threshold. The cross-correlation coefficients are shown for lead hours 12, 24, 72, and 120. Each line represents one of the u thresholds of X at which an estimator is formed. They are shaded from black to light gray, with increasing threshold value. Figure 6b shows the same information, only presented by threshold value for each ensemble member, $\{Z_1, \ldots, Z_m\}$, where $Z_{i-1} \leq Z_i$, $i = 2, \ldots, m$. Thus, both figures show the predictive performance of the GEFS forecasts in terms of cross correlations; however, Fig. 6a shows this predictive performance by ranked ensemble member across thresholds, whereas Fig. 6b shows the performance as a function of increasing threshold across ensemble members.

The convexity of the curves in Fig. 6b stems from the quadratic dependence of the indicator variance on the indicator mean, as well as the convergence of the marginal and joint probabilities to 1 as the indicator threshold increases. Clearly, the choice of plotting position will also impact the limiting behavior of the cross correlations (see section 3). As shown in both Figs. 6a and 6b, the indicator

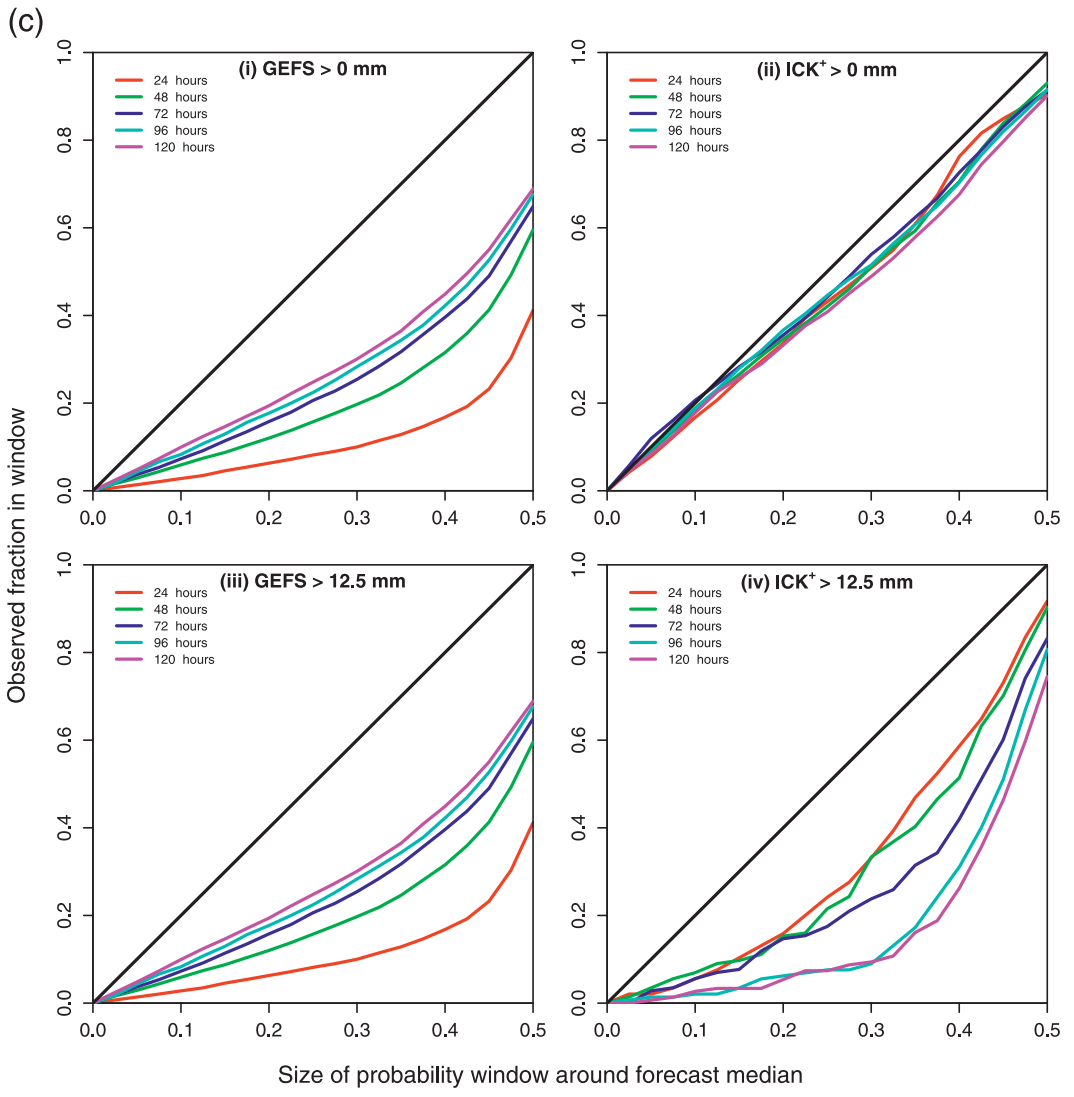

FIG. 7. (*Continued*)

cross correlations are subject to increased sampling uncertainty at the highest thresholds of the observed variable. While the predictive performance of the GEFS forecasts is reasonably consistent between the ranked ensemble members at short lead times, the intermember variability increases at long lead times, particularly at low and high threshold values (Fig. 6b). These patterns are consistent with the SVD spectra shown in Fig. 2. As shown in Fig. 6a, the indicator cross correlations are greatest for the small ensemble members and lowest thresholds at lead hours 12 and 24. This alludes to a concentration of skill in the PoP and light-precipitation forecasts at short lead times.

Figure 7a plots the conditional reliability of the raw forecasts and ICK$^+$ forecasts for precipitation amounts exceeding 0.0 (PoP) and 2.5 mm at lead hours 12, 24, 48, 72, 96, and 120. Figure 7b shows the corresponding results for precipitation amounts exceeding 5 and 12.5 mm. The sampling uncertainties were too large to evaluate conditional reliability at thresholds exceeding 12.5 mm. Figure 7c shows another measure of statistical reliability, namely, the "spread-bias plot." It is similar to the cumulative rank histogram (Anderson 1996; Hamill 1997; Talagrand 1997) or the probability integral transform histogram (Gneiting et al. 2005) and shows the fraction of observations that fall within an interval of fixed probability around the forecast median. Results are shown for observed precipitation amounts exceeding 0 and 12.5 mm at lead hours 12, 24, 48, 72, 96, and 120. Table 3 shows the area under the

TABLE 3. ROC areas for GEFS and (ICK) under independent validation ($\rho^+ = 0.2$).

Lead time (h)	Event (exceedance of precipitation amount in mm)				
	0.0	2.5	5	12.5	25
12	0.87 (0.89)	0.88 (0.91)	0.87 (0.93)	0.84 (0.94)	0.78 (0.91)
24	0.86 (0.88)	0.89 (0.91)	0.87 (0.92)	0.87 (0.94)	0.81 (0.92)
36	0.86 (0.87)	0.89 (0.90)	0.88 (0.92)	0.83 (0.93)	0.79 (0.86)
48	0.84 (0.85)	0.88 (0.89)	0.85 (0.91)	0.86 (0.93)	0.82 (0.89)
60	0.83 (0.84)	0.87 (0.88)	0.85 (0.89)	0.83 (0.92)	0.77 (0.92)
72	0.81 (0.83)	0.85 (0.87)	0.84 (0.88)	0.80 (0.89)	0.74 (0.90)
84	0.80 (0.81)	0.82 (0.84)	0.82 (0.86)	0.81 (0.86)	0.77 (0.86)
96	0.77 (0.78)	0.79 (0.81)	0.79 (0.83)	0.75 (0.84)	0.72 (0.75)
108	0.75 (0.76)	0.77 (0.79)	0.77 (0.82)	0.71 (0.82)	0.61 (0.75)
120	0.73 (0.74)	0.75 (0.76)	0.73 (0.78)	0.72 (0.80)	0.64 (0.73)

ROC curve (AUC) for precipitation amounts exceeding 0.0, 2.5, 5.0, 12.5, 20, and 25 mm. In general, the ICK forecasts were both more reliable and more discriminatory than the raw GEFS forecasts. As indicated in Fig. 7c, the lack of spread in the GEFS forecasts was largely corrected by ICK^+ for positive precipitation amounts. It was also substantially improved for precipitation amounts exceeding 12.5 mm at lead hours 12 and 24. For example, only 40% and 60% of the raw GEFS forecasts captured their verifying observation at lead hours 12 and 24, respectively (i.e., the y axis intercept at $x = 0.5$), while ~90% of the observations were captured following ICK^+. However, there was a loss of conditional reliability following ICK^+ at lead hours 96 and 120 for precipitation amounts exceeding 12.5 mm, where the probabilities *within* the extreme prediction intervals were insufficiently spread (the total fraction of captured observations was slightly improved). This reflects the difficulties associated with statistical postprocessing when the forecasts have limited skill, as evidenced by the reduced indicator cross correlations at longer lead times (Fig. 6). Overall, the gains from ICK declined systematically with increasing precipitation amount. This reflects a combination of sampling uncertainty and the inability of correlation inflation to fully remove the conditional biases in ICK. Future work will consider refinements to the simple correlation inflation adopted here (such as nonuniform correlation inflation), and the sampling uncertainties will be examined for larger datasets, such as the GFS reforecast dataset (Hamill et al. 2006). Given a larger sample size, a seasonal or regime-dependent prior distribution might also be considered. For example, the observed climatology may be conditioned on the ensemble mean of the uncorrected forecasts. Indeed, it is not surprising that the ICK forecasts are significantly more reliable for light precipitation when using the unconditional climatology as the prior distribution.

c. *Streamflow forecasts from the NWS ESP system*

Mean daily inflows were hindcast for a 23-yr period from 1 January 1979 to 31 December 2002 at the North Fork Dam, California (NWS basin NFDC1, U.S. Geological Survey (USGS) station 11427000). The hindcasts were produced with the NWS ESP system, which implements the NWSRFS in an ensemble framework (Schaake et al. 2007). The NWSRFS was forced with ensemble hindcasts of temperature and precipitation from the frozen GEFS (Toth et al. 1997; Hamill et al. 2006; Schaake et al. 2007; Wei et al. 2008). Each streamflow hindcast comprised 54 ensemble members. The hindcasts were aggregated from a 6-hourly time step to daily averages for comparison with the observed flows, which were only available as daily averages. The observed flows are based on stage observations, which were converted to flows using a measured stage-discharge relation (Kennedy 1983). Figure 1 shows the climatological flow regime at NFDC1.

Bias-corrected ESP forecasts were generated for lead times of 24–120 h in 24-hourly increments, using a separate ICK model for each lead time, and independent validation was conducted with a 90/10 split sample. Two auxiliary thresholds were included alongside the main indicator threshold in each estimator. These corresponded to 0.8 and 1.4 times the main threshold. The main thresholds were fixed at increasing quantiles of the climatological distribution, with more thresholds at low flows than high flows, and 200 thresholds in total. The thresholds were chosen by visually inspecting the bias-corrected probabilities for noise, and by checking for stability of the mean CRPS and other verification statistics. The ICK was solved through SVD with a singularity threshold of 95%. Figure 8 shows the

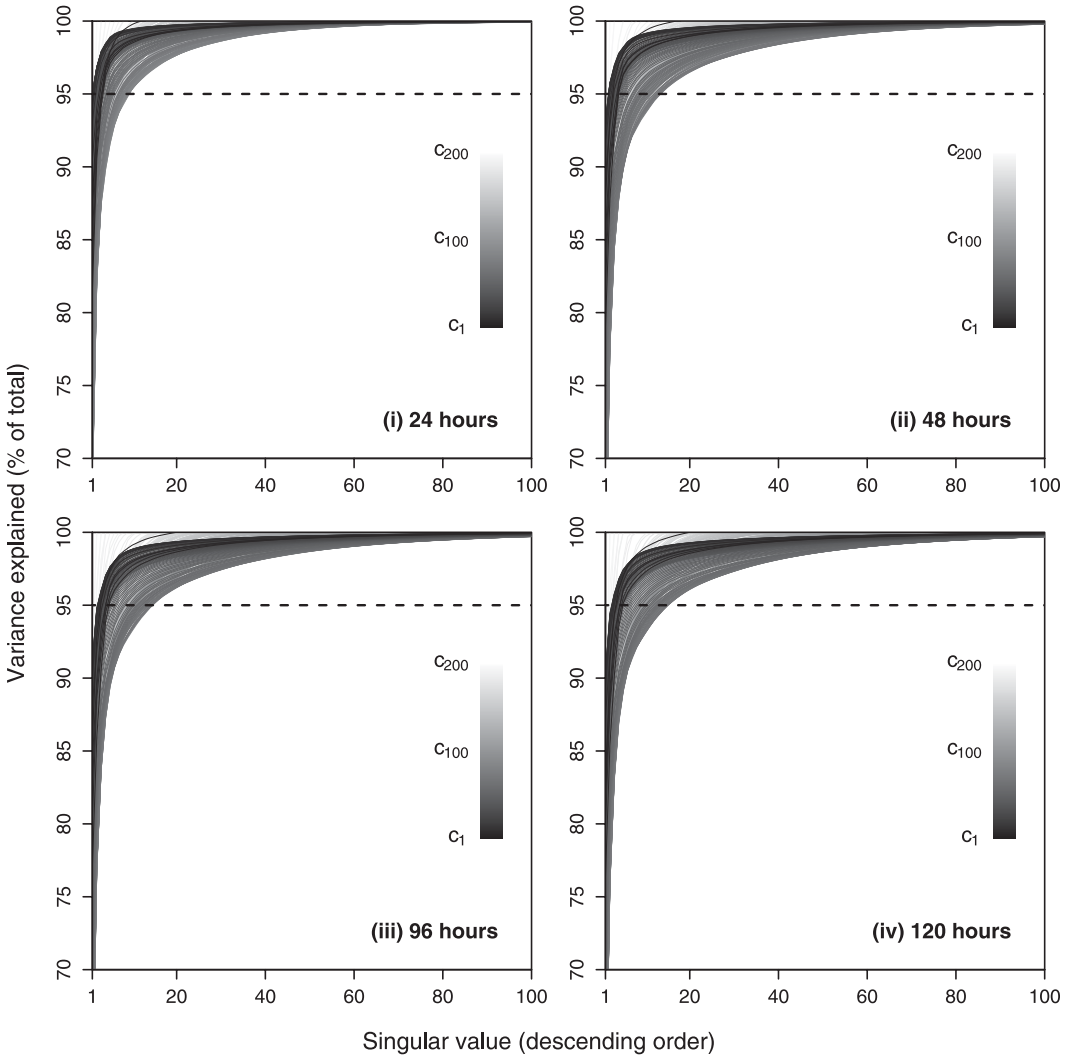

FIG. 8. Singular values by thresholds of X for ESP streamflow forecasts.

proportion of the total indicator variance explained by each singular value. The singular values are arranged in descending order of contribution and by increasing indicator threshold (darker shades). The results are shown for lead hours 24, 48, 96, and 120. As before, most of the total indicator variance is captured by a small number of orthogonal components. More orthogonal variables are required to explain a given proportion of the total indicator variance at longer lead times and at ''medium flows.'' This probably reflects the divergence of ensemble members with increasing lead time as they undergo more variable dynamics from the hydrologic model, particularly in the predominantly occurring low to medium flows. Again, the singularity threshold was chosen by comparing the performance of ICK under dependent and independent validation, and the discrete ICK estimates were corrected and smoothed with a constrained quadratic spline. However, no correlation inflation was required, as the conditional biases were much smaller in the ESP forecasts than the GEFS forecasts.

Examples of the ICK estimates and fitted ccdfs are shown in Fig. 9 for different streamflow amounts, together with the observed cdfs (Heaviside function) and the climatological prior. As with the GEFS precipitation forecasts, the raw probabilities from ICK become increasingly noisy at high flows. Figure 10 shows the average cdfs of the raw and bias-corrected forecasts against the climatological cdf. Again, the ICK estimates are unbiased in a climatological sense, because the observed climatology forms the prior distribution in ICK. Table 4 shows the mean CRPS of the raw and bias-corrected forecasts at each lead time under dependent and independent

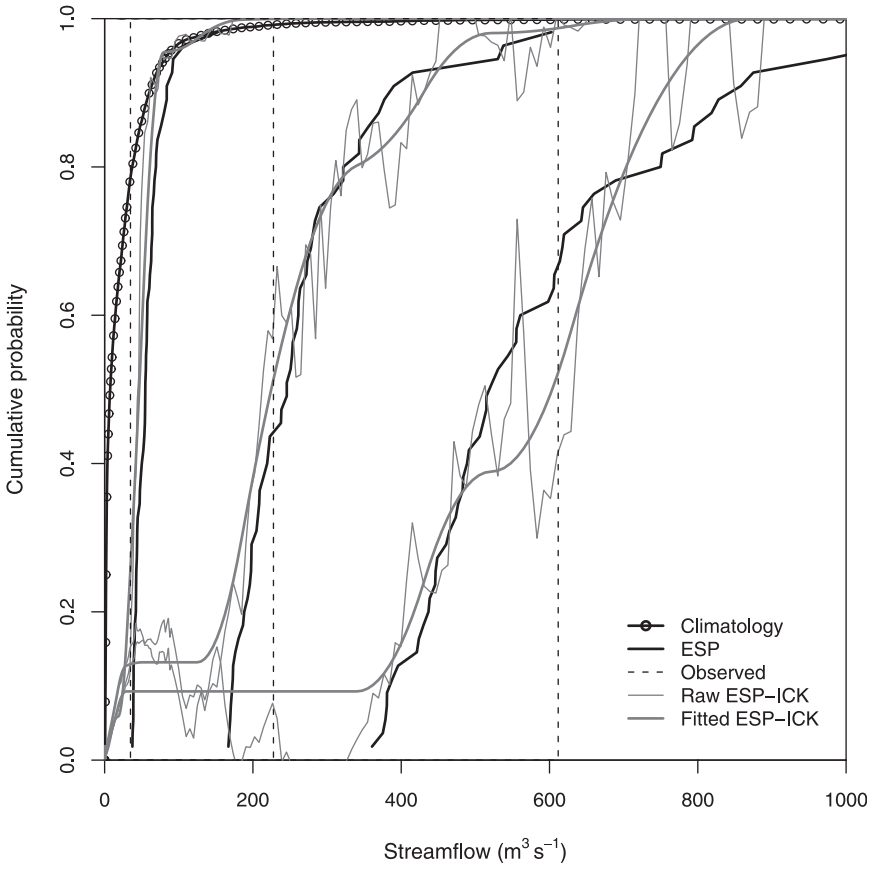

FIG. 9. Three example cdfs for ESP streamflow forecasts.

validation. Figure 11 shows the difference in CRPS between the raw and bias-corrected forecasts as a function of observed streamflow. The ICK forecasts were significantly more skillful than the raw ESP forecasts in terms of CRPS, although the margin for improvement was lessened by their initial high quality. For example, the ICK forecasts were 9%–21% more skillful than the raw ESP forecasts under dependent validation and 5%–16% more skillful under independent validation (Table 4). The conditional biases that affected the ICK precipitation forecasts were not apparent in the ICK streamflow forecasts, as the raw forecast means were both relatively (conditionally) unbiased and strongly correlated with the observations. Indeed, the ICK forecasts were consistently more skillful than the ESP forecasts at the highest observed flows. For example, 20 of the 25 forecasts whose paired observations exceeded 800 $m^3 s^{-1}$, which corresponds to a climatological exceedance probability of 0.001, gained in CRPSS following ICK (Fig. 10).

Figure 12a shows the reliability of the forecasts at lead days 1–5. The results are shown for climatological exceedance probabilities of 0.5 (10 $m^3 s^{-1}$) and 0.75 (33 $m^3 s^{-1}$). Figure 12b shows the corresponding results for exceedance probabilities of 0.9 (63 $m^3 s^{-1}$) and 0.99 (210 $m^3 s^{-1}$). The forecast reliabilities were significantly improved for streamflows with a climatological exceedance probability of 0.9 and smaller. For streamflows with an exceedance probability of 0.99, the ESP forecasts are already very reliable. This is because the hydrologic models are well calibrated for moderately high flows, and the associated forcing ensembles are generally reliable. The noisiness of the reliability curve for streamflows exceeding a climatological probability of 0.99 reflects the small number of events in the flow archive that exceed 210 $m^3 s^{-1}$ (Fig. 1) and the correspondingly high sampling uncertainties for these events (Fig. 9). For most flow thresholds, the raw ESP forecasts are underspread or ''overconfident,'' particularly at short lead times (Figs. 12a and 12b). This is understandable because the current generation of ESP ignores uncertainty in the initial states, parameters, and structure of the hydrologic model (Seo et al. 2006). The lack of uncertainty in the initial states is also reflected in the relative sharpness of the forecast probabilities at the 24-h lead time (Figs. 12a and 12b). Table 5 shows the areas

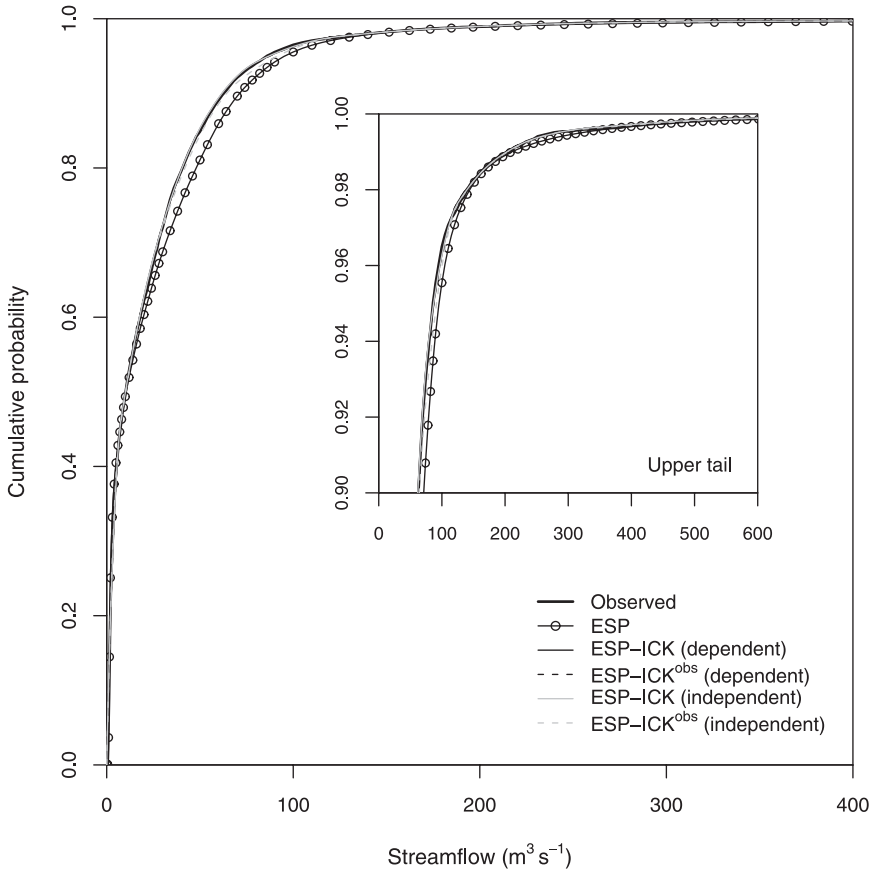

FIG. 10. Climatological cdfs for NFDC1 streamflow.

under the ROC curve for the raw and bias-corrected forecasts. Overall, the reliability of the ESP forecasts was significantly improved following ICK, and the discrimination was improved slightly. However, large improvements in discrimination should not be expected, as statistical postprocessing is concerned with the *calibration*-refinement factorization of $F(x, y)$, not the likelihood–base-rate factorization (Murphy and Winkler 1987). The largest gains in reliability occurred at ''small sized'' and ''medium sized'' flows (Figs. 12a and 12b), with sampling uncertainty preventing analysis at very high flows.

5. Conclusions

Ensemble forecasts of hydrometeorological and hydrologic variables typically contain biases in the mean, spread, and higher moments. These biases can be removed if the conditional cumulative distribution function (ccdf) of the observed variable, given the ensemble forecast, can be modeled reliably. When the shape of this probability distribution is unknown, a nonparametric estimate is desired. In this paper, the ccdf is modeled with a nonparametric technique that is analogous to indicator cokriging (ICK) in geostatistics (Journel and Huijbregts 1978; Isaaks and Strivastava 1989; Cressie 1993). Specifically, the probability of exceeding a discrete threshold of the observed variable, such as flood stage, is modeled as a linear function of the indicator variables of the forecast ensemble members. In terms of minimum error variance, the ICK estimator is the Bayesian optimal linear

TABLE 4. Mean CRPS and associated skill (%) of the ESP streamflow forecasts.

Lead time (h)	ESP	Dependent validation		Independent validation	
		ICK	Skill (%)	ICK	Skill (%)
24	6.77	5.32	21	5.66	16
48	6.9	6.03	13	6.37	8
72	6.96	6.35	9	6.58	6
96	7.41	6.71	10	7.04	5
120	7.83	7.11	9	7.42	5

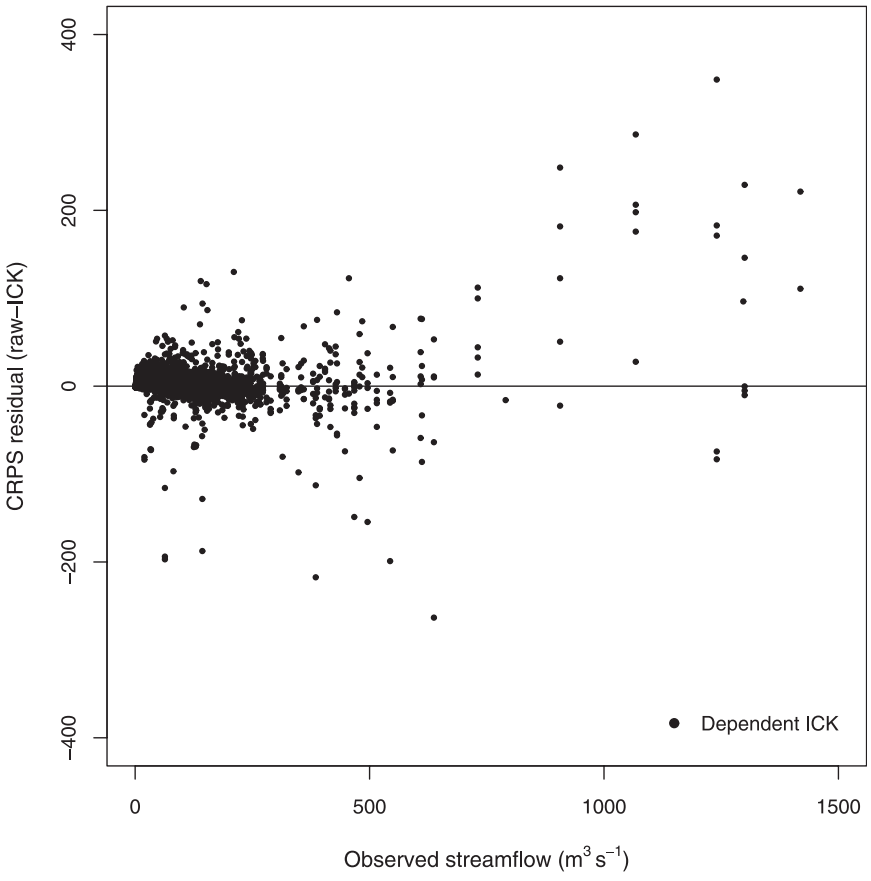

FIG. 11. CRPS residuals (raw-ICK) for streamflow forecasts at all lead times.

estimator of the conditional observed probability at the chosen threshold. In the examples presented, ICK was solved through an orthogonal decomposition of the covariances between the indicator variables of the forecast ensemble members. Because of the strength of these covariances, an orthogonal decomposition was found to significantly reduce the dimensionality of the estimation problem. This regularized form of ICK is analogous to cokriging with the first few principal components of the ensemble forecast (see Deutsch and Journel 1992). By developing an estimator for several thresholds, ICK provides a discrete approximation of the full ccdf. A smooth function is then interpolated through the estimated probabilities to provide a full ccdf, and a quadratic smoothing spline was adopted here (He and Ng 1999).

The ICK technique was used to bias-correct precipitation ensembles from the NCEP Global Ensemble Forecast System (GEFS) and streamflow ensembles from the NWS River Forecast Centers (RFCs). In general, the forecast biases were substantially reduced following ICK. By design, the unconditional probabilities are unbiased, as the observed climatology forms the prior distribution in ICK. The reliability of the forecast

TABLE 5. ROC areas for ESP and (ICK) under independent validation.

	Event (climatological exceedance probability)				
Lead time (h)	0.5	0.75	0.9	0.975	0.99
24	0.97 (0.98)	0.94 (0.97)	0.95 (0.98)	0.91 (0.96)	0.95 (0.98)
48	0.97 (0.98)	0.96 (0.97)	0.97 (0.98)	0.96 (0.98)	0.96 (0.98)
72	0.97 (0.98)	0.96 (0.97)	0.96 (0.97)	0.96 (0.97)	0.95 (0.96)
96	0.97 (0.98)	0.96 (0.97)	0.96 (0.97)	0.96 (0.96)	0.96 (0.97)
120	0.97 (0.98)	0.96 (0.96)	0.96 (0.96)	0.94 (0.95)	0.96 (0.96)

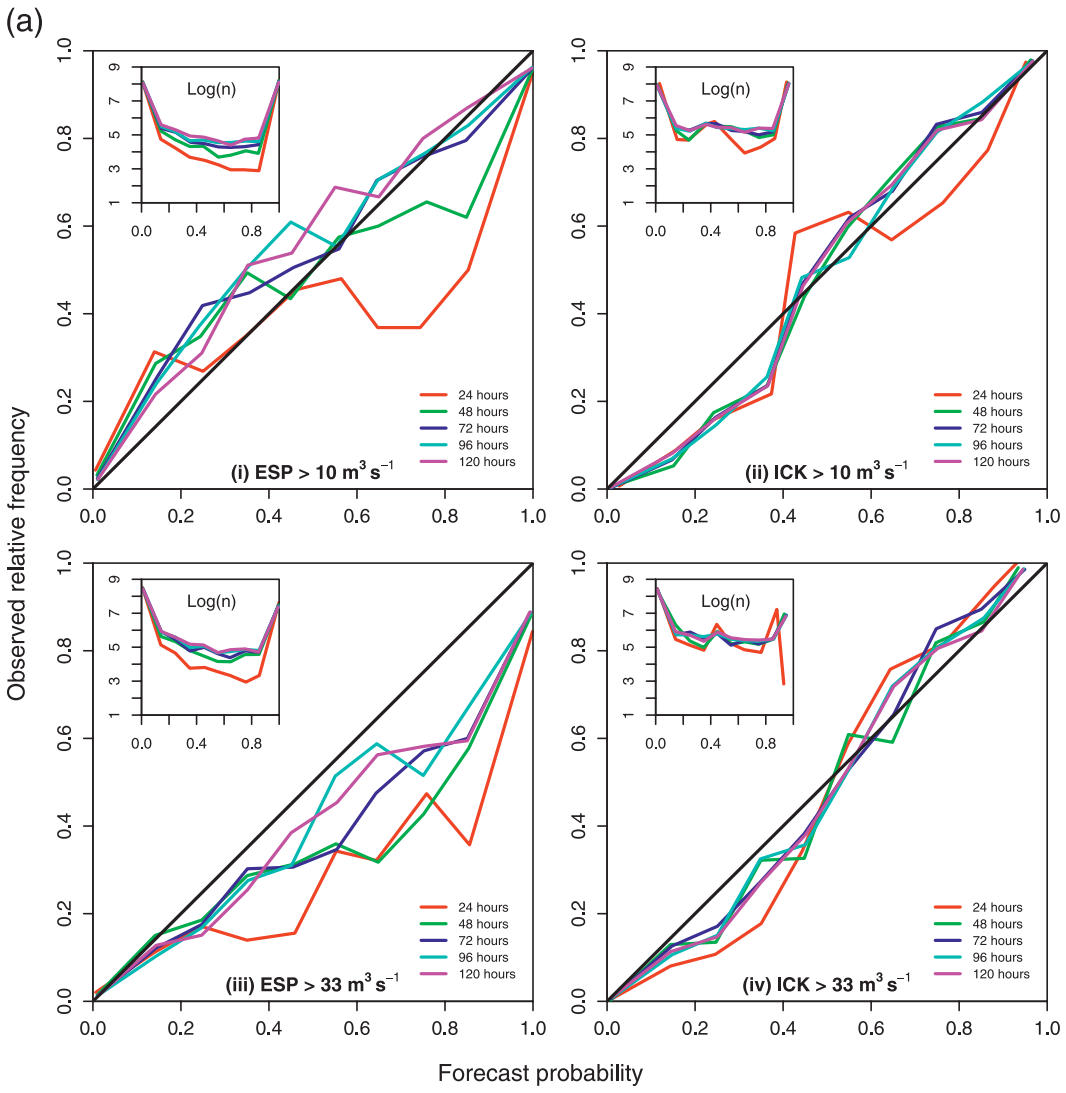

FIG. 12. (a) Reliability diagrams for raw ESP and ICK forecasts (low/all flows). (b) Reliability diagrams for raw ESP and ICK forecasts (medium flows).

probabilities and the mean continuous ranked probability score (CRPS) were also improved by ICK. The CRPS provides a critical check on the performance of ICK, as the objective function minimizes the CRPS on a per quantile basis. The CRPS skill of the ICK precipitation forecasts was 17%–32% (i.e., the mean CRPS of ICK was 17%–32% better than the raw GEFS), and the CRPS skill of the ICK streamflow forecasts was 5%–16%. In both cases, the improvements from ICK declined systematically with forecast lead time, due to the reduced indicator cross correlations at longer lead times. Despite the overall gains in mean CRPS, the ICK forecasts consistently underestimated large precipitation amounts. This was attributed to an attenuation effect or errors in variables, which is well known in least squares estimation. A simple adaptation of ICK was proposed, whereby the indicator cross correlations were uniformly inflated to compensate for their attenuation by sampling error. This substantially reduced the conditional biases at the expense of a small increase in unconditional bias and mean CRPS. In both case studies, the improvements in mean CRPS and other verification statistics were obtained from relatively small sample sizes and uninformative prior distributions, that is, an unconditional climatology with no stratification by season or other environmental conditions. Given a larger sample size, a seasonal or regime-dependent climatology might be considered.

Future work will focus on incorporating additional (prior) knowledge into the ICK estimators via auxiliary

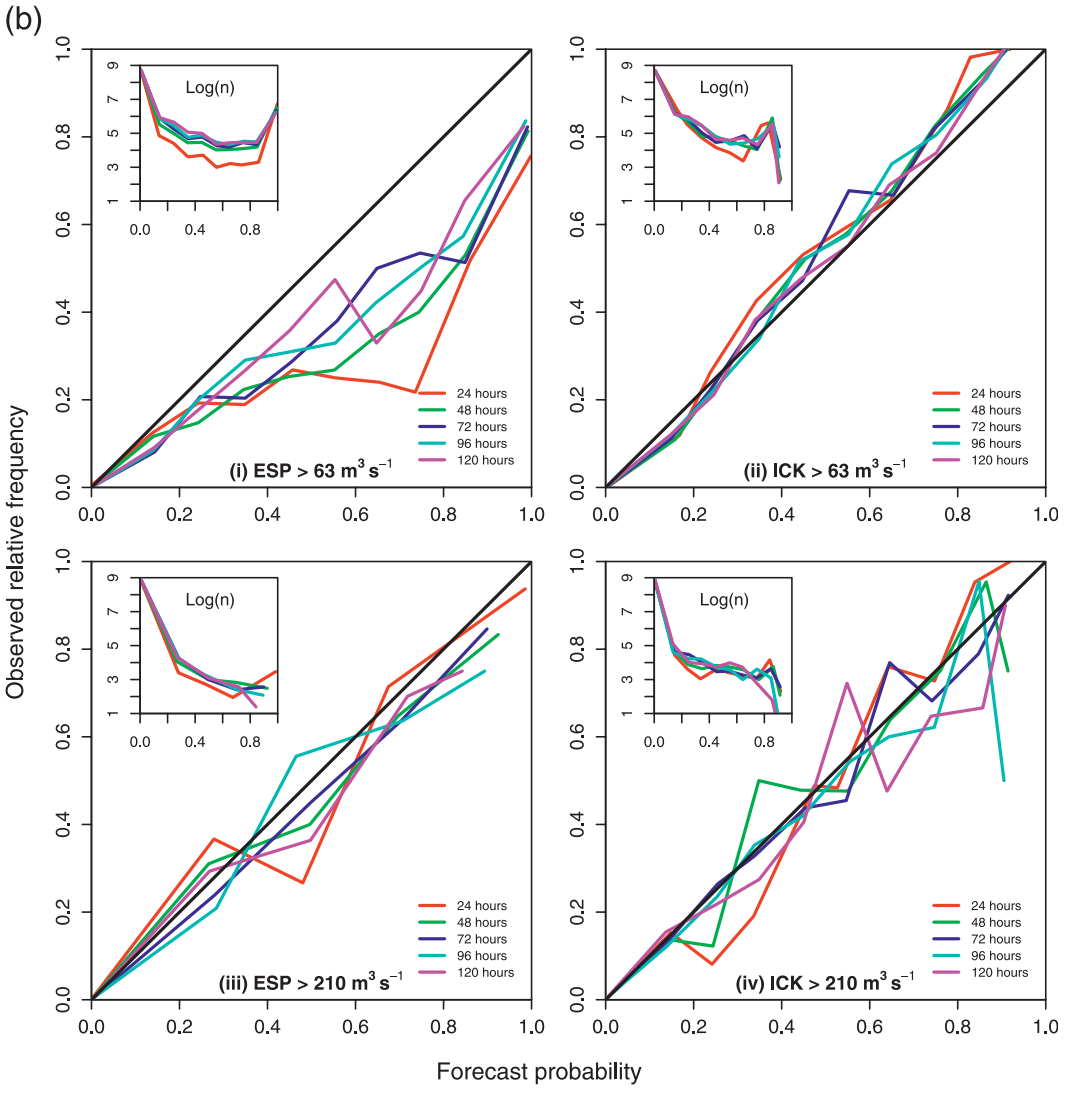

FIG. 12. (*Continued*)

variables; comparing its performance to other statistical postprocessors, such as spline smoothing and kernel density estimation; and testing the technique in an operational setting. Currently, ICK does not model the joint distribution of X across several forecast lead times. This is necessary for dynamic modeling with bias-corrected forcing, such as hydrologic routing of precipitation. One extension of ICK involves integrating forcing ensembles from several meteorological models into the ESP process, including precipitation ensembles from the NCEP GEFS and NCEP short-range ensemble forecast (SREF) models. Such multimodel ensembles are naturally handled in ICK. The additional members simply contribute additional covariates on which to condition the ICK estimate, and any information shared between models is eliminated through an orthogonal decomposition of the indicator covariances. Another study involves structure identification and bias-correction using hindcasts from the frozen GEFS, which will help to quantify the sampling uncertainties of ICK, particularly for extreme events.

Acknowledgments. This work was supported by the National Oceanic and Atmospheric Administration (NOAA) through the Advanced Hydrologic Prediction Service (AHPS) and the Climate Prediction Program for the Americas (CPPA). We thank Yuejian Zhu of the National Centers for Environmental Prediction (NCEP) for providing the GEFS precipitation forecasts and Dingchen Hou, also of NCEP, for providing comments on an earlier draft of this manuscript.

REFERENCES

Ajami, N., Q. Duan, and S. Sorooshian, 2007: An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resour. Res.,* **43,** W01403, doi:10.1029/2005WR004745.

Ali, A. I., and U. Lall, 1996: A kernel estimator for stochastic subsurface characterization from drill-log data. *Ground Water,* **34,** 647–658.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Anderson, M. G. and P. D. Bates, Eds., 2001: *Model Validation: Perspectives in Hydrological Science*. John Wiley and Sons, 512 pp.

Beven, K., and A. Binley, 1992: The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Processes,* **6,** 279–298.

Borga, M., and A. Vizzaccaro, 1997: On the interpolation of hydrologic variables: Formal equivalence of multiquadratic surface fitting and kriging. *J. Hydrol.,* **195,** 160–171.

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

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

Bröcker, J., and L. A. Smith, 2007: Scoring probabilistic forecasts: On the importance of being proper. *Wea. Forecasting,* **22,** 382–388.

Brown, J. D., and G. Heuvelink, 2005: Assessing uncertainty propagation through physically based models of soil water flow and solute transport. *The Encyclopedia of Hydrological Sciences,* M. Anderson, Ed., John Wiley and Sons, 1181–1195.

Buja, A., T. J. Hastie, and R. J. Tibshirani, 1989: Linear smoothers and additive models. *Ann. Stat.,* **17,** 453–510.

Ciach, G. J., M. L. Morrissey, and W. F. Krajewski, 2000: Conditional bias in radar rainfall estimation. *J. Appl. Meteor.,* **39,** 1941–1946.

Cressie, N. A. C., 1993: *Statistics for Spatial Data*. rev. ed. John Wiley and Sons, 900 pp.

Demargne, J., M. Mullusky, K. Werner, T. Adams, S. Lindsey, N. Schwein, W. Marosi, and E. Welles, 2009: Application of forecast verification science to operational river forecasting in the U.S. National Weather Service. *Bull. Amer. Meteor. Soc.,* **90,** 779–784.

Deutsch, C. V., and A. G. Journel, 1992: *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press, 340 pp.

Draper, N. R., and H. Smith, 1998: *Applied Regression Analysis*. 3rd ed. John Wiley and Sons, 736 pp.

Dubrule, O., 1983: Two methods with different objectives: Splines and kriging. *Math. Geol.,* **15,** 245–257.

Eckel, F. A., and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting,* **13,** 1132–1147.

Fawcett, T., 2006: An introduction to ROC analysis. *Pattern Recognit. Lett.,* **27,** 861–874.

Franz, K. J., H. C. Hartmann, S. Sorooshian, and R. Bales, 2003: Verification of National Weather Service ensemble streamflow predictions for water supply forecasting in the Colorado River basin. *J. Hydrometeor.,* **4,** 1105–1118.

Fuller, W. A., 1987: *Measurement Error Models*. John Wiley and Sons, 440 pp.

Georgakakos, K. P., 2003: Probabilistic climate-model diagnostics for hydrologic and water resources impact studies. *J. Hydrometeor.,* **4,** 92–105.

Glahn, H., and D. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Gneiting, T., E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.,* **69B,** 243–268.

Golub, G. H., and C. F. Van Loan, 1996: *Matrix Computations*. 3rd ed. Johns Hopkins, 642 pp.

Goovaerts, P., 1997: *Geostatistics for Natural Resource Evaluation*. Oxford University Press, 483 pp.

Green, D. M., and J. M. Swets, 1966: *Signal Detection Theory and Psychophysics*. Wiley and Sons Inc., 455 pp.

Gupta, H. V., K. J. Beven, and T. Wagener, 2005: Model calibration and uncertainty estimation. *The Encyclopedia of Hydrological Sciences,* M. Anderson, Ed., John Wiley & Sons, 2015–2032.

Hamill, T. M., 1997: Reliability diagrams for multicategory probabilistic forecasts. *Wea. Forecasting,* **12,** 736–741.

——, J. S. Whittaker, and S. L. Mullen, 2006: Reforecasts: An important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.,* **87,** 33–46.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.,* **136,** 2620–2632.

Hansen, P. C., 1987: The truncated *SVD* as a method for regularization. *BIT Numer. Math.,* **27,** 534–553.

Hashino, T., A. A. Bradley, and S. S. Schwartz, 2006: Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci. Discuss.,* **3,** 561–594.

He, X., and P. Ng, 1999: COBS: Qualitatively constrained smoothing via linear programming. *Comput. Stat.,* **14,** 315–337.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570.

Hsu, W.-R., and A. H. Murphy, 1986: The attributes diagram: A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.,* **2,** 285–293.

Isaaks, E. H., and R. M. Strivastava, 1989: *An Introduction to Applied Geostatistics*. Oxford University Press, 561 pp.

Jolliffe, I. T. and D. B. Stephenson, Eds., 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley and Sons, 240 pp.

Journel, A. G., and Ch J. Huijbregts, 1978: *Mining Geostatistics*. Academic Press, 600 pp.

Katz, R. W., and A. H. Murphy, 1997: *Economic Value of Weather and Climate Forecasts*. Cambridge University Press, 238 pp.

Kelly, K. S., and R. Krzysztofowicz, 1997: A bivariate meta-Gaussian density for use in hydrology. *Stochastic Hydrol. Hydraul.,* **11,** 17–31.

Kennedy, E. J., 1983: Computation of continuous records of streamflow. Techniques of Water-Resources Investigations of the United States Geological Survey Rep. 3-A13, 52 pp. [Available online at http://pubs.usgs.gov/twri/twri3-a13/pdf/TWRI_3-A13.pdf.]

Koenker, R., 2005: *Quantile Regression*. Cambridge University Press, 370 pp.

Larson, L. W., 1976: *Precipitation and its Measurement: A State of the Art*. Water Resources Series, Vol. 24, University of Wyoming, 71 pp.

Mason, S. J., and N. E. Graham, 2002: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. *Quart. J. Roy. Meteor. Soc.,* **128,** 2145–2166.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.,* **22,** 1087–1095.

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

NRC, 2006: *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts*. National Academies Press, 112 pp.

O'Conner, P. D. T., 2002: *Practical Reliability Engineering*. 4th ed. John Wiley and Sons, 540 pp.

Olsson, J., and G. Lindström, 2008: Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *J. Hydrol.,* **350,** 14–24.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

Saito, H., and P. Goovaerts, 2002: Accounting for measurement error in uncertainty modeling and decision-making using indicator kriging and *p*-field simulation: Application to a dioxin contaminated site. *Environmetrics,* **13,** 555–567.

Schaake, J., and Coauthors, 2007: Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci.,* **4,** 655–717.

Schweppe, F. C., 1973: *Uncertain Dynamic Systems*. Prentice-Hall, 576 pp.

Seo, D.-J., 1996: Nonlinear estimation of spatial distribution of rainfall - An indicator cokriging approach. *Stochastic Hydrol. Hydraul.,* **10,** 127–150.

——, H. D. Herr, and J. C. Schaake, 2006: A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci.,* **3,** 1987–2035.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.,* **135,** 3209–3220.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

Talagrand, O., 1997: Assimilation of observations, an introduction. *J. Meteor. Soc. Japan,* **75,** 191–209.

Toth, Z., E. Kalnay, S. M. Tracton, R. Wobus, and J. Irwin, 1997: A synoptic evaluation of the NCEP ensemble. *Wea. Forecasting,* **12,** 140–153.

Watson, G. S., 1984: Smoothing and interpolation by kriging and with splines. *Math. Geol.,* **16,** 601–615.

Wei, M., Z. Toth, R. Wobus, and Y. Zhu, 2008: Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus,* **60A,** 62–79.

Wilczak, J., and Coauthors, 2006: Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004. *J. Geophys. Res.,* **111,** D23S28, doi:10.1029/2006JD007598.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.

——, 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.,* **16,** 361–368.

Yandell, B. S., 1993: Smoothing splines—A tutorial. *Statistician,* **42,** 317–319.