Calibration of Probabilistic Quantitative Precipitation Forecasts

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 6 April 1998, in final form 10 December 1998)

ABSTRACT

From 1 August 1990 to 31 July 1995, the Weather Service Forecast Office in Pittsburgh prepared 6159 *probabilistic quantitative precipitation forecasts.* Forecasts were made twice a day for 24-h periods beginning at 0000 and 1200 UTC for two river basins. This is the first in a series of articles devoted to a comprehensive verification of these forecasts. The property verified herein is *calibration*: a match between forecast probabilities and empirical frequencies of events.

Monthly time series of calibration statistics are analyzed to infer (i) trends in calibration over time, (ii) the forecasters' skill in quantifying uncertainty, (iii) the adaptability of forecasters' judgments to nonstationarities of the predictand, (iv) the possibility of reducing biases through dynamic recalibration, and (v) the potential for improving calibration through individualized training.

1. Introduction

a. Project background

Since 1990 the Eastern Region of the National Weather Service (NWS) and the University of Virginia have cooperated on the development of a methodology for producing *probabilistic quantitative precipitation forecasts* (PQPFs) for river basins. An advantage of a probabilistic forecast is that it allows the forecaster to quantify, and convey to users, the degree of uncertainty about the predictand (Murphy and Winkler 1979). This degree of uncertainty is likely to vary from occasion to occasion, but a deterministic forecast, which specifies only a single estimate, does not convey this information. The primary purpose of a PQPF is to serve as an input to a hydrologic model that will produce probabilistic forecasts of river stages. The methodology has been tested operationally at the Weather Service Forecast Office (WSFO) in Pittsburgh. Testing began on 1 May 1990 and continues to this day.

Earlier articles described the methodology (Krzysztofowicz et al. 1993) and reported initial results (Krzysztofowicz and Drake 1992). This article is the first in a series devoted to a comprehensive verification of 6159 forecasts prepared by the WSFO Pittsburgh as part of routine operational forecasting during 5 yr between 1 August 1990 and 31 July 1995.

b. River basins

The predictand is the 24-h *basin average precipitation amount.* Forecasts are made twice a day for 24-h periods beginning at 0000 and 1200 UTC 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 and (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, these basins have somewhat different weather characteristics.

c. Testing setup

The optimal implementation of the PQPF methodology should take the form of a human–computer system (Krzysztofowicz et al. 1993). Through an interactive interface, the computer should support judgmental tools for guiding the forecaster's reasoning and estimation procedures (mostly graphical) for eliciting judgmental estimates from the forecaster. The actual implementation of the methodology during the testing has been rather

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

© 1999 American Meteorological Society

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

primitive. It was manual during the first year and later became computerized, but only partly and without an interactive interface. Because proper tools and procedures are essential to maximizing human judgmental performance, it is safe to say that the testing has been performed in a suboptimal setup.

d. Verification samples

During the 5 yr from 1 August 1990 to 31 July 1995, a total of 6159 joint observations of the forecast and predictand have been collected, 3090 for the Monongahela basin and 3069 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, Ohio, 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 (NOAA 1972).

Forecasts for each basin are verified 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 58 values beginning in October 1990 and ending in July 1995. Because of the 2-month stagger of the verification periods, any observation from month 3 through 58 of the 5-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 period to period because of missing observations—a predicament typical in operational forecasting. The minimum, average, and maximum sample sizes are, respectively, 70, 153, and 182 for each of the basins.

2. Forecast and verification

a. Judgmental probabilistic forecast

Let ω denote the 24-h *basin average precipitation amount.* At the forecast time, this amount is uncertain and thus is treated as a random variable, denoted W. Ideally, the PQPF would specify the entire exceedance function of W, as shown by a broken line in Fig. 1. For any fixed $\omega \geq 0$, the exceedance function gives $P(W > \omega)$, the probability of observing a precipitation amount W larger than ω.

An operational approximation to the exceedance function is obtained by having a forecaster judgmentally assess three *exceedance fractiles* of W defined as follows. Let p denote a probability number such that $p \in \{0.75, 0.50, 0.25\}$. The $100p$% exceedance fractile of W is an estimate x_{100p} such that the exceedance probability is

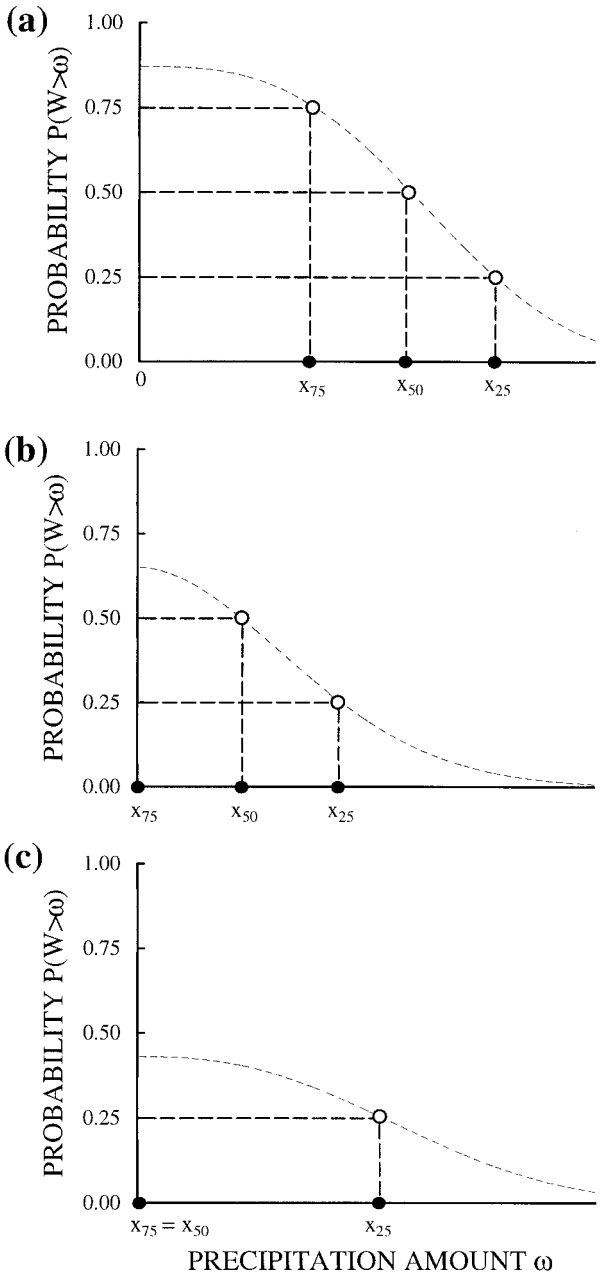

FIG. 1. Probabilistic quantitative precipitation forecast, specified by three exceedance fractiles (x_{75}, x_{50}, x_{25}), and an interpolated exceedance function (dashed line). The three examples (a), (b), and (c) correspond to orders R0, R1, and R2 of the exceedance fractiles.

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

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 that is three times more likely to be exceeded than not; that is, $P(W > x_{75}) = 0.75$. The *25% exceedance fractile* x_{25} is an estimate that is three times less

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

TABLE 1. Examples of PQPFs prepared for 24-h periods beginning at 1200 UTC for the Monongahela basin.

Date	Exceedance fractiles (in.)			Implied range of PoP
	x_{75}	x_{50}	x_{25}	
23 Feb 1992	0.20	0.36	0.45	R0
8 May 1992	0.00	0.13	0.26	R1
25 Jul 1992	0.00	0.00	0.30	R2

likely to be exceeded than not; that is, $P(W > x_{25}) = 0.25$. Exceedance fractiles x_{75} and x_{25} define a *50% credible interval* about the median; that is, $P(x_{75} < W \leq x_{25}) = 0.50$. It is this credible interval that quantifies, and conveys to the user, the uncertainty about W on a particular occasion. The larger the uncertainty in the forecaster's judgment, the wider the credible interval. [A detailed explanation of the assessment protocol can be found in Krzysztofowicz et al. (1993, appendix A).]

The exceedance fractile x_{100p} is positive if and only if the forecast *probability of precipitation* occurrence over the basin, PoP for short, is such that $P(W > 0) > p$. The three values of p used operationally partition the domain of PoP into four ranges, denoted by Ri, $i = 0$, 1, 2, 3, and specified as follows:

$$\text{R0:} \quad 0.75 < P(W > 0) \leq 1, \tag{2a}$$

$$\text{R1:} \quad 0.50 < P(W > 0) \leq 0.75, \tag{2b}$$

$$\text{R2:} \quad 0.25 < P(W > 0) \leq 0.50, \tag{2c}$$

$$\text{R3:} \quad 0 \leq P(W > 0) \leq 0.25. \tag{2d}$$

Each range of PoP is in one-to-one correspondence with an order of the exceedance fractiles:

$$\text{R0} \Leftrightarrow 0 < x_{75} < x_{50} < x_{25}, \tag{3a}$$

$$\text{R1} \Leftrightarrow 0 = x_{75} < x_{50} < x_{25}, \tag{3b}$$

$$\text{R2} \Leftrightarrow 0 = x_{75} = x_{50} < x_{25}, \tag{3c}$$

$$\text{R3} \Leftrightarrow 0 = x_{75} = x_{50} = x_{25}. \tag{3d}$$

For example, if in the forecaster's judgment the PoP for a basin is in the range R1, then $x_{75} = 0$, and the forecaster should assess the other two exceedance fractiles such that $0 < x_{50} < x_{25}$. And vice versa; if the forecaster assessed the exceedance fractiles such that $0 = x_{75} < x_{50} < x_{25}$, then he implies that the PoP for the basin is in the range R1.

Table 1 reports three PQPFs for the Monongahela basin, and Fig. 1 shows the interpolated exceedance functions. The normative interpretation of these PQPFs is as follows. On 23 February 1992, the probability of precipitation occurrence over the basin was between 0.75 and 1; there was a 75% chance that the 24-h basin average precipitation amount would exceed 0.20 in., a 50% chance that it would exceed 0.36 in., and a 25% chance that it would exceed 0.45 in.; also, there was a 50% chance that the amount would be between 0.20 in. and 0.45 in. On 8 May 1992, the PoP was between 0.50 and 0.75; there was a 50% chance that 0.13 in. would be exceeded and a 25% chance that 0.26 in. would be exceeded. On 25 July 1992, the PoP was between 0.25 and 0.50, and there was a 25% chance that 0.30 in. would be exceeded.

b. Decision-theoretic verification

The verification reported herein is founded on principles derived from Bayesian decision theory (Krzysztofowicz 1992, 1996). The underlying viewpoint is that of potential users of forecasts—the decision makers who have to bear all consequences of actions taken in response to forecasts. From this viewpoint, PQPFs should be verified with respect to two properties: calibration and informativeness.

The purpose of informativeness measures is to tell us whether the economic value has increased or decreased from one set of forecasts to another, and in relation to worthless forecasts and perfect forecasts. A subsequent article will be devoted to the informativeness of PQPFs.

In order that users can take a forecast at its face value, the *normative interpretation* of the forecast, as defined by (1)–(3), should be consistently adhered to over time. Consistent interpretability of forecasts requires good calibration (Murphy and Winkler 1974; Lichtenstein and Fischhoff 1980; Wallsten et al. 1993). The purpose of calibration measures is to tell us whether or not the normative interpretation has been maintained in a set of forecasts.

Section 3 describes statistical measures for verifying the calibration of PQPFs. Section 4 presents results of calibration analyses for the two river basins. Section 5 introduces the concept of dynamic recalibration and demonstrates its effectiveness. Section 6 reports calibration analyses for sets of PQPFs prepared by individual forecasters; the aim of these analyses is to uncover any learning effects and training needs. Section 7 summarizes the findings.

3. Calibration measures

a. Empirical verification

There is an established empirical procedure for verifying the calibration property of probabilistic forecasts of a predictand whose probability distribution is strictly continuous (Murphy and Winkler 1974, 1979; Alpert and Raiffa 1982); an example of such a predictand is temperature. Precipitation amount W has a mixed (discrete-continuous) probability distribution because $P(W = 0)$ may be positive. This fact substantially complicates the verification problem for PQPF. To the best of our knowledge, this problem has not received a rigorous treatment in the literature. Our methodology, developed from the viewpoint of Bayesian decision theory, is presented in appendix A. It leads to an empirical verification procedure that consists of two analyses.

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

b. Probability of precipitation

The purpose of the following analysis is to verify the normative correspondence (2)–(3) between the order of the exceedance fractiles and the implied PoP. Toward this end, define $P(W > 0 | \mathrm{R}i)$, the probability of precipitation on those occasions on which the order of the exceedance fractiles is Ri. It is said that the PoP implied by a PQPF is *well calibrated* if for every $i = 0, 1, 2, 3$, the probability $P(W > 0 | \mathrm{R}i)$ falls within the range Ri.

The empirical verification of a set of forecasts requires two counts: M_i—the number of forecasts specifying order Ri, and m_i—the number of forecasts that specify order Ri and are followed by the occurrence of precipitation, $\omega > 0$. Therefrom, the frequency estimator of probability $P(W > 0 | \mathrm{R}i)$ is

$$q_i = \frac{m_i}{M_i}, \qquad i = 0, 1, 2, 3. \tag{4}$$

The PoP implied by a PQPF from the given set of forecasts is said to be well calibrated if for every $i = 0, 1, 2, 3$, the frequency q_i falls within the range Ri.

Within the Bayesian inference framework presented in appendix B, estimator (4) constitutes the posterior mean of probability $P(W > 0 | \mathrm{R}i)$. The posterior standard deviation of this probability is

$$u_i = \left[\frac{m_i(M_i - m_i)}{M_i^2(M_i + 1)}\right]^{1/2}, \qquad i = 0, 1, 2, 3. \tag{5}$$

The standard deviation u_i offers a measure of the sampling uncertainty about q_i. As the sample size increases, $M_i \to \infty$, the uncertainty decreases, $u_i \to 0$.

c. Exceedance fractiles

The purpose of the following analysis is to verify the normative interpretation of the exceedance fractiles (1). Toward this end, define $P(W > X_{100p} | X_{100p} > 0)$, the probability of observing precipitation amount W larger than X_{100p} on those occasions on which a positive exceedance fractile $X_{100p} > 0$ is assessed. It is said that the exceedance fractiles specified by a PQPF are *well calibrated* if $P(W > X_{100p} | X_{100p} > 0) = p$ for $p = 0.75, 0.50, 0.25$.

The empirical verification of a set of forecasts requires two counts: N_{100p}—the number of forecasts specifying fractile $x_{100p} > 0$, and n_{100p}—the number of forecasts that specify fractile $x_{100p} > 0$ and are followed by an observation of precipitation amount $\omega > x_{100p}$. It should be noted that $N_{75} = M_0$, $N_{50} = M_0 + M_1$, and $N_{25} = M_0 + M_1 + M_2$. Therefrom, the frequency estimator of probability $P(W > X_{100p} | X_{100p} > 0)$ is

$$r_{100p} = \frac{n_{100p}}{N_{100p}}, \qquad p = 0.75, 0.50, 0.25. \tag{6}$$

The exceedance fractiles specified by a PQPF from the given set of forecasts are said to be well calibrated if $r_{100p} \simeq p$ for $p = 0.75, 0.50, 0.25$.

Within the Bayesian inference framework presented in appendix B, estimator (6) constitutes the posterior mean of probability $P(W > X_{100p} | X_{100p} > 0)$. The posterior standard deviation of this probability is

$$v_{100p} = \left[\frac{n_{100p}(N_{100p} - n_{100p})}{N_{100p}^2(N_{100p} + 1)}\right]^{1/2}, \qquad p = 0.75, 0.50, 0.25. \tag{7}$$

The standard deviation v_{100p} offers a measure of the sampling uncertainty about r_{100p}. As the sample size increases, $N_{100p} \to \infty$, the uncertainty decreases, $v_{100p} \to 0$.

A measure summarizing the calibration of all three fractiles (x_{75}, x_{50}, x_{25}) is the *calibration score*, CS, defined as the root-mean-square difference between the empirical frequency r_{100p} and the forecast probability p:

$$\mathrm{CS} = \left\{\frac{1}{3}[(r_{75} - 0.75)^2 + (r_{50} - 0.50)^2 + (r_{25} - 0.25)^2]\right\}^{1/2}. \tag{8}$$

The score is bounded, $0 \le \mathrm{CS} < 0.677$, with CS = 0 being the best. The upper bound arises when $r_{75} = 0$, $r_{50} = 1$, and $r_{25} = 1$. The CS is similar to the calibration component of the Brier score (DeGroot and Fienberg 1983).

4. Calibration of forecasts

a. Behavior of the predictand

The purpose of the following analysis is (i) to characterize behavior of the predictand alone and, possibly, (ii) to establish a basis for explaining the verification results. Toward this end, observations of W for the 24 h beginning at 1200 UTC were extracted from the 3-month verification sample and were used to estimate the following elements: (i) the probability of precipitation $\pi = P(W > 0)$ and (ii) the 75%, 50%, and 25% exceedance fractiles of W, conditional on precipitation occurrence; denoted $\omega_{100p|0}$, the conditional exceedance fractile is defined by $P(W > \omega_{100p|0} | W > 0) = p$ for $p = 0.75, 0.50, 0.25$. Together, these elements characterize the distribution of the 24-h basin average precipitation amount on any day during the 3-month verification period.

The four elements (the probability of precipitation π and the conditional exceedance fractiles $\omega_{75|0}$, $\omega_{50|0}$, and $\omega_{25|0}$) provide a more complete characterization of the distribution of precipitation amount than the three elements (the unconditional exceedance fractiles x_{75}, x_{50}, and x_{25}). The three-element characterization is used in the PQPF in order to keep the number of assessments small. The four-element characterization is used in the local climatic guidance (Krzysztofowicz and Sigrest

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

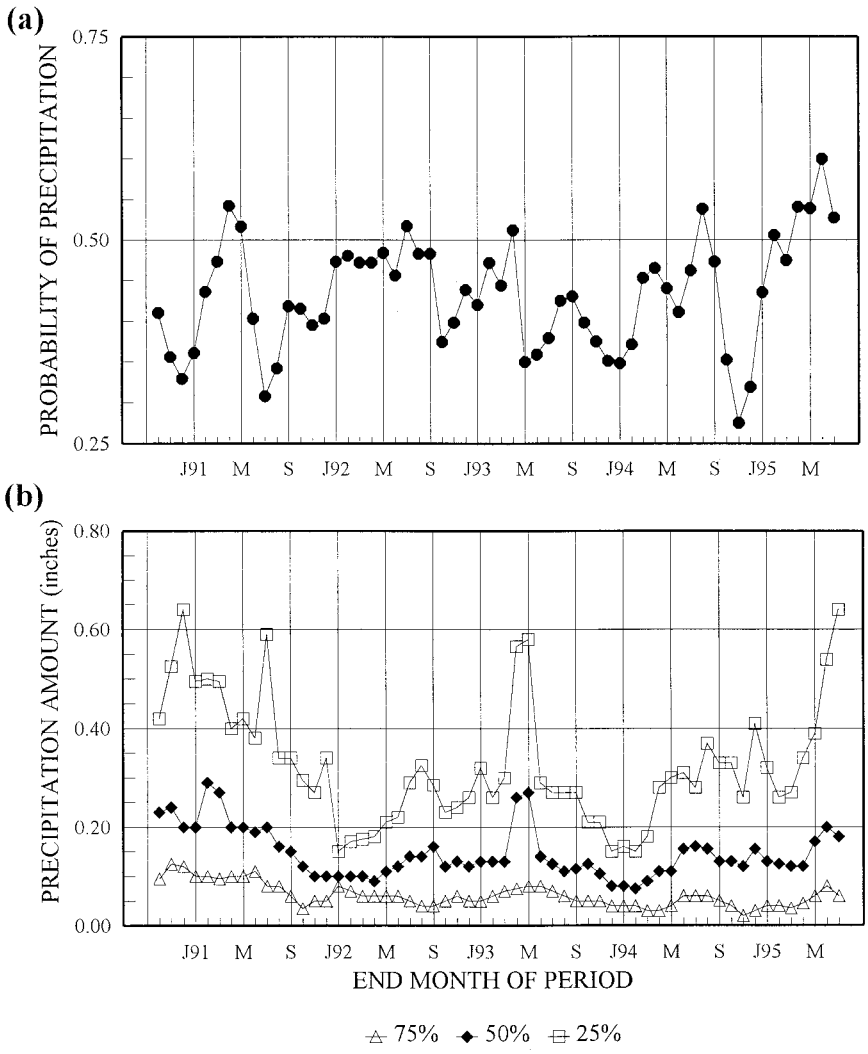

FIG. 2. Statistics of the 24-h basin average precipitation amount observed from 1200 UTC during the 3-month verification periods: (a) probability of precipitation occurrence π, and (b) exceedance fractiles of the amount, conditional on precipitation occurrence ($\omega_{75|0}$, $\omega_{50|0}$, $\omega_{25|0}$); Monongahela basin, Aug 1990–Jul 1995.

1997) as one of several products that may be requested by forecasters. Figures 2 and 3 display the time series of the four elements (π, $\omega_{75|0}$, $\omega_{50|0}$, $\omega_{25|0}$) for each of the basins during the pentad August 1990–July 1995.

The first striking observation is that the within-year seasonality of precipitation was weak and dominated by multiyear trends. In the Monongahela basin (Fig. 2), the probability of precipitation fluctuated within a narrow range, 0.34–0.54, from August 1991 to October 1994, but with larger amplitudes before and after. The distribution of the precipitation amount per rainy day was nonstationary. The time series of the conditional median amount and the variability of the amount (as measured by the difference between the 25% and 75% conditional exceedance fractiles) reveal three trend periods. From the onset of the pentad to January 1992 the median and the variability were decreasing, then went through a 2-yr cycle of an increase followed by a decrease, and finally in March 1994 began to increase again until the end of the pentad.

In the Allegheny basin (Fig. 3), the probability of precipitation was decreasing until June 1991, then entered an increasing trend with seasonal fluctuations throughout the rest of the pentad. The distribution of the precipitation amount per rainy day was nonstationary and exhibited trends generally similar to those observed in the Monongahela basin.

In conclusion, the probability distribution of the 24-h precipitation amount in the two basins during the pentad August 1990–July 1995 lacked any decisive annual cycle associated with seasons and instead exhibited a nonstationarity with multiyear trends. Consequently, climatic PQPFs estimated for each 3-month verification period under the assumption of the year-to-year sta-

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

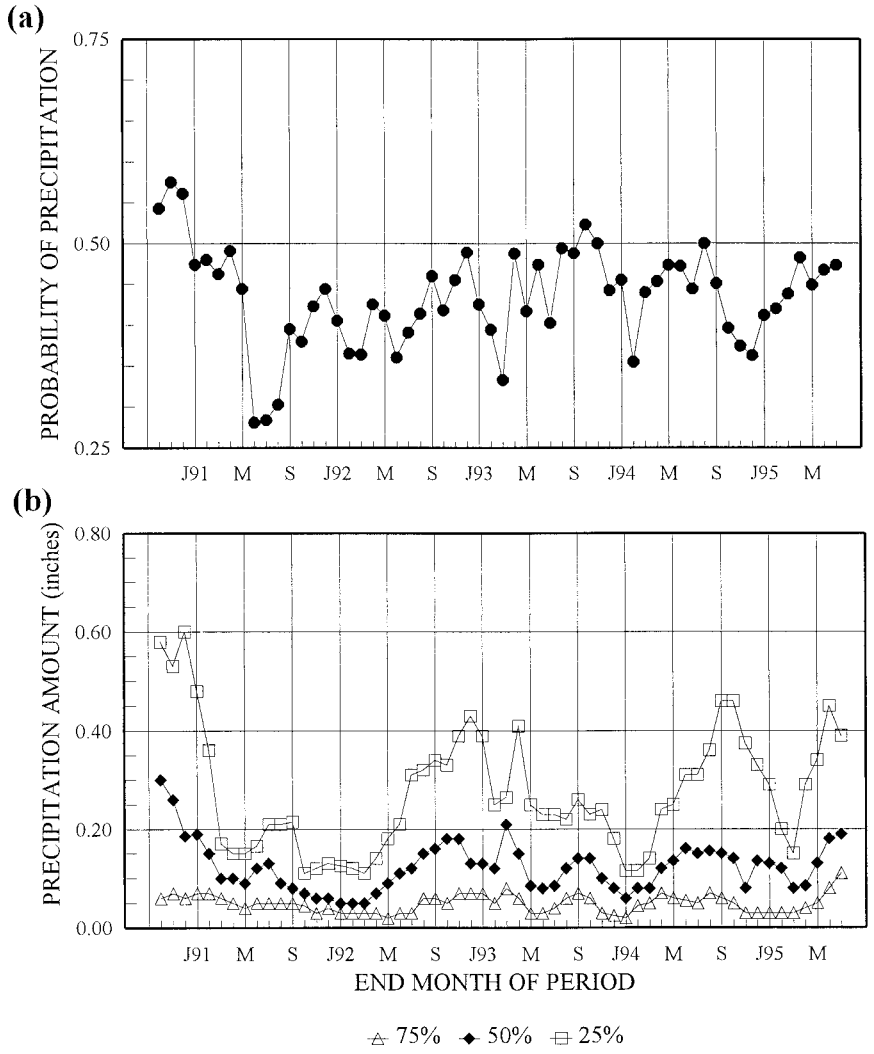

FIG. 3. Statistics of the 24-h basin average precipitation amount observed from 1200 UTC during the 3-month verification periods: (a) probability of precipitation occurrence π, and (b) exceedance fractiles of the amount, conditional on precipitation occurrence ($\omega_{75|0}$, $\omega_{50|0}$, $\omega_{25|0}$); Allegheny basin Aug 1990–Jul 1995.

tionarity would provide rather poor predictions of W during this particular pentad. One may also hypothesize that nonstationarities caused by a trend, turns of the trend, or sudden breaks in the trend would present a particular challenge to the forecasters.

b. Calibration of probability of precipitation

Calibration of the PoP implied by a PQPF is shown in Figs. 4 and 5. Frequencies q_0 and q_3 fall within their normative ranges R0 and R3, respectively, most of the time, implying generally good skill in judging the high and low PoPs for the basins. The averages of the time series of q_0 and q_3 are reported in Table 2. The few large miscalibrations can usually be associated with sudden breaks in a trend of the predictand. For the Monongahela basin, q_0 dips below 0.75 in July 1991 (Fig. 4a) just when the probability of precipitation π falls sharply below the trend (Fig. 2a) while the 25% conditional exceedance fractile $\omega_{25|0}$ skyrockets (Fig. 2b); the departure of q_3 above 0.25 in April 1993 (Fig. 4a) coincides with a sudden jump of both the conditional median $\omega_{50|0}$ and the 25% conditional exceedance fractile $\omega_{25|0}$ of the actual amount (Fig. 2b). Likewise, for the Allegheny basin, the largest departure of q_0 below 0.75 in February–March 1993 (Fig. 5a) coincides with a sudden fall of the probability of precipitation π below the trend (Fig. 3a) and a simultaneous break in the trend of the 25% conditional exceedance fractile $\omega_{25|0}$ of the actual amount (Fig. 3b).

Frequencies q_1 and q_2 are plotted in Figs. 4b and 5b. The averages of the time series are 0.51 and 0.40 for the Monongahela basin, and 0.55 and 0.39 for the Allegheny basin (Table 2). The averages fall within their

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

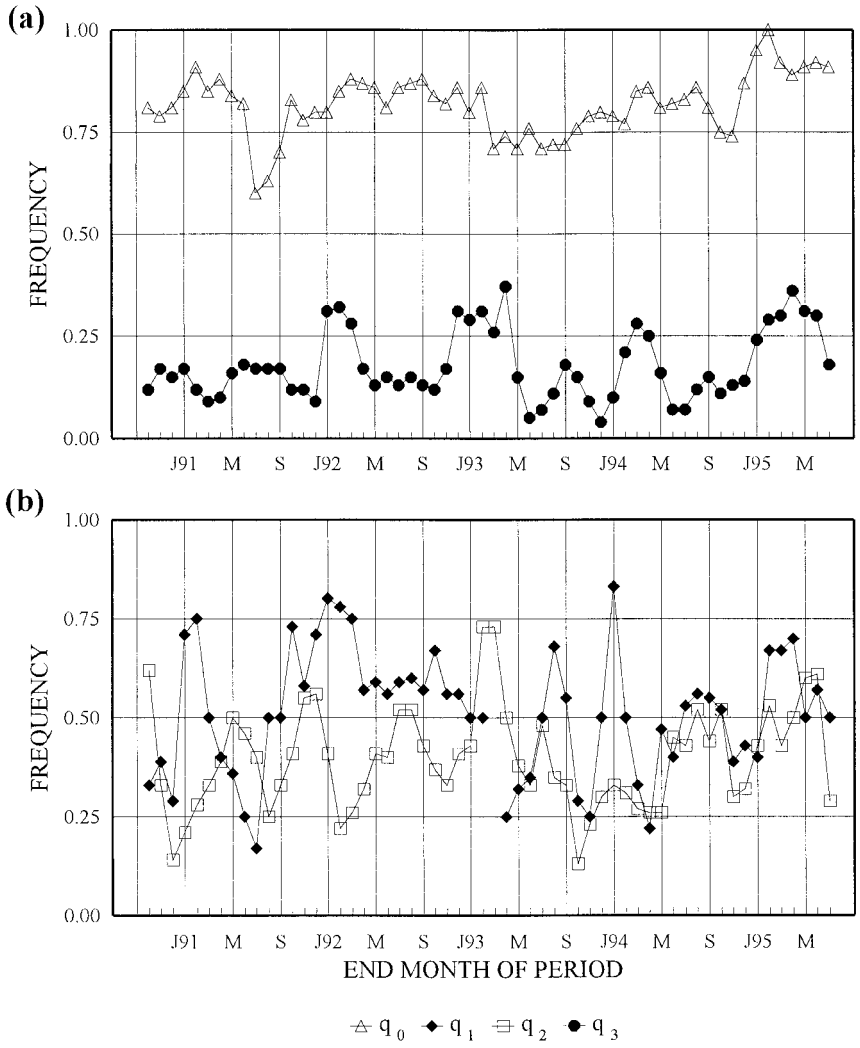

FIG. 4. Calibration of the PoP implied by a PQPF: (a) frequencies q_0 and q_3, and (b) frequencies q_1 and q_2; Monongahela basin, 3-month verification periods.

normative ranges R1 and R2, respectively. However, the time series themselves show several large fluctuations outside their normative ranges, especially the time series of q_1. These fluctuations may be explained by two causes. The first cause may be the nonstationarity of the distribution of the predictand, which at times made it difficult for the forecasters to maintain good calibration, as discussed earlier. The second cause may be the small sample sizes M_1 and M_2 (Table 2). Therefore, the calibration of the PoP within the ranges R1 and R2 should not be inferred from individual points but rather from the averages of the time series.

The small sample sizes M_1 and M_2, relative to the sample sizes M_0 and M_3, have yet another implication. Imagine a semi-clairvoyant who perfectly detects whether or not precipitation will occur but is unsure about the amount. Such a clairvoyant would use only the fractile orders R0 and R3, and the implied PoP would be well calibrated; specifically, $q_0 = 1$ and $q_3 = 0$, whereas q_1 and q_2 would be undefined. Vis-a-vis this benchmark, the finding that the fractile orders R0 and R3 are specified more frequently than the fractile orders R1 and R2 (Table 2), and the implied PoP is well calibrated, attests to the forecasters' skill in detecting rainy versus nonrainy days.

In conclusion, good calibration of the PoP has been maintained most of the time within the ranges R0 and R3, and on the average within the ranges R1 and R2. Thus overall, the forecasters well judged when to specify a particular order of the exceedance fractiles. The forecasters have also been able to adapt their judgments to nonstationarities of the predictand such as trends and slow turns of a trend; however, sudden breaks in a trend apparently caused difficulties in adapting judgments as they were followed by periods of poor calibration.

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

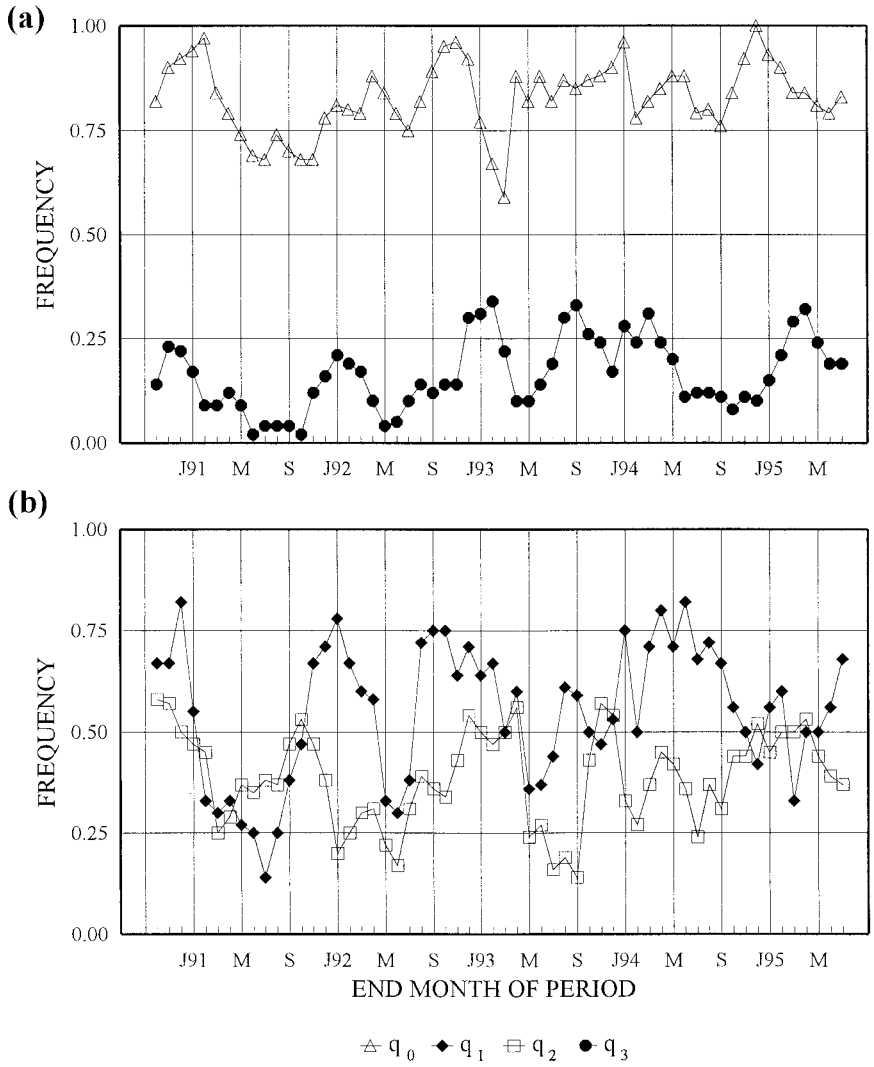

FIG. 5. Calibration of the PoP implied by a PQPF: (a) frequencies q_0 and q_3, and (b) frequencies q_1 and q_2; Allegheny basin, 3-month verification periods.

c. Calibration of exceedance fractiles

Calibration of the exceedance fractiles specified by a PQPF is shown in Figs. 6 and 7. The figures display the time series of frequencies r_{75}, r_{50}, r_{25}, and the calibration score CS. The averages of the time series are reported in Table 3. (Unlike the conditional exceedance fractiles $\omega_{100p|0}$ of the predictand analyzed in section 4a, the exceedance fractiles x_{100p} specified by a PQPF are unconditional, and they are verified as such.)

TABLE 2. Averages of the time series of the sample sizes M_i, frequencies q_i, and standard deviations u_i (i = 0, 1, 2, 3), which verify calibration of the PoP implied by a PQPF.

		Index i of the PoP range Ri			
River basin	Statistic	0	1	2	3
Monongahela	M_i	50	13	24	67
	q_i	0.82	0.51	0.40	0.18
	u_i	0.05	0.14	0.10	0.05
Allegheny	M_i	46	15	25	67
	q_i	0.83	0.55	0.39	0.17
	u_i	0.05	0.12	0.10	0.04

For the Monongahela basin, Fig. 6a reveals that the exceedance fractile x_{25} was well calibrated oftentimes, whereas the exceedance fractiles x_{50} and x_{75} were well calibrated only a few times. In 56 out of 58 cases, $r_{50} < 0.50$ and $r_{75} < 0.75$, implying a bias toward overestimating x_{50} and x_{75}. For example, consider the verification period ending in March 1991. The forecast probability of a precipitation amount higher than x_{50} is always 0.50. Yet during the verification period, the frequency of a precipitation amount higher than x_{50} was only $r_{50} = 0.39$. In order to match the empirical frequency with the forecast probability, one may say in

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

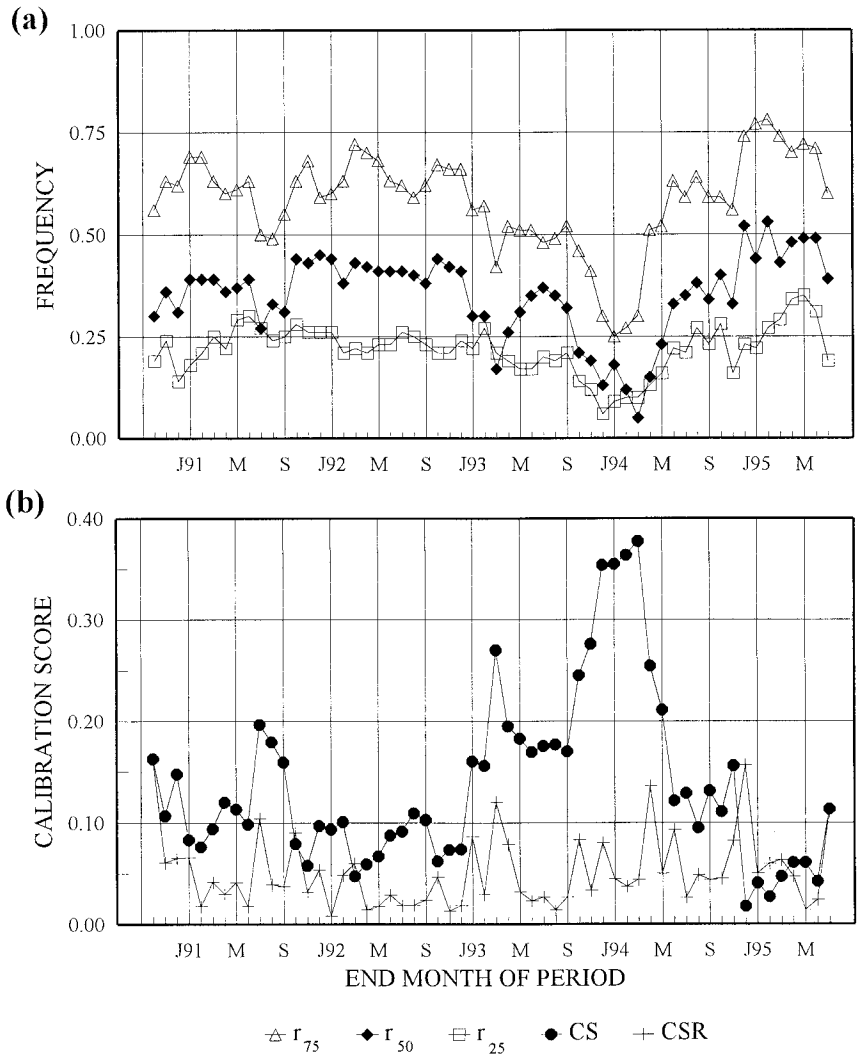

FIG. 6. Calibration of the exceedance fractiles of a PQPF: (a) frequencies r_{75}, r_{50}, r_{25}, and (b) calibration score CS and calibration score after dynamic recalibration CSR; Monongahela basin, 3-month verification periods.

retrospect that, on the average, the median x_{50} should have been lower. Likewise, the exceedance fractile x_{75} was overestimated, on the average, as the frequency of a precipitation amount higher than x_{75} was only $r_{75} = 0.63$. On the other hand, the exceedance fractile x_{25} was well calibrated as $r_{25} = 0.25$.

Normatively, the difference between frequencies should be $r_{75} - r_{25} = 0.50$. Figure 6a reveals that most of the time $r_{75} - r_{25} < 0.50$. This implies that the 50% credible interval (x_{75}, x_{25}) was too narrow, on the average. In other words, the forecasters were overconfident in judging how far from x_{50} the actual amount may fall with probability 0.50.

The time series of the CS, plotted in Fig. 6b, summarizes the overall calibration of the exceedance fractiles during the pentad. There were three epochs. From October 1990 to December 1992, the CS trended downward, which indicates improving calibration. This trend was briefly interrupted in July 1991 when a sudden jump in the variability of the predictand (Fig. 2b) apparently caused the forecasters some difficulty. On the other hand, a sudden decrease in the variability of the predictand in January 1992 and the subsequent turn of the trend (Figs. 2a and 2b) were handled by the forecasters remarkably well as the CS remained low. From January 1993 to March 1994, the CS trended upward, indicating substantial deterioration of the calibration. Its cause is unknown. A decrease of both the median and the variability of the actual amount (Fig. 2b) may be partly responsible for this miscalibration because all three exceedance fractiles were substantially overestimated, on the average, during this epoch (Fig. 6a). From April 1994 to July 1995, the CS trended downward (Fig. 6b), while both the median and the variability of the actual

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

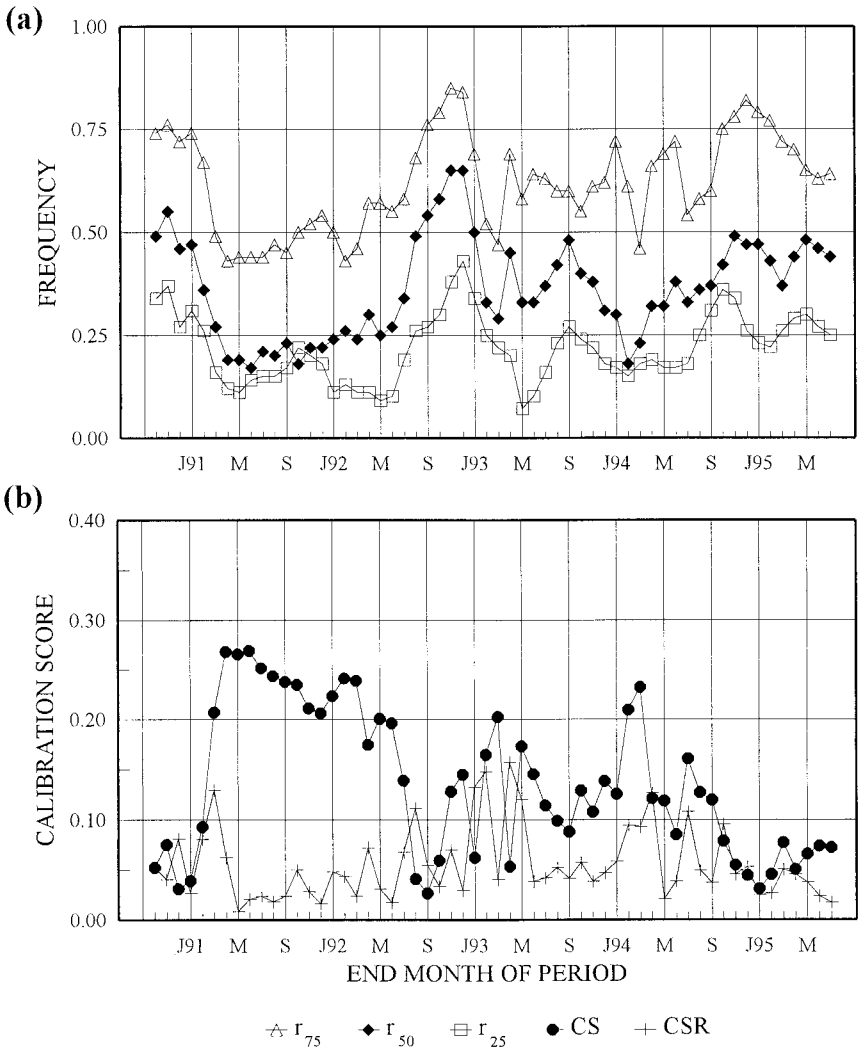

FIG. 7. Calibration of the exceedance fractiles of a PQPF: (a) frequencies r_{75}, r_{50}, r_{25}, and (b) calibration score CS, and calibration score after dynamic recalibration CSR; Allegheny basin, 3-month verification periods.

amount trended upward (Fig. 2b). Apparently, the forecasters were able to correctly adapt their judgments from period to period as dictated by the nonstationarity of the predictand.

TABLE 3. Averages of the time series of the sample sizes N_{100p}, frequencies r_{100p}, standard deviations υ_{100p} (p = 0.75, 0.50, 0.25), and the CS, which verify calibration of the exceedance fractiles of a PQPF.

		Exceedance probability p			
River basin	Statistic	0.75	0.50	0.25	CS
Monongahela	N_{100p}	50	63	87	
	r_{100p}	0.58	0.35	0.22	0.14
	υ_{100p}	0.07	0.06	0.04	
Allegheny	N_{100p}	46	61	85	
	r_{100p}	0.62	0.36	0.22	0.14
	υ_{100p}	0.07	0.06	0.04	

For the Allegheny basin, Fig. 7a reveals biases similar to those seen earlier: (i) most often $r_{75} < 0.75$, $r_{50} < 0.50$, and $r_{25} < 0.25$, indicating that the exceedance fractiles were overestimated, on the average, and (ii) most often $r_{75} - r_{25} < 0.50$, indicating that the 50% credible interval was too narrow, on the average. Interestingly, the time series of frequencies r_{75}, r_{50}, and r_{25} (Fig. 7a) follow the pattern of the conditional exceedance fractiles of the actual amounts (Fig. 3b). Generally, calibration was better when both the median and the variability of the actual amount were high, while it was worse when both the median and the variability of the actual amount were low.

The time series of the calibration score CS, plotted in Fig. 7b, shows a drastic deterioration of the overall calibration of the exceedance fractiles around March 1991, February 1993, and February 1994, just when the

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

25% conditional exceedance fractile $\omega_{25|0}$ of the actual amount suddenly decreased (Fig. 3b) implying reduced variability of the predictand. However, the overall trend in the CS during the pentad has been downward. Thus the calibration has generally been improving despite nonstationarity of the predictand.

The overall conclusions are as follows: (i) The calibration of the exceedance fractiles during any 3-month verification period has been different for each river basin. This suggests that calibration depends upon some characteristics of the local weather regime which now and again elude the forecasters' skills. (ii) The average calibration during the pentad has been identical for each basin. This suggests that, in the long run, the forecasters' skills are independent of the characteristics of these particular basins, such as location, size, and elevation. (iii) The 25% exceedance fractile x_{25} has been calibrated better than the other two fractiles. (iv) There has been a bias toward overestimating the exceedance fractiles, especially x_{75} and x_{50}. (v) There has been a bias toward overconfidence as the 50% credible interval (x_{75}, x_{25}) has been too narrow, on the average. (vi) There was a hypothesis at the onset of the test that the calibration may consistently deteriorate during summer, when precipitation becomes more isolated in nature and more difficult to forecast. The analyses presented herein do not support this hypothesis. Factors other than seasonal weather regime seem to have dominated the forecaster's ability to maintain good calibration. (vii) The calibration has been better during periods with high median actual amounts than during periods with low median actual amounts. (viii) The forecasters have been able to adapt their judgments to nonstationarities of the predictand such as trends and slow turns of a trend; however, sudden breaks in a trend, especially decreases in both the median and the variability of the actual amount, often have led to drastic deterioration of the calibration. (ix) The general trend during the pentad has been toward improved calibration (except for an unexplained huge deterioration registered in the Monongahela basin from January 1993 to March 1994).

5. Recalibration of forecast

When miscalibration of the exceedance fractiles is consistent or exhibits a trend, as in Figs. 6a and 7a, it may be possible to improve calibration by recalibrating the probabilities attached to the exceedance fractiles. To test such a possibility, a *dynamic recalibration* is performed as follows. Let m be the index of months, $m = 1, 2, \ldots,$ and let $r_{100p}(m)$ denote the frequency estimated via (6) based on a sample from the verification period ending in month m. Suppose that during month $m \in \{2, 3, \ldots\}$ the forecaster prepares PQPFs in the same manner as in the previous months, but a user of forecasts interprets x_{100p} as an estimate of W such that the exceedance probability is

$$P(W > x_{100p} \mid x_{100p} > 0) = r_{100p}(m - 1), \quad p = 0.75, 0.50, 0.25. \quad (9)$$

In other words, the stated forecast probability p is replaced by the empirical frequency $r_{100p}(m - 1)$ from the preceding verification period. This replacement is called recalibration.

The exceedance fractiles specified by a PQPF from the set of forecasts prepared during the verification period ending in month m are *dynamically well calibrated* if $r_{100p}(m) \simeq r_{100p}(m - 1)$ for $p = 0.75, 0.50, 0.25$. The *calibration score after dynamic recalibration,* CSR(m), is defined analogously to (8) as the root-mean-square difference between the empirical frequency $r_{100p}(m)$ and the recalibrated forecast probability $r_{100p}(m - 1)$:

$$\mathrm{CSR}(m) = \left\{\frac{1}{3}\sum_{p}[r_{100p}(m) - r_{100p}(m-1)]^2\right\}^{1/2}. \quad (10)$$

The score is bounded, $0 \le \mathrm{CSR}(m) < 1$, with $\mathrm{CSR}(m) = 0$ being the best. It is calculated for $m = 1, 2, \ldots$ with the initial probabilities $r_{100p}(0) = p$.

The dynamic recalibration offers an improvement if $\mathrm{CSR}(m) < \mathrm{CS}(m)$, where CS($m$) is the calibration score computed according to (8) using differences $r_{100p}(m) - p$. Because the improvement is not guaranteed, the performance of the dynamic recalibration should be continuously monitored.

For each basin, Fig. 6b or Fig. 7b shows the time series of the calibration score after dynamic recalibration $\{\mathrm{CSR}(m): m = 1, \ldots, 58\}$, which should be compared with the time series of the original calibration score $\{\mathrm{CS}(m): m = 1, \ldots, 58\}$. The inequality $\mathrm{CSR}(m) < \mathrm{CS}(m)$ holds in 51 months for the Monongahela basin and in 49 months for the Allegheny basin. The average scores are 0.05 versus 0.14 for the Monongahela basin, and 0.06 versus 0.14 for the Allegheny basin. Clearly, the dynamic recalibration has led to an improvement over the original calibration.

Scatterplots of the empirical frequency $r_{100p}(m)$ versus the recalibrated forecast probability $r_{100p}(m - 1)$ for $m = 2, 3, \ldots, 58$ and $p = 0.75, 0.50, 0.25$, called *dynamic calibration diagrams,* are shown in Fig. 8 for the Monongahela basin. The diagrams reveal that the points cluster along the diagonal; ideally, they would lie on the diagonal. However, without the dynamic recalibration, the original calibration diagrams would show these points lying on three vertical lines: all $r_{75}(m)$ would have abscissa 0.75 (Fig. 8a), all $r_{50}(m)$ would have abscissa 0.50 (Fig. 8b), and all $r_{25}(m)$ would have abscissa 0.25 (Fig. 8c).

In conclusion, the concept of recalibration should be further studied as a possible means of ensuring that operational PQPFs are well calibrated and hence can be interpreted consistently over time by users.

6. Individual calibration

The calibration of PQPFs issued by an office depends upon skills of individual forecasters. It is therefore in-

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

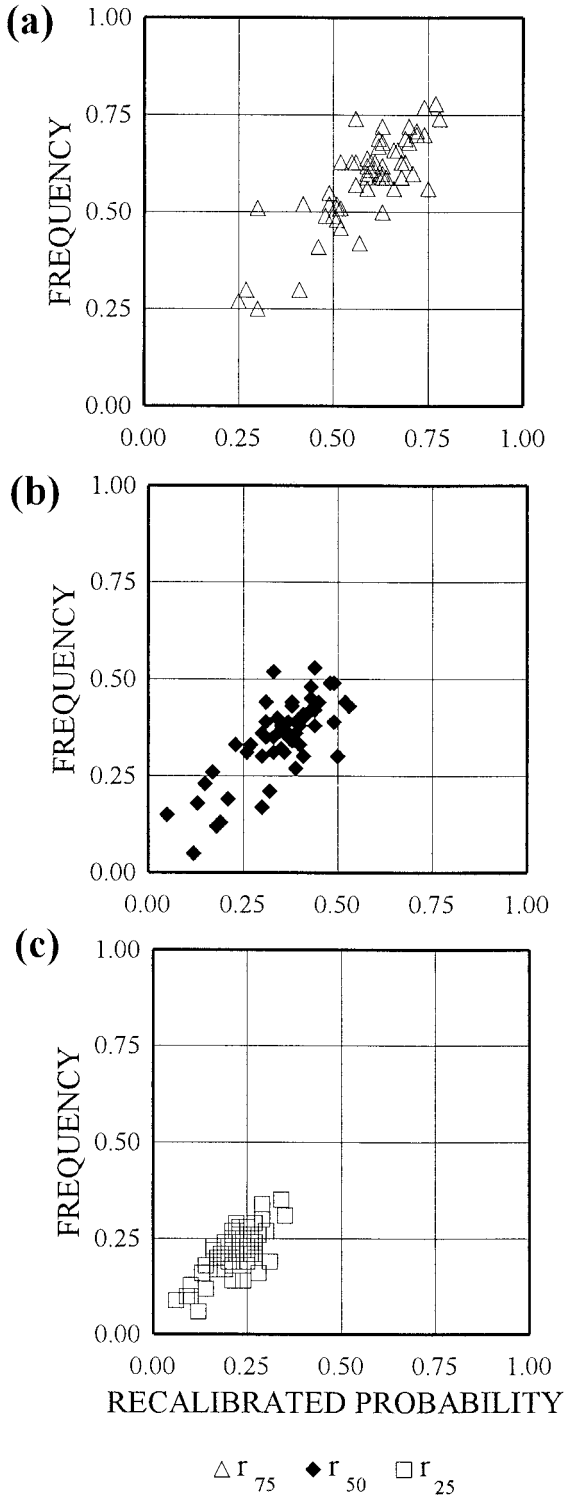

FIG. 8. Dynamic calibration diagrams resulting from monthly recalibration of PQPFs based on 3-month verification statistics: (a) calibration of x_{75}, (b) calibration of x_{50}, and (c) calibration of x_{25}; Monongahela basin.

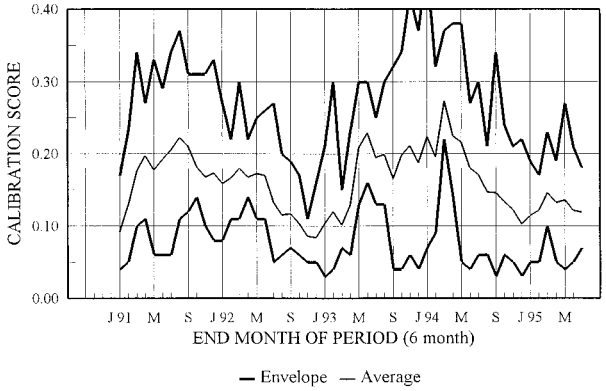

FIG. 9. Envelope and average of individual calibration scores CS for nine forecasters; both basins, 6-month verification periods.

structive to analyze the calibration of PQPFs prepared by each forecaster. There were nine of them who rotated duties at WSFO Pittsburgh during the entire pentad and prepared 86.6% of all PQPFs. In order to obtain samples large enough, the verification period is stretched to 6 months and forecasts for both basins are combined. Hence a verification statistic for a forecaster has a time series of 55 values beginning in January 1991 and ending in July 1995.

The averages of the nine time series of the CS for individual forecasters are 0.12, 0.14, 0.15, 0.16, 0.16, 0.17, 0.18, 0.18, 0.19. The time series are also summarized in terms of an envelope and an average time series shown in Fig. 9. The envelope pinpoints a consensus of forecasters' judgments that there was one ''easy'' epoch around November 1992, when each forecaster attained CS below 0.12, and two ''difficult'' epochs around June 1993 and March 1994, when none of the forecasters attained CS below 0.16 and 0.22, respectively. The envelope also pinpoints epochs when individual calibrations diverged. For example, after June 1993, the lower bound decreases while the upper bound continues to increase, implying that some forecasters were improving the calibration while others were deteriorating. The net effect, as depicted by the average time series, was a deterioration.

Overall, the lower bound of the CS envelope, which reaches below 0.05, delineates the potential for good calibration of PQPFs. At the same time, the predominantly large width of the CS envelope implies a wide variation of the calibration skill among the forecasters. This suggests a potential for improving the calibration through individualized training and exchange of operational experience among forecasters.

7. Summary

A forecast must have a well-defined normative interpretation. The skill of producing a set of forecasts that consistently adhere to this interpretation is termed

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

good calibration. This property is necessary in order that users can take a forecast at its face value.

Calibration measures reveal the degree to which the normative interpretation has been maintained in a set of forecasts. Calibration of a PQPF has two components: (i) calibration of the PoP, and (ii) calibration of the exceedance fractiles. In this paper, time series of calibration statistics from 6159 pilot PQPFs have been analyzed.

To draw overall conclusions, one aspect of PQPF production must be recalled (Krzysztofowicz and Drake 1993). As a primary guidance, Pittsburgh forecasters use the 24-h QPF prepared by the Hydrometeorological Prediction Center (formerly the Weather Forecast Branch in the Meteorological Operations Division of the National Meteorological Center). This guides the forecaster in assessing the median x_{50}. There is no central guidance for assessing the exceedance fractiles x_{75} and x_{25}. A local climatic guidance (Krzysztofowicz and Sigrest 1997) is available, but otherwise the forecaster must rely on his judgment to quantify the uncertainty in terms of x_{75} and x_{25}.

The calibration statistics attest to the forecasters' skill in quantifying the uncertainty about a precipitation amount; this confirms findings of Murphy et al. (1985). Noteworthy is the fact that x_{25} has been calibrated better than x_{75} and x_{50}. Also noteworthy is the forecasters' ability to adapt their judgments to nonstationarities of the predictand. Finally, the general trend during the pentad toward improved calibration corroborates the significance of training and experience (Murphy and Daan 1984).

The calibration statistics also reveal several weak aspects of forecasters' judgment, such as a bias toward overestimating x_{75} and x_{50}, and deterioration of the calibration after sudden breaks in a trend and during epochs of low precipitation. The challenge before designers of an operational PQPF system will be to harness the strong aspects while alleviating the weak ones. In particular, a future design should consider (i) interactive (graphical) procedures for eliciting judgmental estimates and verifying their internal coherence in real time, (ii) dynamic recalibration of forecasts, (iii) individualized training of forecasters to improve their calibration through feedback, and (iv) probabilistic central guidance.

Acknowledgments. This article was completed while Roman Krzysztofowicz was on assignment with the National Weather Service, Eastern Region, under an Intergovernmental Personnel Act agreement. Research leading to this article was supported by the National Weather Service, under the project ''Development of a Prototype Probabilistic Forecasting System.'' The leadership of Gary Carter in promoting this project and fostering a collaborative research environment within the Eastern Region is gratefully acknowledged.

APPENDIX A

Bayesian Theory of Calibration

a. Bayesian formulation

Let $\mathbf{x} = (x_{75}, x_{50}, x_{25})$ denote the vector of exceedance fractiles specified by a PQPF. According to Bayesian decision theory (Morris 1977; Krzysztofowicz 1987; Murphy and Winkler 1987), $\mathbf{x}$ is viewed as an observation of random vector $\mathbf{X} = (X_{75}, X_{50}, X_{25})$. A decision maker (an external evaluator of forecasts) does not take $\mathbf{x}$ at its face value, but uses it to construct, via Bayes' theorem, the *posterior exceedance function.* For every $\omega \geq 0$, this function specifies $P(W > \omega | \mathbf{X} = \mathbf{x})$, the probability of event $W > \omega$, conditional on forecast $\mathbf{X} = \mathbf{x}$. In order to completely characterize the stochastic dependence between $\mathbf{X}$ and W, one must construct a *family of posterior exceedance functions* for all $\mathbf{x}$.

The family of posterior exceedance functions provides a basis for defining and verifying the calibration property of forecasts (DeGroot and Fienberg 1983).

b. Definition of calibration

Predictand W has the sample space $\Omega = \{\omega: 0 \leq \omega < \infty\}$. Forecast $\mathbf{X} = (X_{75}, X_{50}, X_{25})$ has the sample space $S = \text{S0} \cup \text{S1} \cup \text{S2} \cup \text{S3}$, where

$$\begin{aligned} \text{S0} &= \{\mathbf{x}: 0 < x_{75} < x_{50} < x_{25} < \infty\}, \\ \text{S1} &= \{\mathbf{x}: 0 = x_{75} < x_{50} < x_{25} < \infty\}, \\ \text{S2} &= \{\mathbf{x}: 0 = x_{75} = x_{50} < x_{25} < \infty\}, \\ \text{S3} &= \{\mathbf{x}: 0 = x_{75} = x_{50} = x_{25}\}. \end{aligned} \tag{A1}$$

The family of posterior exceedance functions specifies probability $P(W > \omega | \mathbf{X} = \mathbf{x})$ for every $\omega \in \Omega$ and every $\mathbf{x} \in S$.

DEFINITION. The PQPF is said to be *well calibrated* if and only if the following two sets of conditions are satisfied (i) by the PoP:

$$\begin{aligned} 0.75 < P(W > 0 | \mathbf{X} = \mathbf{x}) \leq 1, &\qquad \mathbf{x} \in \text{S0}, \\ 0.50 < P(W > 0 | \mathbf{X} = \mathbf{x}) \leq 0.75, &\qquad \mathbf{x} \in \text{S1}, \\ 0.25 < P(W > 0 | \mathbf{X} = \mathbf{x}) \leq 0.50, &\qquad \mathbf{x} \in \text{S2}, \\ 0 < P(W > 0 | \mathbf{X} = \mathbf{x}) \leq 0.25, &\qquad \mathbf{x} \in \text{S3}; \end{aligned} \tag{A2}$$

and (ii) by the exceedance fractiles:

$$\begin{aligned} P(W > x_{75} | \mathbf{X} = \mathbf{x}) = 0.75, &\qquad \mathbf{x} \in \text{S0}, \\ P(W > x_{50} | \mathbf{X} = \mathbf{x}) = 0.50, &\qquad \mathbf{x} \in \text{S0} \cup \text{S1}, \\ P(W > x_{25} | \mathbf{X} = \mathbf{x}) = 0.25, &\qquad \mathbf{x} \in \text{S0} \cup \text{S1} \cup \text{S2}. \end{aligned} \tag{A3}$$

Condition (A2) should be obvious as it parallels (2)–(3). Condition (A3) may need some explanation. In particular, one may ask why is the calibration of X_{75} verified for only forecasts $\mathbf{x} \in \text{S0}$, which specify $0 < x_{75} < x_{50} < x_{25}$. Consider a counterexample. Suppose the forecast

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

is $\mathbf{x} \in$ S2 so that $0 = x_{75} = x_{50} < x_{25}$. This implies that, in the forecaster's judgment, $P(W > 0) \leq 0.50$. If one insisted on using this forecast in the verification of X_{75}, then one would require that $P(W > x_{75} | \mathbf{X} = \mathbf{x}) = P(W > 0 | \mathbf{X} = \mathbf{x}) = 0.75$. Clearly, this required posterior probability of event $W > 0$ contradicts the forecaster's probability. Hence, such a verification procedure would be improper because the forecaster could never be well calibrated.

c. Necessary conditions

To verify the sufficient conditions (A2)–(A3), one must construct $P(W > \omega | \mathbf{X} = \mathbf{x})$ for $\omega = 0$, x_{75}, x_{50}, x_{25}, and for every $\mathbf{x} \in S$. While theoretically straightforward, practically it is rather a difficult task. Helpful necessary conditions are established next.

THEOREM 1. If the PoP implied by a PQPF is *well calibrated*, then the following conditions hold:

$$\begin{aligned} 0.75 < P(W > 0 | \mathbf{X} \in \text{S0}) &\leq 1, \\ 0.50 < P(W > 0 | \mathbf{X} \in \text{S1}) &\leq 0.75, \\ 0.25 < P(W > 0 | \mathbf{X} \in \text{S2}) &\leq 0.50, \\ 0 < P(W > 0 | \mathbf{X} \in \text{S3}) &\leq 0.25. \end{aligned} \quad \text{(A4)}$$

PROOF. For any $\omega \in \Omega$, the posterior exceedance probability is specified by Bayes' theorem:

$$P(W > \omega | \mathbf{X} = \mathbf{x}) = \frac{f(\mathbf{x} | W > \omega) P(W > \omega)}{\kappa(\mathbf{x})}, \qquad \mathbf{x} \in \text{S}i, \quad i = 0, 1, 2, 3,$$

where $f(\cdot | W > \omega)$ is a generalized probability density function of $\mathbf{X}$, conditional on the hypothesis that event $W > \omega$ occurs; $P(W > \omega)$ is the prior probability of the event; and κ is a generalized probability density function of $\mathbf{X}$ specified by

$$\kappa(\mathbf{x}) = f(\mathbf{x} | W > \omega) P(W > \omega) + f(\mathbf{x} | W \leq \omega) P(W \leq \omega).$$

Suppose that $P(W > \omega | \mathbf{X} = \mathbf{x}) \leq p$ for some p and every $\mathbf{x} \in$ Si. [Each right inequality in (A2) is of this form.] Then via Bayes' theorem,

$$\frac{f(\mathbf{x} | W > \omega) P(W > \omega)}{\kappa(\mathbf{x})} \leq p.$$

When both sides are multiplied by $\kappa(\mathbf{x})$ and integrated over subspace Si, one obtains

$$\int_{\text{S}i} f(\mathbf{x} | W > \omega)\, d\mathbf{x}\, P(W > \omega) \leq p \int_{\text{S}i} \kappa(\mathbf{x})\, d\mathbf{x},$$

$$P(\mathbf{X} \in \text{S}i | W > \omega) P(W > \omega) \leq p P(\mathbf{X} \in \text{S}i),$$

$$\frac{P(\mathbf{X} \in \text{S}i | W > \omega) P(W > \omega)}{P(\mathbf{X} \in \text{S}i)} \leq p.$$

The left side is Bayes' theorem, which yields $P(W > \omega | \mathbf{X} \in \text{S}i) \leq p$. This inequality holds for any $\omega \in \Omega$, in particular for $\omega = 0$. This proves each right inequality in (A4). A parallel argument proves each left inequality. QED.

THEOREM 2. If the exceedance fractiles specified by a PQPF are *well calibrated*, then the following conditions hold:

$$\begin{aligned} P(W > X_{75} | \mathbf{X} \in \text{S0}) &= 0.75, \\ P(W > X_{50} | \mathbf{X} \in \text{S0} \cup \text{S1}) &= 0.50, \\ P(W > X_{25} | \mathbf{X} \in \text{S0} \cup \text{S1} \cup \text{S2}) &= 0.25. \end{aligned} \quad \text{(A5)}$$

PROOF. For $p = 0.75$, via derivation analogous to that used in the proof of Theorem 1, one can establish that the first line of (A3) implies

$$\overline{H}(x_{75} | \mathbf{X} \in \text{S0}, X_{75} = x_{75}) = P(W > x_{75} | \mathbf{X} \in \text{S0}, X_{75} = x_{75}) = 0.75.$$

It now follows that

$$P(W > X_{75} | \mathbf{X} \in \text{S0}) = E[\overline{H}(X_{75} | \mathbf{X} \in \text{S0}, X_{75}) | \mathbf{X} \in \text{S0}] = 0.75,$$

where the conditional expectation is taken with respect to X_{75}. This proves the first line of (A5). For $p = 0.50$ and $p = 0.25$, the proofs are analogous. QED.

d. Empirical verification procedure

Because the sufficient conditions (A2)–(A3) are difficult to verify without a parametric model for the posterior exceedance function, the empirical verification procedure is formulated using the necessary conditions (A4)–(A5). This procedure, which is described in section 3, arises as follows. First, notation is simplified by noting that

$$\begin{aligned} &\mathbf{X} \in \text{S}i \Leftrightarrow \text{R}i, \qquad i = 0, 1, 2, 3, \\ &\mathbf{X} \in \text{S0} \Leftrightarrow X_{75} > 0, \\ &\mathbf{X} \in \text{S0} \cup \text{S1} \Leftrightarrow X_{50} > 0, \\ &\mathbf{X} \in \text{S0} \cup \text{S1} \cup \text{S2} \Leftrightarrow X_{25} > 0. \end{aligned}$$

Second, conditions (A4)–(A5) are adopted as an operational definition of a well-calibrated PQPF.

When predictand W has a strictly continuous distribution with support Ω, then $P(W > 0) = 1$ and $S =$ S0. Consequently, (A4) vanishes and (A5) reduces to

$$\begin{aligned} P(W > X_{75}) &= 0.75, \\ P(W > X_{50}) &= 0.50, \\ P(W > X_{25}) &= 0.25. \end{aligned} \quad \text{(A6)}$$

Condition (A6) has been widely used as the definition of a well-calibrated probabilistic forecast of a predictand having a strictly continuous probability distribution. Examples of verification analyses based on (A6) can be

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

found in Murphy and Winkler (1974, 1979) and Alpert and Raiffa (1982). Hence, the empirical verification procedure developed herein for the PQPF is a generalization of an established approach. Such a generalization is necessary for any predictand having a discrete-continuous probability distribution.

APPENDIX B

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

$$\alpha = t, \qquad \beta = T - t. \tag{B1}$$

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

$$E(\Theta \,|\, T, t) = \frac{t}{T}, \tag{B2}$$

$$\mathrm{Var}(\Theta \,|\, T, t) = \frac{t(T - t)}{T^2(T + 1)}. \tag{B3}$$

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 $H(\theta) = P(\Theta \le \theta \,|\, T, t)$, 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 $P(\Theta \le \theta_{100p} \,|\, T, t) = p$ and such that

$$\theta_{100p} = H^{-1}(p), \qquad 0 < p < 1, \tag{B4}$$

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

$$C_{100p} = (\theta_{100(1-p)/2}, \theta_{100(1+p)/2}) \tag{B5}$$

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).

c. Application to verification

In the context of the verification procedure described in section 3, probability Θ is interpreted either as $P(W > 0 \,|\, \mathrm{R}i)$, when one verifies the probability of precipitation, or as $P(W > X_{100p} \,|\, X_{100p} > 0)$, when one verifies the exceedance fractiles. In each case, (B2)–(B3) provide a basis for defining the frequency estimator of Θ and the associated standard deviation. A central credible interval can be calculated according to (B1), (B4), and (B5). The inverse H^{-1} of the incomplete beta function can be evaluated approximately via formulas from Abramowitz and Stegun (1972, p. 945).

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.

DeGroot, M. H., and S. E. Fienberg, 1983: The comparison and evaluation of forecasters. *Statistician,* **32,** 12–22.

Krzysztofowicz, R., 1987: Markovian forecast processes. *J. Amer. Stat. Assoc.,* **82,** 31–37.

——, 1992: Bayesian correlation score: A utilitarian measure of forecast skill. *Mon. Wea. Rev.,* **120,** 208–219.

——, 1996: Sufficiency, informativeness, and value of forecasts. *Proc. Workshop on the Evaluation of Space Weather Forecasts,* Boulder, CO, Space Environment Center, National Oceanic and Atmospheric Administration, 103–112.

——, and T. R. Drake, 1992: Probabilistic quantitative precipitation forecasts for river forecasting. Preprints, *Symp. on Weather Forecasting,* Atlanta, GA, Amer. Meteor. Soc., 66–71.

——, and ——, 1993: Usage of guidance products in preparing probabilistic QPFs for river basins. Post-prints, Third National Heavy Precipitation Workshop, 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, 1997: Local climatic guidance for proba-

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC

bilistic quantitative precipitation forecasting. *Mon. Wea. Rev.*, **125,** 305–316.

——, 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.

Morris, P. A., 1977: Combining expert judgments: A Bayesian approach. *Manage. Sci.*, **23,** 679–693.

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

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

——, and H. Daan, 1984: Impacts of feedback and experience on the quality of subjective probability forecasts: Comparison of results from the first and second years of the Zierikzee experiment. *Mon. Wea. Rev.*, **112,** 413–423.

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

——, W.-R. Hsu, R. L. Winkler, and D. S. Wilks, 1985: The use of probabilities in subjective quantitative precipitation forecasts: Some experimental results. *Mon. Wea. Rev.*, **113,** 2075–2089.

NOAA, 1972: National Weather Service river forecast system forecast procedures. NOAA Tech. Memo. NWS Hydro-14. [Available from National Weather Service, Hydrological Research Laboratory, 1325 East–West Highway, Silver Spring, MD 20910.]

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.

Unauthenticated | Downloaded 08/02/21 03:46 AM UTC