Comparative Verification of Guidance and Local Quantitative Precipitation Forecasts: Calibration Analyses

ROMAN KRZYSZTOFOWICZ

Department of Systems Engineering and Division of Statistics, University of Virginia, Charlottesville, Virginia

ASHLEY A. SIGREST

Department of Systems Engineering, University of Virginia, Charlottesville, Virginia

(Manuscript received 20 August 1998, in final form 10 December 1998)

ABSTRACT

A comparative verification is reported of 2631 matched pairs of *quantitative precipitation forecasts* (QPFs) prepared daily from 1 October 1992 to 31 October 1996 by the Hydrometeorological Prediction Center (HPC) and the Weather Service Forecast Office in Pittsburgh (PIT). The predictand is the 24-h spatially averaged precipitation amount. The property of QPF being verified is *calibration.* Four interpretations of each QPF are hypothesized and verified: an exceedance fractile, a conditional exceedance fractile, the mean, and the conditional mean (with conditioning on precipitation occurrence).

Time series of calibration statistics support the following conclusions. (i) The HPC QPF, which lacks an official interpretation, is calibrated as the 18%–19% exceedance fractile and as the conditional median, on average. (ii) It serves as a useful guidance to local forecasters. (iii) Pittsburgh forecasters adjust the guidance in the correct direction to produce PIT QPF, whose official interpretation is the (unconditional) median. (iv) Relative to this interpretation, HPC QPF has a substantial overestimation bias, which hampers the calibration of PIT QPF. (v) The calibration of each QPF lacks consistency over time. (vi) To improve the potential for good calibration, the guidance QPF and the local QPF should be given the same probabilistic interpretation; the conditional median of the spatially averaged precipitation amount is recommended.

1. Introduction

The premise underlying the flow of forecast information within the National Weather Service is that each stage of processing improves a forecast. This premise is embodied in a concept of the *end-to-end forecast process*—a blueprint for an integrated hydrometeorological forecasting system (Wernly and Uccellini 1998; Office of Meteorology 1999). In order to optimize system integration, forecasts should be evaluated at each processing stage.

An evaluation of forecasts input to and output from a particular processing stage can be accomplished via a *matched comparative verification.* This article reports such a verification for two forecasts of precipitation amount: (i) HPC QPF—quantitative precipitation forecast prepared by the Hydrometeorological Prediction Center of the National Centers for Environmental Prediction (formerly the Weather Forecast Branch in the Meteorological Operations Division of the National Meteorological Center), which provides guidance to local forecasters; and (ii) PIT QPF—quantitative precipitation forecast prepared by the Weather Service Forecast Office (WSFO) in Pittsburgh, Pennsylvania. Subject to verification are 2631 matched pairs of QPFs prepared during the four years from 1 October 1992 to 31 October 1996 for two river basins in Pennsylvania.

The property of forecasts that is verified herein is called *calibration.* Broadly speaking, good calibration means that a forecast has a well-defined interpretation which is consistently maintained over time. Good calibration is necessary in order that users can take a forecast at its face value. This verification study addresses six questions. (i) What is the actual probabilistic interpretation of a QPF? Four interpretations are hypothesized and verified: an exceedance fractile, a conditional exceedance fractile, the mean, and the conditional mean. (ii) Is a given interpretation consistently maintained over time? (iii) Does the guidance provide significant information to local forecasters? (iv) Do local forecasters improve upon calibration of the guidance? (v) What should be the official probabilistic interpretation of a QPF? (vi) What improvements to the forecasting system are called for?

Corresponding author address: Professor Roman Krzysztofowicz, University of Virginia, Thornton Hall, SE, Charlottesville, VA 22903.

2. Predictand and forecasts

a. Hydrologic predictand

Let ω denote the *basin average precipitation amount* accumulated during 24 h beginning at 1200 UTC. At the forecast time, this amount is uncertain and thus is treated as a random variable, denoted *W*. The importance of this predictand arises from a hydrologic forecasting problem: the primary source of uncertainty in short-term river stage forecasts is the unknown future volume of water precipitating over the river basin. By inputting a forecast of *W* to a hydrologic model, one may expect to increase the lead time and reliability of river stage forecasts.

Forecasts of *W* have been made for two river basins: (i) the Lower Monongahela River basin above Connellsville, which covers 3429 km^2 (1324 mi^2) in Pennsylvania and Maryland, with the elevation ranging from 262 m (860 ft) to 979 m (3213 ft) at Mount Davis—the highest point in Pennsylvania; (ii) the Upper Allegheny River basin above the Kinzua dam, which covers 5853 km^2 (2260 mi^2) in Pennsylvania and New York, with the elevation ranging from about 366 m (1200 ft) to 762 m (2500 ft). Besides different sizes and elevations, the basins have somewhat different weather characteristics.

b. Guidance forecast

The QPF guidance is prepared judgmentally by HPC forecasters who apply knowledge, experience, and techniques to observations, analyses, and model outputs (Funk 1991). The guidance is for all 48 conterminous states and has graphical form: it shows isopleths of *spatially averaged precipitation* (SAP) fields, which can be interpolated to ''basin or even subbasin averages'' (Olson et al. 1995).

For the purpose of a comparative verification, the graphical guidance is transformed into a forecast of the basin average precipitation amount as follows. River basin boundary is overlaid, SAP isopleths are interpolated to create a grid of the field, the field is numerically integrated over the basin, and the result is divided by the basin area; the output is an estimate *y*. This estimate constitutes a deterministic forecast of *W*, to be called HPC QPF or guidance QPF.

c. Local forecast

Forecast of *W* is produced by the WSFO Pittsburgh according to a Bayesian methodology (Krzysztofowicz et al. 1993). The forecaster prepares a *source forecast* in a graphical form. It consists of three sets of isopleths of the SAP field over the forecast area. Each set of isopleths corresponds to a specified *exceedance probability p* ($p = 0.75, 0.50, 0.25$). The defining property of a set of isopleths for a given *p* is that the integral of the field delineated by the isopleths over a river basin, divided by the basin area, gives an estimate x_{100p} of the $100p\%$ exceedance fractile of the basin-average precipitation amount *W*. In other words, x_{100p} is an estimate such that, in the forecaster's judgment, the exceedance probability is

$$P(W > x_{100p}) = p, \qquad p = 0.75, 0.50, 0.25. \quad (1)$$

Specifically, the *50% exceedance fractile* x_{50}, also called the *median*, is an estimate that is equally likely to be exceeded or not exceeded; that is, $P(W > x_{50}) = P(W \leq x_{50}) = 0.50$. The *75% exceedance fractile* x_{75} is an estimate such that $P(W > x_{75}) = 0.75$. The 25% *exceedance fractile* x_{25} is an estimate such that $P(W > x_{25}) = 0.25$.

When preparing the source forecast, the Pittsburgh forecaster usually takes the graphical HPC guidance as initial isopleths for $p = 0.50$ (Krzysztofowicz and Drake 1993). He may then adjust the isopleths and/or isopleths values so that they produce an estimate x_{50} that conforms to his judgment. With notation simplified to $x = x_{50}$, this estimate constitutes a deterministic forecast of *W*, to be called PIT QPF or local QPF.

d. Verification samples

During the four years from 1 October 1992 to 31 October 1996, a total of 2631 joint observations of the two forecasts and the predictand have been collected, 1317 for the Monongahela basin and 1314 for the Allegheny basin. Inasmuch as the predictand cannot be observed directly, its ''observation'' is estimated as a weighted average of rain gauge reports within and near the basin. Estimation is performed daily by the Ohio River Forecast Center in Wilmington as part of routine operational forecasting. Estimation procedures account for the topology of the rain gauge network, missing reports, erroneous reports, and the conversion of snowfall into its water equivalent.

Matched comparative verifications of the two forecasts are performed for each basin separately. Verification statistics are computed monthly for a set of forecasts prepared during the last three months. Thus each statistic has a time series of 47 values beginning in December 1992 and ending in October 1996. Because of the 2-month stagger of the verification periods, any observation from month 3 through 47 of the 4-yr epoch affects three consecutive values of the statistic. As a result, the time series of a statistic behaves similarly to a moving average. The *3-month verification period* was chosen in order to accumulate a sample of reasonable size. The sample size varies from 42 to 92 and has an average of 81 matched observations per basin.

e. Behavior of the predictand

Before verifying forecasts, it is desirable to characterize behavior of the predictand alone. Toward this end,

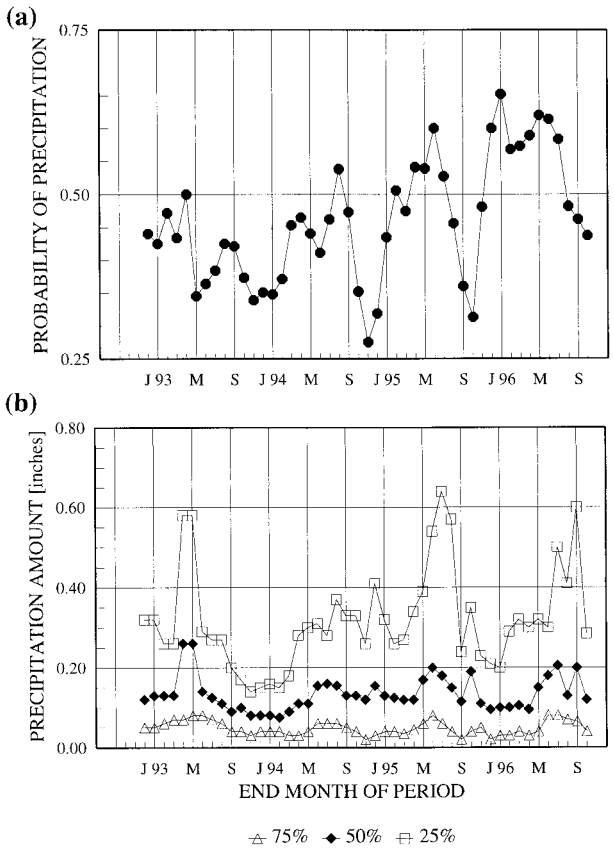

FIG. 1. Statistics of the 24-h basin average precipitation amount observed from 1200 UTC during the 3-month verification periods: (a) probability of precipitation occurrence π, (b) exceedance fractiles of the amount, conditional on precipitation occurrence ($\omega_{75|0}$, $\omega_{50|0}$, $\omega_{25|0}$); Monongahela basin, Oct 1992–Oct 1996.

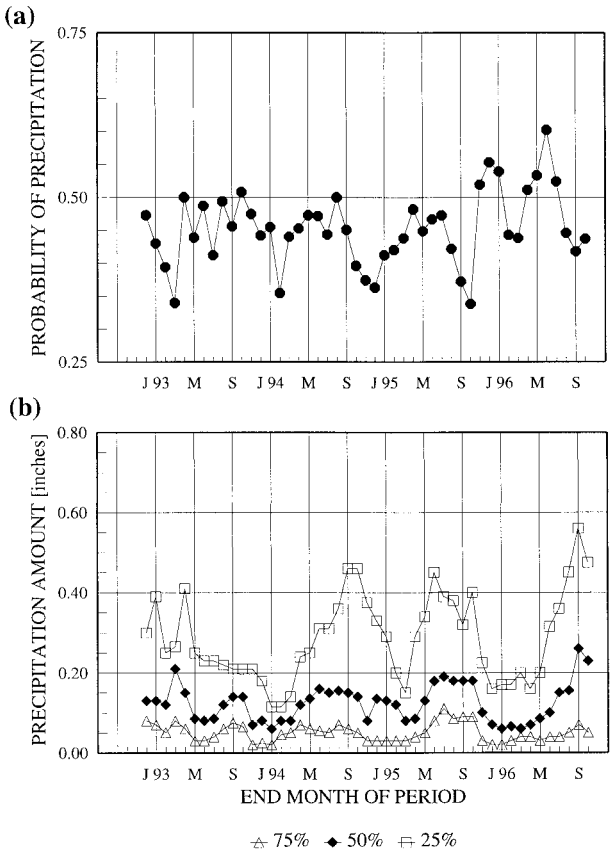

FIG. 2. Statistics of the 24-h basin average precipitation amount observed from 1200 UTC during the 3-month verification periods: (a) probability of precipitation occurrence π, (b) exceedance fractiles of the amount, conditional on precipitation occurrence ($\omega_{75|0}$, $\omega_{50|0}$, $\omega_{25|0}$); Allegheny basin, Oct 1992–Oct 1996.

observations of W from the 3-month verification sample were used to estimate the probability of precipitation,

$$\pi = P(W > 0), \tag{2}$$

and the 75%, 50%, and 25% exceedance fractiles of W, conditional on precipitation occurrence; denoted $\omega_{100p|0}$, the conditional exceedance fractile is defined by

$$P(W > \omega_{100p|0} \,|\, W > 0) = p,$$
$$p = 0.75,\ 0.50,\ 0.25. \tag{3}$$

Together, (2)–(3) characterize the probability distribution of the 24-h basin average precipitation amount on any day during the 3-month verification period. Figures 1 and 2 display the time series of these elements for each of the basins.

The 4-yr average probability of precipitation was about the same in the two basins (0.46 in Monongahela and 0.45 in Allegheny); however, the probability varied over time more in the Monongahela basin (between 0.28 and 0.65) than it did in the Allegheny basin (between 0.34 and 0.60). Moreover, the Monongahela basin experienced a trend of increasing annual average probability of precipitation from 0.41 in 1993 to 0.56 in 1996.

The conditional exceedance fractiles $\omega_{75|0}$ and $\omega_{25|0}$ define a 50% *conditional credible interval* about the conditional median $\omega_{50|0}$; that is, $P(\omega_{75|0} < W \leq \omega_{25|0} \,|\, W > 0) = 0.50$. This credible interval, $\omega_{25|0} - \omega_{75|0}$, characterizes the variability of the precipitation amount per wet day. As Figs. 1b and 2b reveal, most of the time increased variability was accompanied by increased skewness toward higher amounts, as $\omega_{50|0} - \omega_{75|0} < \omega_{25|0} - \omega_{50|0}$. Both the median amount per wet day and the variability of the amount exhibited within year seasonality in the Allegheny basin. In the Monongahela basin, the seasonality was disrupted by a trend, from November 1993 to July 1995, of increasing median and variability. If nonstationarity of the predictand raises the difficulty of a forecasting task, precipitation over the Monongahela basin offered a test of this hypothesis.

3. Verification methodology

a. Hypothesized interpretations of forecasts

In order that users can take a forecast at its face value, *normative interpretation* of the forecast must be well

defined and consistently adhered to over time. Consistent interpretability of forecasts requires good calibration (Murphy and Winkler 1974; Lichtenstein and Fischhoff 1980; Wallsten et al. 1993), also termed reliability or unbiasedness. The local QPF has the normative interpretation as a median of predictand W. The guidance QPF has no official probabilistic interpretation.

The objective of this study is to compare the calibration of the two QPFs under the same interpretation and, perhaps, to uncover a probabilistic interpretation of the guidance QPF. Toward this end four alternative interpretations of forecast x (and y) are hypothesized:

(i) the $100p$% exceedance fractile of W,

$$P(W > x) = p, \qquad 0 < p < 1; \tag{4}$$

(ii) the $100q$% exceedance fractile of W, conditional on precipitation occurrence,

$$P(W > x \,|\, W > 0) = q, \qquad 0 < q < 1; \tag{5}$$

(iii) the mean (expectation) of W,

$$x = E(W); \tag{6}$$

(iv) the mean of W, conditional on precipitation occurrence,

$$x = E(W \,|\, W > 0). \tag{7}$$

It is said that QPF is *well calibrated* with respect to a given interpretation if the defining equality holds always.

b. Necessary conditions for calibration

To conclude that a QPF is well calibrated under a hypothesized interpretation, one must empirically verify that the equation defining the interpretation is satisfied for every forecast value x. While theoretically straightforward, practically it is rather a difficult task because x can take on any of the infinite number of values. For this reason, when predictand W is a continuous variate, it has been almost a universal practice in meteorology as well as other fields to verify only a necessary condition for calibration (Alpert and Raiffa 1982; Murphy and Winkler 1974, 1979).

For precipitation amount as the predictand, the situation is even more complicated because W is a mixed (discrete-continuous) variate: $P(W = 0) > 0$ implies that the distribution of W is continuous only with probability $P(W > 0) < 1$. A suitable theory of calibration was derived by Krzysztofowicz and Sigrest (1999, appendix A) and will be applied herein.

With forecast x viewed as an observation of random variable X, the necessary condition for calibration of a QPF under interpretation (4), (5), (6), or (7), respectively, takes the following form:

$$P(W > X \,|\, X > 0) = p, \qquad 0 < p < 1; \tag{8}$$

$$P(W > X \,|\, W > 0) = q, \qquad 0 < q < 1; \tag{9}$$

$$E(X) = E(W); \tag{10}$$

$$E(X \,|\, W > 0) = E(W \,|\, W > 0). \tag{11}$$

The left side of (8) or (9) defines the probability that on a randomly chosen day the observation of predictand W is larger than the observation of forecast variate X. The conditioning on $X > 0$ in (8) arises because whenever, in the forecaster's judgment, the probability of event $W > 0$ is *equal to* or *less than* p, the forecaster should set $X = 0$. Hence, for a fixed p, forecast $X = 0$ could never be well calibrated. Ergo, it would be improper to verify its calibration as a $100p$% exceedance fractile of W. The conditioning on $W > 0$ in (9) follows directly from (5). The left side of (10) or (11) defines the mean of forecast variate X, unconditional in (10) and conditional on $W > 0$ in (11).

Each of the equations, (8)–(11), constitutes only a necessary condition for calibration because the corresponding interpretation of a QPF given by (4)–(7) is verified not for every forecast value x, but only on average. Therefore, one can infer that if definition (4), (5), (6), or (7) holds for every x, then the corresponding condition (8), (9), (10), or (11) holds, but not vice versa. Nonetheless, it has been almost a universal practice to use a necessary condition as the definition of a well-calibrated forecast. We shall follow this practice.

In summary, (8)–(11) constitute a basis for calibration analyses reported in the subsequent section. Inasmuch as the average sample size is 81, and the time series of calibration statistics reveal a rather consistent dominance order (one lying above the other), inferences are described without reference to any measures of uncertainty or confidence. Results of significance tests are included as footnotes to tables, and an analysis of the sampling uncertainty for two calibration statistics is presented in the appendix.

4. Calibration analyses

a. Calibration of exceedance fractiles

The empirical verification of a set of forecasts requires two counts: N—the number of forecasts $x > 0$, and n—the number of forecasts $x > 0$ that are followed by an observation of precipitation amount $\omega > x$. Therefrom, the frequency estimator of probability $P(W > X \,|\, X > 0)$ is $r = n/N$. Based on (8), the QPFs in a given set are said to be well calibrated as the $100p$% exceedance fractile of the predictand if $r \simeq p$.

The time series of exceedance frequency r for HPC QPF and PIT QPF are plotted in Figs. 3a and 4a. Annual averages and standard deviations of r computed from these time series are reported in Tables 1 and 2. At first glance, neither forecast has been calibrated consistently

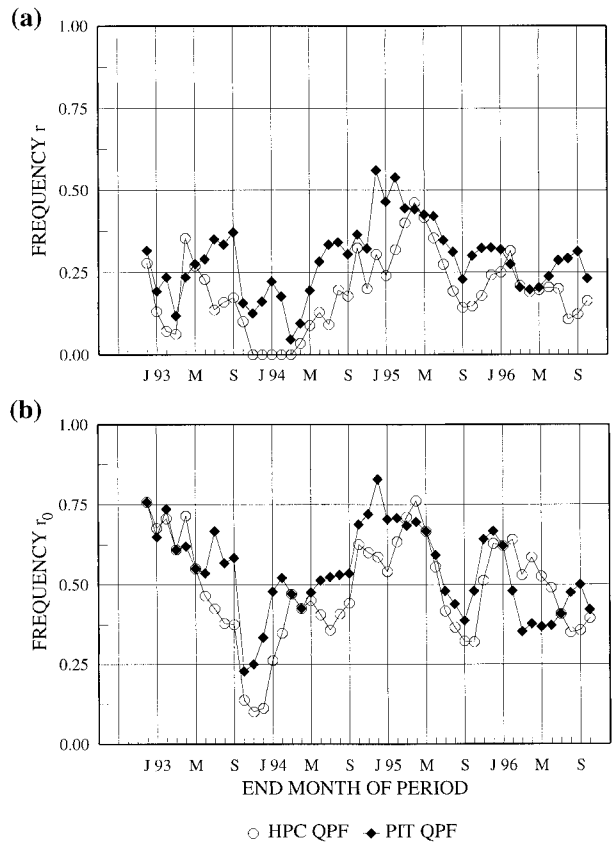

FIG. 3. Calibration of QPF as (a) an exceedance fractile of W, and (b) an exceedance fractile of W, conditional on $W > 0$; Monongahela basin, 3-month verification periods.

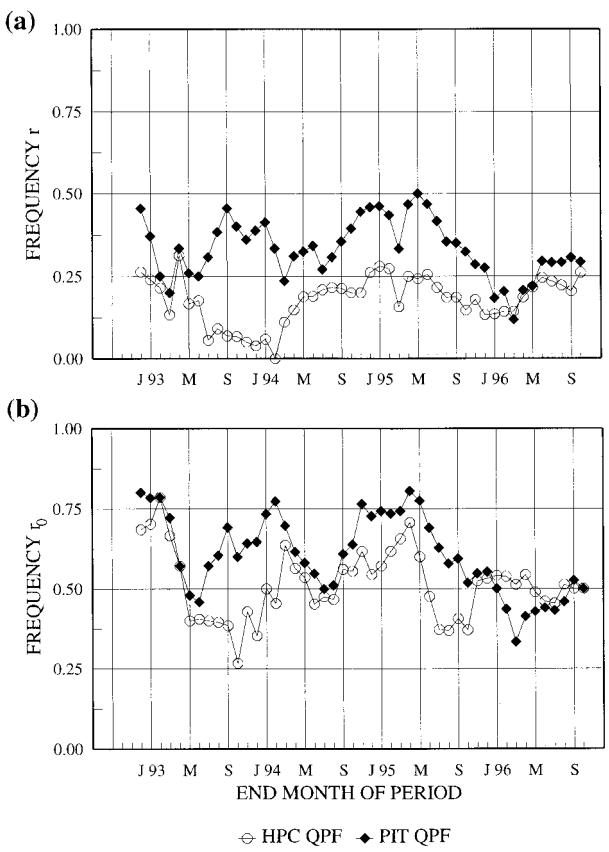

FIG. 4. Calibration of QPF as (a) an exceedance fractile of W, and (b) an exceedance fractile of W, conditional on $W > 0$; Allegheny basin, 3-month verification periods.

over time. Fluctuations of the exceedance frequency have been large, especially for the Monongahela basin.

On the average, HPC QPF is calibrated as the 19% exceedance fractile of W for the Monongahela basin, and the 18% exceedance fractile of W for the Allegheny basin. Thus the guidance substantially overestimates the median of the predictand. This bias was detected by Pittsburgh forecasters a long time ago, with the implication that in order to estimate the median of W, the guidance QPF must be decreased, on the average. As Figs. 3a and 4a reveal, Pittsburgh forecasters have been making skillful adjustments. In all but five verification periods for both basins, PIT QPF has the exceedance frequency r closer to 0.50 than HPC QPF. Noteworthy in the Allegheny basin is the period from July 1993 to February 1994 when the calibration of PIT QPF stayed near average r while the calibration of HPC QPF deteriorated. Still, most of the time the adjustments have been insufficient. Instead of being the median, PIT QPF is calibrated, on the average, as the 29% exceedance fractile of W for the Monongahela basin, and the 33% exceedance fractile of W for the Allegheny basin.

The time series of the exceedance frequencies for PIT QPF and HPC QPF parallel each other most of the time. This is not unexpected because a direct survey of Pittsburgh forecasters found that the HPC QPF had the highest utility among all guidance products (Krzysztofowicz and Drake 1993). Overall, local forecasters improve the calibration of their QPF relative to the calibration of the guidance QPF, especially for the Allegheny basin. Nonetheless, an overestimation bias remains and calibration is inconsistent over time.

b. Calibration of conditional exceedance fractiles

The empirical verification of a set of forecasts requires two counts: N_0—the number of observations of precipitation amount $\omega > 0$, and n_0—the number of observations of precipitation amount that is greater than zero, $\omega > 0$, and greater than the forecast, $\omega > x$. Therefrom, the frequency estimate of probability $P(W > X \mid W > 0)$ is $r_0 = n_0/N_0$. Based on (9), the QPFs in a given set are said to be well calibrated as the $100q\%$ conditional exceedance fractile of the predictand if $r_0 \simeq q$.

The time series of exceedance frequency r_0, plotted for both forecasts in Figs. 3b and 4b, oscillate around probability 0.50. Annual averages of r_0 are reported in Tables 1 and 2. The grand averages, 0.48 and 0.51, imply

TABLE 1. Annual statistics of calibration of HPC QPF and PIT QPF as median, conditional median, mean, and conditional mean of the predictand; Monongahela basin.

		Exceedance frequency		Relative bias (%)	
Year/statistic	Forecast	r	r_0	b	b_0
Dec 92–Nov 93					
avg	HPC	0.16	0.49	66.3	38.4
	PIT	0.25	0.56	21.7	6.1
std dev	HPC	0.10	0.22	73.4	70.1
	PIT	0.09	0.17	49.7	43.9
Dec 93–Nov 94					
avg	HPC	0.10	0.41	101.7	79.9
	PIT	0.24	0.52	40.8	28.9
std dev	HPC	0.10	0.14	73.3	81.5
	PIT	0.10	0.10	51.2	52.4
Dec 94–Nov 95					
avg	HPC	0.29	0.53	45.2	29.5
	PIT	0.40	0.61	8.5	0.8
std dev	HPC	0.11	0.15	41.4	32.1
	PIT	0.10	0.13	25.8	25.1
Dec 95–Oct 96					
avg	HPC	0.20	0.50	45.6	27.6
	PIT	0.26	0.46	41.0	29.4
std dev	HPC	0.06	0.11	22.9	21.9
	PIT	0.05	0.10	16.9	15.9
Dec 92–Oct 96					
avg	HPC	0.19[a]	0.48[c]	65.1[a]	44.2[b]
	PIT	0.29	0.54	27.7	16.0
std dev	HPC	0.12	0.16	60.3	59.8
	PIT	0.11	0.14	40.4	38.7

Based on a test of the Behrens–Fisher problem (Lindgren 1976, p. 352), the difference between the average statistic for HPC and the average statistic for PIT is significant at [a] $\alpha < 0.001$, [b] $\alpha = 0.01$, and [c] $\alpha = 0.1$.

that HPC QPF was well calibrated as the conditional median of W, especially for the Allegheny basin where the annual averages are 0.51, 0.51, 0.52, and 0.51. The grand averages, 0.54 and 0.61, imply that PIT QPF slightly underestimated the conditional median of W. This should not be surprising because the normative task of Pittsburgh forecasters was to estimate the median of W, which is always smaller than the conditional median of W.

None of the forecasts was calibrated consistently. The exceedance frequency r_0 of HPC QPF oscillated in a range 0.10–0.76 for the Monongahela basin and 0.27–0.79 for the Allegheny basin. The exceedance frequency r_0 of PIT QPF oscillated in a range 0.23–0.83 for the Monongahela basin and 0.33–0.80 for the Allegheny basin. These ranges, and the standard deviations of r_0 given in Tables 1 and 2, reveal that each forecast was calibrated more inconsistently for the Monongahela basin than for the Allegheny basin. A possible cause is the nonstationarity of W over the Monongahela basin, which is documented in Fig. 1.

The time series of r_0 and r, for the same basin and forecast, have somewhat similar patterns. This similarity has a theoretical explanation. If a forecast specifies the probability of precipitation occurrence, $\pi = P(W > 0)$, and an estimate y which constitutes the $100q$% conditional exceedance fractile of W, such that $P(W > y \mid W > 0) = q$, then y constitutes the $100p$% exceedance fractile of W, such that $P(W > y) = p$, where

$$p = \pi q, \qquad 0 < q < 1. \tag{12}$$

For example, if $\pi = 0.70$ and $q = 0.50$, then $p = 0.35$. In other words, if the guidance specifies an estimate y which is well calibrated as the conditional median of W on a day on which the forecast probability of precipitation is 0.70, then in effect y constitutes a 35% (unconditional) exceedance fractile of W. When a local forecaster receives an estimate y without such a probabilistic interpretation and must estimate the median of W, he starts with a biased guidance.

In summary, HPC QPF could be interpreted as the conditional median of the basin average precipitation amount. As such, it was well calibrated, on the average, although the calibration lacked consistency over time. In retrospect, if this interpretation of the guidance were known to local forecasters, then perhaps the overestimation bias discussed in the preceding sections could have been avoided, while modifications to the guidance could have been more effective in producing local QPFs, which are calibrated consistently.

TABLE 2. Annual statistics of calibration of HPC QPF and PIT QPF as median, conditional median, mean, and conditional mean of the predictand; Allegheny basin.

		Exceedance frequency		Relative bias (%)	
Year/statistic	Forecast	r	r_0	b	b_0
Dec 92–Nov 93					
avg	HPC	0.15	0.51	44.5	21.0
	PIT	0.34	0.64	−4.2	−15.4
std dev	HPC	0.09	0.17	49.2	43.5
	PIT	0.08	0.12	24.2	22.3
Dec 93–Nov 94					
avg	HPC	0.15	0.51	41.1	21.7
	PIT	0.34	0.63	−8.9	−17.5
std dev	HPC	0.08	0.08	38.7	41.3
	PIT	0.06	0.09	26.7	25.3
Dec 94–Nov 95					
avg	HPC	0.22	0.52	41.7	71.4
	PIT	0.40	0.67	−6.5	−19.9
std dev	HPC	0.05	0.12	35.6	19.4
	PIT	0.07	0.10	16.3	13.8
Dec 95–Oct 96					
avg	HPC	0.19	0.51	61.6	38.1
	PIT	0.24	0.46	44.3	24.3
std dev	HPC	0.05	0.03	14.4	14.2
	PIT	0.06	0.06	38.6	30.8
Dec 92–Oct 96					
avg	HPC	0.18*	0.51*	46.9*	24.3*
	PIT	0.33	0.61	5.4	−7.8
std dev	HPC	0.07	0.11	36.8	32.5
	PIT	0.09	0.12	34.3	29.1

* Based on a test of the Behrens–Fisher problem (Lindgren 1976, p. 352), each difference between the average statistic for HPC and the average statistic for PIT is significant at $\alpha < 0.001$.

c. Calibration of means

The empirical verification of a set of forecasts is accomplished by estimating the mean of the forecast, $\overline{x}$, and the mean of the actual precipitation amount, $\overline{\omega}$. Next, a relative bias is calculated as $b = 100(\overline{x} - \overline{\omega})/\overline{\omega}$. Based on (10), a necessary condition for a QPF to be calibrated as the mean of the predictand is $b \simeq 0\%$.

The time series of bias b are plotted in Figs. 5a and 6a, and averages of b are reported in Tables 1 and 2. Over four years, the average mean $\overline{\omega}$ was 0.112″ for the Monongahela basin and 0.104″ for the Allegheny basin.

The plots, averages, and standard deviations of b reveal that HPC QPF had predominantly positive bias whose magnitude fluctuated widely. The average bias over four years was 65.1% for the Monongahela basin and 46.9% for the Allegheny basin. Likewise, PIT QPF had a tendency toward positive bias although Pittsburgh forecasters were able to substantially reduce both its magnitude and variability. The average bias over four years was 27.7% and 5.4%, respectively. Thus on average, PIT QPF could possibly be interpreted as an approximate mean of W for the Allegheny basin. However, on a monthly basis, this interpretation would be subject to a large variability: the average standard deviation of b over four years was 40.4% and 34.3%, respectively.

d. Calibration of conditional means

To verify a set of forecasts empirically, a subset of joint observations of the forecast and the predictand is extracted such that each actual precipitation amount is greater than zero. Using the observations from this subset, the mean of the forecast, $\overline{x}_0$, and the mean of the actual amount, $\overline{\omega}_0$, are estimated. Next, a relative conditional bias is calculated as $b_0 = 100(\overline{x}_0 - \overline{\omega}_0)/\overline{\omega}_0$. Based on (11), a necessary condition for a QPF to be calibrated as the conditional mean of the predictand is $b_0 \simeq 0\%$.

The time series of conditional bias b_0 are plotted in Figs. 5b and 6b, and averages of b_0 are reported in Tables 1 and 2. Over four years, the average conditional mean $\overline{\omega}_0$ was 0.241″ for the Monongahela basin and 0.232″ for the Allegheny basin.

Overall, HPC QPF overestimated the conditional mean of W; however, the 4-yr average of conditional bias b_0, which was 44.2% for the Monongahela basin and 24.3% for the Allegheny basin, was smaller than the 4-yr average of bias b. The 4-yr average b_0 of PIT QPF was 16.0% and −7.8%, respectively. Thus by modifying the guidance, Pittsburgh forecasters were able to substantially reduce the magnitude of the conditional bias.

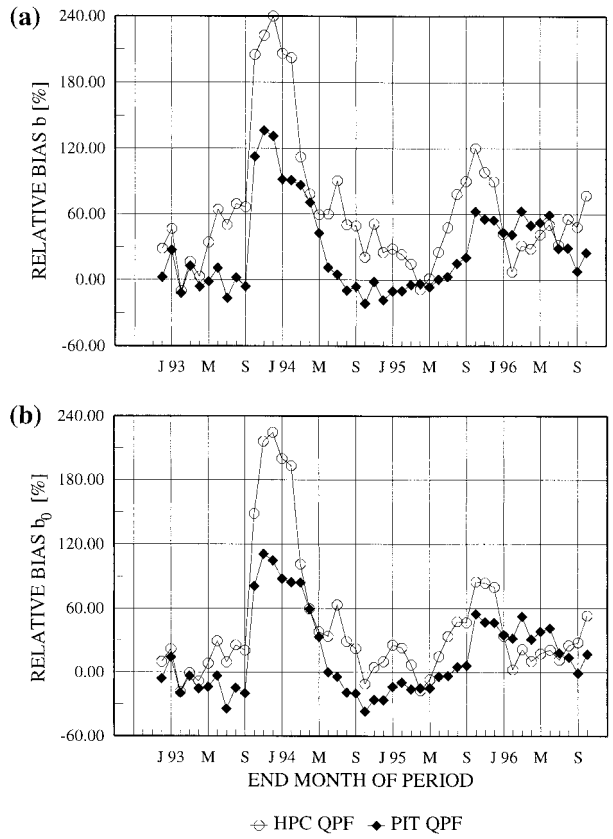

FIG. 5. Calibration of QPF as (a) a mean of W, and (b) a mean of W, conditional on $W > 0$; Monongahela basin, 3-month verification periods.

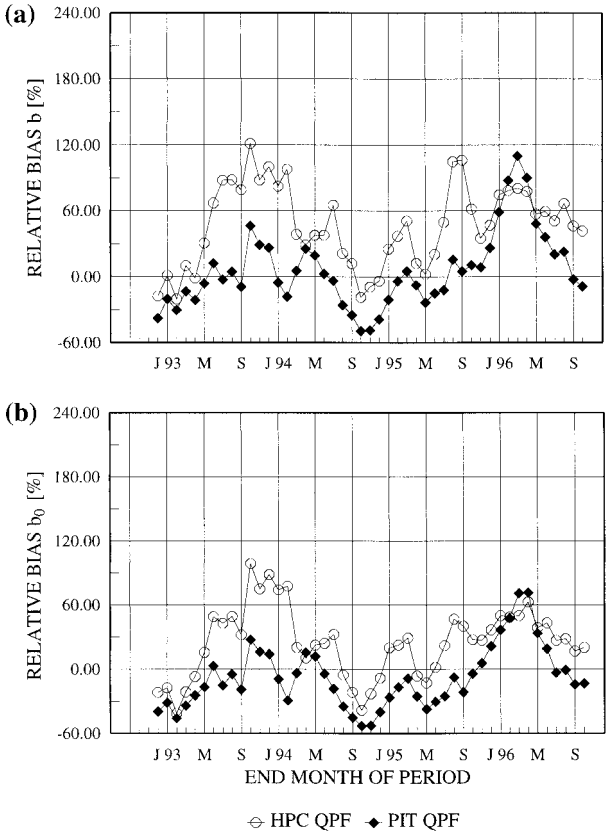

FIG. 6. Calibration of QPF as (a) a mean of W, and (b) a mean of W, conditional on $W > 0$; Allegheny basin, 3-month verification periods.

The time series of b_0 and b, for the same basin and forecast, have similar patterns. This similarity has a theoretical explanation. If a forecast specifies the probability of precipitation occurrence, $\pi = P(W > 0)$, then

$$E(W) = \pi E(W \mid W > 0). \tag{13}$$

For example, if $\pi = 0.30$ and $E(W \mid W > 0) = 1.20''$, then $E(W) = 0.36''$. In other words, if a well-calibrated forecaster specifies $1.20''$ as the conditional mean of W on a day on which the forecast probability of precipitation is 0.30, then in effect the (unconditional) mean of W is $0.36''$.

In summary, HPC QPF cannot be interpreted as the conditional mean of the basin average precipitation amount. On average, PIT QPF could possibly be interpreted as an approximate conditional mean of W for the Allegheny basin. (In fact, PIT QPF for the Allegheny basin falls above the mean of W, with average bias of 5.4%, and below the conditional mean of W, with average conditional bias of -7.8%.)

5. Explanatory analyses

a. Miscalibration and nonstationarity

Suppose that a QPF is interpreted as a $100p\%$ exceedance fractile of W, where p equals the average of the time series of exceedance frequency r (Tables 1 and 2). Large fluctuations of the time series of r around p are indicative of an inconsistent calibration over time (provided that sampling variability is relatively small, which it is, as concluded in the appendix).

Inasmuch as a *consistent calibration* of the guidance QPF is essential to improving calibration of a local QPF, causes of miscalibration should be researched. One hypothesis emerges from a comparison of Fig. 3a with Fig. 1b for the Monongahela basin. The worst calibration of HPC QPF occurred from November 1993 to March 1994, just when the median and the variability of the precipitation amount per wet day were the lowest in four years. Thereafter, from April 1994 to April 1995, the median and the variability of the actual amount trended upward and so did the exceedance frequency r. Other minima and maxima of the frequency r also coincide with sudden fluctuations of the median and the variability of the actual amount.

A comparison of Fig. 4a with Fig. 2b for the Allegheny basin yields similar observations. The worst calibration of HPC QPF occurred from July 1993 to February 1994 and coincided with a downward trend of the median and the variability of the actual amount, both of which reached their lowest values in four years during

TABLE 3. Spearman's rank correlation coefficients ρ between exceedance frequency, r or r_0, and median precipitation amount per wet day, $\omega_{50|0}$.

Basin	Forecast	Correlation ρ between $\omega_{50\|0}$ and r	Correlation ρ between $\omega_{50\|0}$ and r_0
Monongahela	HPC	0.44*	0.22**
	PIT	0.33*	0.19**
Allegheny	HPC	0.40*	0.18
	PIT	0.23**	0.02

Based on Fisher's two-sided test of $\rho = 0$ (Lindgren 1976, p. 478), ρ is significant at * $\alpha = 0.02$, and ** $\alpha = 0.2$.

the 3-month verification periods ending in January and February 1994. The most consistent calibration was from May 1994 to October 1995, a period of relatively high median actual amounts and high variability of the amounts. And again the exceedance frequency r suddenly dropped when both the median and the variability reached their local minimum in March 1995.

In summary, the inconsistent calibration of HPC QPF seems to be associated with a nonstationarity of the predictand. In particular, the tendency toward overestimation increases during periods when both the median and the variability of the actual amount decrease.

b. Calibration and actual amount

A hypothesis was raised by forecasters that the calibration statistic may be associated with the actual precipitation amount. A formal statement of this hypothesis is $\rho \neq 0$, where ρ is Spearman's rank correlation coefficient between the time series of exceedance frequency r and the time series of median precipitation amount per wet day, $\omega_{50|0}$. The correlations reported in Table 3 suggest that a positive association between r and $\omega_{50|0}$ exists but is weak. It is stronger for the HPC QPF than it is for the PIT QPF.

The analysis was repeated under the hypothesis that each QPF is interpreted as a $100q\%$ conditional exceedance fractile of W. In all cases, the correlations between the time series of r_0 and the time series of $\omega_{50|0}$, reported in Table 3, are quite weak.

Normatively, the correlation should be zero because r (or r_0) should be constant over time while $\omega_{50|0}$ may vary from one verification period to the next. Vis-a-vis this norm, the results in Table 3 corroborate the conclusion that the HPC QPF is more appropriately interpreted as a conditional exceedance fractile rather than an (unconditional) exceedance fractile.

c. Calibration and seasonality

Another hypothesis raised by forecasters was that the calibration statistic may exhibit seasonality: better calibration was expected during the cool season dominated by stratiform precipitation than during the warm season

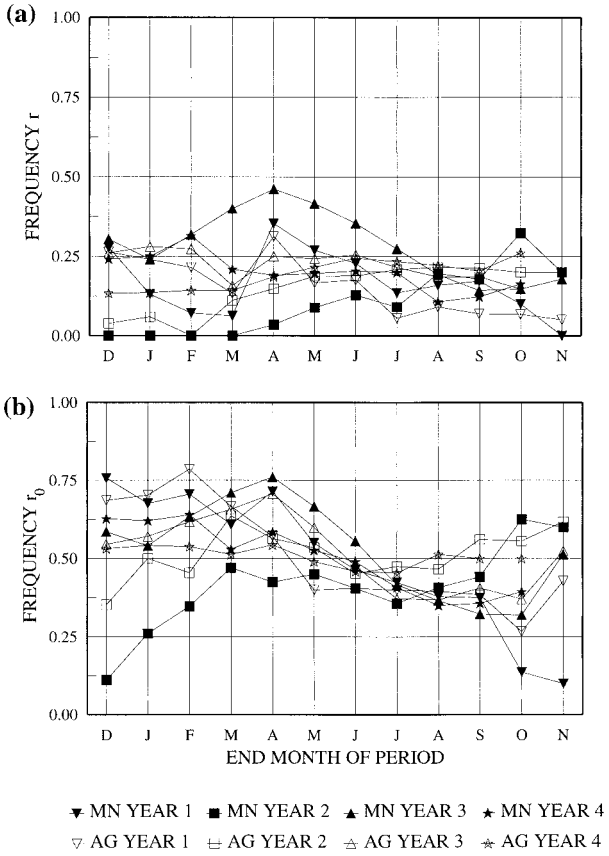

FIG. 7. Seasonality of calibration of HPC QPF as (a) an exceedance fractile of W, and (b) an exceedance fractile of W, conditional on $W > 0$; MN—Monongahela basin, AG—Allegheny basin; 3-month verification periods.

dominated by convective precipitation. To examine this hypothesis, annual time series of the exceedance frequency, r or r_0, for both basins and the same forecast are plotted together in Figs. 7 and 8. The plots do not support the original hypothesis, but instead reveal an association between the season and the consistency of calibration. (The smaller the spread of r, or r_0, in a given month, the more consistent the calibration.)

The HPC QPF has the smallest spread of r in August (Fig. 7a) and the smallest spread of r_0 in July (Fig. 7b). The PIT QPF has the smallest spread of r and r_0 in August (Figs. 8a and 8b). It should be recalled now, that the exceedance frequency, r or r_0, plotted against August verifies the calibration of forecasts prepared during the 3-month period June–July–August. The emerging conclusion is that QPFs prepared during summer have the most consistent year-to-year calibration. In fact, this consistency seems remarkable when compared with the inconsistent calibration of forecasts prepared during October–November–December or February–March–April. Thus, contrary to intuitive expectations, the convective precipitation regime is not an obstacle to preparing consistently calibrated QPFs.

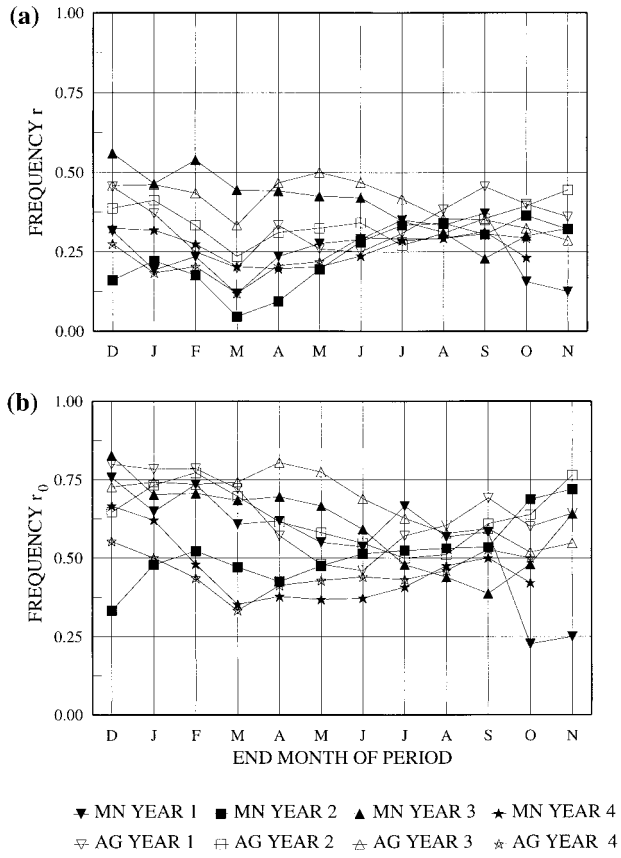

FIG. 8. Seasonality of calibration of PIT QPF as (a) an exceedance fractile of *W*, and (b) an exceedance fractile of *W*, conditional on $W > 0$; MN—Monongahela basin, AG—Allegheny basin; 3-month verification periods.

6. Conclusions

The HPC QPF does not have an official probabilistic interpretation. The calibration analyses reported herein have revealed alternative probabilistic interpretations of HPC QPF when the predictand is the spatially averaged precipitation amount *W*. The conclusions are as follows. (i) On average, the forecast is calibrated as the 18%–19% exceedance fractile of *W* and as the conditional median of *W* (average conditional exceedance frequency 48%–51%). (ii) On average, the forecast is calibrated neither as the mean of *W* nor as the conditional mean of *W*; each mean is substantially overestimated. (iii) These results suggest that HPC forecasters make their QPF as if they judged the median of the precipitation amount, conditional on the hypothesis that precipitation will occur. (That HPC forecasters make a conditional judgment of the precipitation amount was corroborated by the forecasters' recall of their thought process: they first delineate an area where significant precipitation is likely, and next estimate the amount over this area, irrespective of the probability of precipitation occurrence; personal communication with senior HPC forecasters.) (iv) Regardless of the interpretation, the calibration of HPC QPF lacks consistency when verified over 3-month periods. The inconsistencies can be partly explained by the nonstationarity of the predictand. On the other hand, annual averages of the 3-month calibration statistics are generally consistent over the four years.

The task of Pittsburgh forecasters has been to produce a QPF interpretable as the median of *W*. The conclusions are as follows. (i) The HPC QPF serves as a useful guidance to local forecasters. (ii) Pittsburgh forecasters skillfully adjust the guidance and thereby improve the calibration of their QPF relative to the calibration of the guidance QPF. (iii) Nevertheless, the local QPF remains hampered by the overestimation bias, and its calibration lacks consistency in the same way as the guidance QPF does.

In order to further improve the calibration of local QPFs, the following steps are recommended. (i) The guidance QPF should be given a probabilistic interpretation that should be routinely verified. Verification results should be used as feedback by HPC forecasters and as interpretive filter by local forecasters when they adjust the guidance QPF. (ii) In light of the consistent annual average calibration of HPC QPF as the conditional median of the spatially averaged precipitation amount, this interpretation should be considered for official adoption. (iii) The guidance QPF and the local QPF should be given the same probabilistic interpretation, so that local forecasters can channel their skills toward improving the calibration of an estimate rather than rescaling the estimate from one interpretation to another.

Acknowledgments. This paper is based on work funded by the National Oceanic and Atmospheric Administration under Award NA67WD0486, ''Probabilistic Hydrometeorological Forecast System.'' Results reported herein were first presented at a workshop held with the senior forecasters of the Hydrometeorological Prediction Center in Camp Springs, Maryland, on 27–28 February 1997. Insightful discussions are acknowledged with William Drzal, Thomas Graziano, James Hoke, Norman Junker, Brian Korty, Theresa Rossi, Bruce Terry, Erich Wolf, and Sondra Young.

APPENDIX

Bayesian Characterization of Sampling Uncertainty

a. Posterior distribution

Let Θ denote the unknown probability of some event, where $0 < \Theta < 1$. Bayesian inference about Θ via a conjugate family of distributions proceeds as follows (Bernardo and Smith 1994). Prior information about Θ is encoded in a beta density with parameters (α', β'). A random sample of size T is collected and the number of events t is counted. The prior density of Θ is revised

based on the *sample statistics* (T, t). The resultant posterior density of Θ is beta with parameters (α, β), where

$$\alpha = \alpha' + t, \qquad \beta = \beta' + T - t.$$

These relations between the *prior parameters* (α', β') and the *posterior parameters* (α, β) suggest that prior information is essentially equivalent to a sample of size $\alpha' + \beta'$ in which α' events were counted. After sample statistics (T, t) have been collected, the total sample size is $\alpha + \beta = \alpha' + \beta' + T$, and the total count of events is $\alpha = \alpha' + t$.

Suppose that prior information is nil, and consequently the equivalent sample size is zero. Letting $\alpha' \to 0$ and $\beta' \to 0$, one obtains

$$\alpha = t, \qquad \beta = T - t. \tag{A1}$$

A beta density with these parameters yields the following expressions for the *posterior mean* and *posterior variance*:

$$E(\Theta \mid T, t) = \frac{t}{T}, \tag{A2}$$

$$\mathrm{Var}(\Theta \mid T, t) = \frac{t(T - t)}{T^2(T + 1)}. \tag{A3}$$

The posterior mean is simply the frequency estimator of probability Θ. The advantage of Bayesian inference is that it also specifies the posterior variance, which characterizes the uncertainty that remains about Θ after a particular sample has been collected.

b. Central credible interval

The posterior distribution of Θ, defined as $H(\theta) = P(\Theta \le \theta \mid T, t)$, is given by

$$H(\theta) = \frac{\Gamma(\alpha, \beta)}{\Gamma(\alpha)\Gamma(\beta)} \int_0^\theta \tau^{\alpha-1}(1 - \tau)^{\beta-1}\, d\tau,$$

where Γ is the gamma function and H is the incomplete beta function. Distribution H characterizes the uncertainty about probability Θ, given nil prior information and sample statistics (T, t). A simplified characterization of this uncertainty may be obtained that is analogous to, yet different from, a classical confidence interval.

Let θ_{100p} denote the $100p\%$ posterior fractile of Θ, defined by $P(\Theta \le \theta_{100p} \mid T, t) = p$ and such that

$$\theta_{100p} = H^{-1}(p), \qquad 0 < p < 1, \tag{A4}$$

where H^{-1} denotes the inverse of H. Then

$$C_{100p} = (\theta_{100(1-p)/2}, \theta_{100(1+p)/2}) \tag{A5}$$

is the $100p\%$ *central credible interval* of Θ. Now one may state that, given the sample statistics (T, t), there is a $100p\%$ chance that the unknown probability Θ lies within the interval C_{100p}. (One should recall that a confidence interval defined in classical statistics does not admit such an interpretation.)

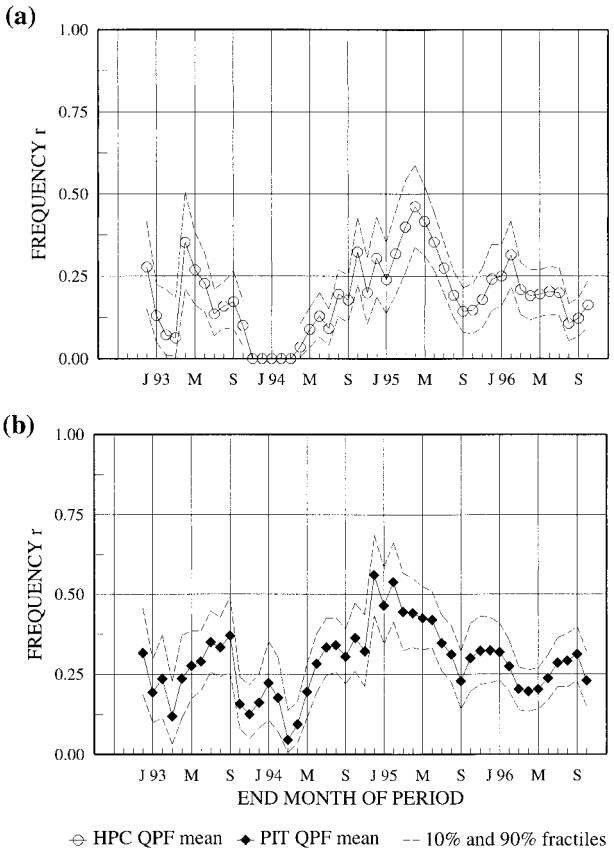

FIG. A1. Mean and 80% central credible interval of probability $P(W > X \mid X > 0)$ that verifies QPF as an exceedance fractile of W: (a) HPC QPF and (b) PIT QPF; Monongahela basin, 3-month verification periods.

c. Application to verification

In the context of the verification procedure described in section 4, probability Θ is interpreted either as $P(W > X \mid X > 0)$, when one verifies QPF as an exceedance fractile, or as $P(W > X \mid W > 0)$, when one verifies QPF as a conditional exceedance fractile. In each case, (A2) provides a basis for defining the frequency estimator of Θ as the posterior mean:

$$r = E[P(W > X \mid X > 0) \mid N, n] = \frac{n}{N}, \tag{A6}$$

$$r_0 = E[P(W > X \mid W > 0) \mid N_0, n_0] = \frac{n_0}{N_0}. \tag{A7}$$

A central credible interval can be calculated according to (A1), (A4), and (A5). The inverse H^{-1} of the incomplete beta function can be evaluated approximately via formulas from Abramowitz and Stegun (1972, p. 945).

d. Uncertainty analyses

Results of the sampling uncertainty analyses are reported for the Monongahela basin; results for the Al-

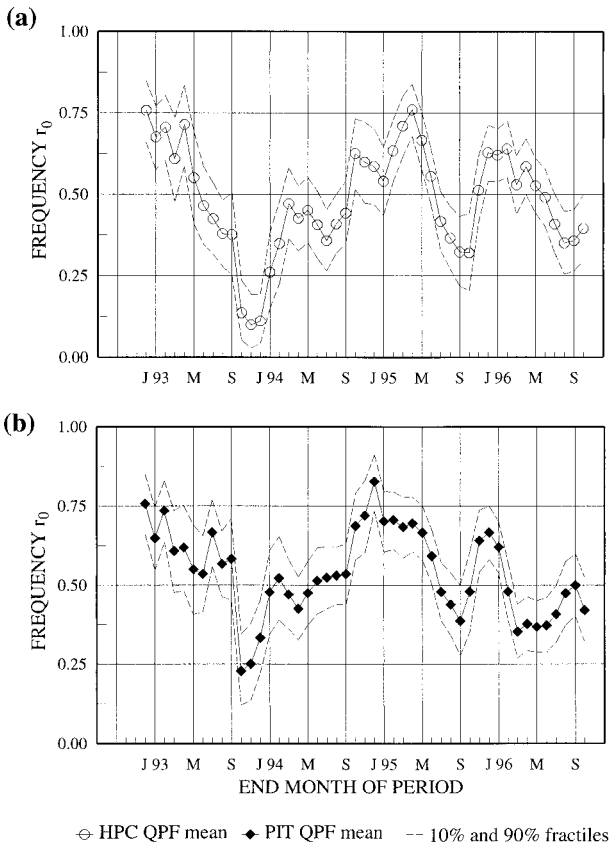

FIG. A2. Mean and 80% central credible interval of probability $P(W > X | W > 0)$ that verifies QPF as a conditional exceedance fractile of W: (a) HPC QPF and (b) PIT QPF; Monongahela basin, 3-month verification periods.

legheny basin are similar. Figures A1 and A2 show time series of the posterior means, r and r_0, and the 80% central credible intervals for HPC QPF and PIT QPF. The time series of r and r_0 are the same as those compared in Fig. 3 and analyzed in section 4. A time series of the 80% credible interval depicts the uncertainty around the time series of r, or r_0, caused by a finite sample size. The size of a sample from the 3-month verification period varies between 42 and 92, and has an average of 81 observations.

All four plots exhibit two features. First, the 80% central credible interval is not always symmetric about r, or r_0, and its width varies from one verification period to the next. The average width is 0.16 and 0.19 about r for HPC QPF and PIT QPF, respectively (Fig. A1), and 0.20 and 0.21 about r_0 for HPC QPF and PIT QPF, respectively (Fig. A2). Second, a time series of the 80% central credible interval closely traces all large fluctuations of a time series of r, or r_0; the amplitudes of these fluctuations are bigger than the width of the credible interval at any period.

In conclusion, large fluctuations of the time series of exceedance frequencies r and r_0 are not an artifact of sampling variability; they may, therefore, be attributed to the performance of the forecasters.

REFERENCES

Abramowitz, M., and I. A. Stegun, 1972: *Handbook of Mathematical Functions.* Dover, 1046 pp.

Alpert, M., and H. Raiffa, 1982: A progress report on the training of probability assessors. *Judgment Under Uncertainty: Heuristics and Biases*, D. Kahneman, P. Slovic, and A. Tversky, Eds., Cambridge University Press, 294–305.

Bernardo, J. M., and A. F. M. Smith, 1994: *Bayesian Theory.* Wiley, 586 pp.

Funk, T. W., 1991: Forecasting techniques utilized by the Forecast Branch of the National Meteorological Center during a major convective rainfall event. *Wea. Forecasting,* **6,** 548–564.

Krzysztofowicz, R., and T. R. Drake, 1993: Usage of guidance products in preparing probabilistic QPFs for river basins. NOAA Tech. Memo. NWS ER-87, 43–50. [Available from National Technical Information Service, U.S. Department of Commerce, Springfield, VA 22161.]

——, and A. A. Sigrest, 1999: Calibration of probabilistic quantitative precipitation forecasts. *Wea. Forecasting,* **14,** 427–442.

——, W. J. Drzal, T. R. Drake, J. C. Weyman, and L. A. Giordano, 1993: Probabilistic quantitative precipitation forecasts for river basins. *Wea. Forecasting,* **8,** 424–439.

Lichtenstein, S., and B. Fischhoff, 1980: Training for calibration. *Organ. Behav. Human Perform.,* **26,** 149–171.

Lindgren, B. W., 1976: *Statistical Theory.* Macmillan, 614 pp.

Murphy, A. H., and R. L. Winkler, 1974: Credible interval temperature forecasting: Some experimental results. *Mon. Wea. Rev.,* **102,** 784–794.

——, and ——, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bull. Amer. Meteor. Soc.,* **60,** 12–19.

Office of Meteorology, 1999: The modernized end-to-end forecast process for quantitative precipitation information: Hydrometeorological requirements, scientific issues, and service concepts. 188 pp. [Available from National Weather Service, 1325 East–West Highway, Silver Spring, MD 20910.]

Olson, D. A., N. W. Junker, and B. Korty, 1995: Evaluation of 33 years of quantitative precipitation forecasting at the NMC. *Wea. Forecasting,* **10,** 498–511.

Wallsten, T. S., D. V. Budescu, and R. Zwick, 1993: Comparing the calibration and coherence of numerical and verbal probability judgments. *Manage. Sci.,* **39,** 176–190.

Wernly, D. R., and L. W. Uccellini, 1998: Storm forecasting for emergency response: Current status and future directions. *Storms,* R. A. Pielke Jr. and R. A. Pielke Sr., Eds., Routledge, in press.