AGU PUBLICATIONS

Water Resources Research

RESEARCH ARTICLE

10.1002/2014WR016617

On the assessment of reliability in probabilistic hydrometeorological event forecasting

Caleb M. DeChant[1] **and Hamid Moradkhani**[1]

[1]Department of Civil and Environmental Engineering, Portland State University, Portland, Oregon, USA

Abstract Probabilistic forecasts are commonly used to communicate uncertainty in the occurrence of hydrometeorological events. Although probabilistic forecasting is common, conventional methods for assessing the reliability of these forecasts are approximate. Among the most common methods for assessing reliability, the decomposed Brier Score and Reliability Diagram treat an observed string of events as samples from multiple Binomial distributions, but this is an approximation of the forecast reliability, leading to unnecessary loss of information. This article suggests testing the hypothesis of reliability via the Poisson-Binomial distribution, which is a generalized solution to the Binomial distribution, providing a more accurate model of the probabilistic event forecast verification setting. Further, a two-stage approach to reliability assessment is suggested to identify errors in the forecast related to both bias and overly/insufficiently sharp forecasts. Such a methodology is shown to more effectively distinguish between reliable and unreliable forecasts, leading to more robust probabilistic forecast verification.

Key Points:

- Current techniques for verifying probabilistic forecasts are approximate
- Poisson-Binomial distribution is the precise solution to forecast verification
- The necessary sample size for reliability assessment is reduced

Correspondence to:
H. Moradkhani,
hamidm@pdx.edu

Citation:
DeChant, C. M., and H. Moradkhani (2015), On the assessment of reliability in probabilistic hydrometeorological event forecasting, *Water Resour. Res.*, *51*, 3867–3883, doi:10.1002/2014WR016617.

Received 7 NOV 2014
Accepted 24 APR 2015
Accepted article online 28 APR 2015
Published online 1 JUN 2015

1. Introduction

Hydrometeorological events (e.g., precipitation occurrence, droughts, floods) are often forecasted as probabilities, representing a forecaster's certainty that a given event will occur [*Murphy et al.*, 1980; *Madadgar and Moradkhani*, 2013; *Wetterhall et al.*, 2013; *Yan and Moradkhani*, 2015]. Such probabilistic forecasts are motivated by the presence of uncertainties in land surface and atmospheric processes, which undermine the ability to precisely predict future event occurrences [*Slingo and Palmer*, 2011; *DeChant and Moradkhani*, 2014]. Since forecasters do not have complete knowledge of future events, hydrologists and meteorologists alike have recognized the benefits of communicating uncertainty in their forecasts [*Hamill*, 2012; *Pappenberger et al.*, 2011]. This is evidenced by the wealth of operational probabilistic forecasting systems [*Buizza et al.*, 1999; *Demargne et al.*, 2014; *Park et al.*, 2008; *Saha et al.*, 2006] and probabilistic forecasting research initiatives [*Schaake et al.*, 2007]. By issuing probabilistic forecasts, the end user is notified of the imperfect nature of the forecast, and therefore should only rely on a forecasted event occurring with the designated probability [*Joslyn and Savelli*, 2010; *Gigerenzer et al.*, 2005]. Further, this communication of forecast uncertainty can improve risk management when resources are in danger, assuming that the forecasts accurately represent the uncertainty of an event occurring [*Carriquiry and Osgood*, 2012]. This necessitates detailed examination of forecast quality to ensure effective management of risk.

Two characteristics indicate the quality of a probabilistic forecast: reliability and sharpness. Reliability, also termed calibration, refers to the accuracy of the forecasted probability in conveying the true probability of an event occurring [*Christensen et al.*, 2015]. For example, an event that is forecasted with a probability of 50% should occur in 50% of instances. Alternatively, sharpness is the level of certainty in the forecast, where greater sharpness indicates a reduction in uncertainty, which may be measured by the forecast variance or entropy [*Machete*, 2013]. A shaper forecast will have a tendency to generate probabilities approaching zero or one, with a perfectly sharp forecast only generating values of zero or one (deterministic forecast). With both the reliability and sharpness components of a forecast being important, it becomes necessary to have a multiobjective verification system for full assessment of forecast quality.

Multiobjectivity in forecast verification may be achieved through either a continuous function or rule based comparison. Continuous functions used for assessing probabilistic event forecasts should be strictly proper [*Bröcker*, 2009; *Christensen et al.*, 2015; *Gneiting and Raftery*, 2007], with typical examples being quadratic, spherical, or logarithmic functions [*Bickel*, 2007]. Of these functions, quadratic is particularly common, which is often referred to as the Brier Score (BS) [*Brier*, 1950]. The BS is a smooth function that is strictly proper, providing a statistically sound method for comparing competing forecasts, but the BS has a complex relationship between sharpness and reliability [*Mason*, 2004]. Alternatively, a rule-based approach may exhibit more control over the interaction between reliability and sharpness. This study takes the perspective that reliability should be held paramount, and therefore follows the paradigm "maximizing sharpness subject to calibration," as stated in *Gneiting et al*. [2007]. Within this paradigm, reliability of a forecast is a requisite condition for acceptability [*Mitchell and Wallis*, 2011]. Although sharper forecasts are desired, it is imperative to ensure that sharpness is not a factor when comparing an unreliable forecast to a reliable forecast. Through this framework, it is essential that reliability assessment be accurate, motivating a detailed look at the typical methods for reliability evaluation. The remainder of this manuscript will examine reliability assessment in probabilistic event forecasting, with the intention of assuring maximum accuracy when assessing reliability.

2. Identifying a Distribution for Reliability Assessment

Assume that some forecast methodology, f, using some information, D_t, estimates the probability of an event, p_t, at time t, as is shown in equation (1).

$$p_t = f(D_t) \tag{1}$$

Likewise, assume that an observation, O_t, is available at each forecast time, which may be either 0 or 1, with 1 indicating event occurrence and 0 indicating event nonoccurrence. This is the typical verification setting, where the forecasted probabilities and observed event occurrences compose all available information. With this information, the forecaster will attempt to determine if the forecast is a reliable predictor of the event of interest.

A probabilistic forecast is deemed reliable if the forecasted event probabilities are statistically indistinguishable from the true event probabilities [*Annan and Hargreaves*, 2010]. Note that the term "true probability" used here refers to the probability that properly represents the uncertainty in the forecast. Reliability assessment therefore becomes an examination of the similarity between the forecasted and true probabilities. Although the true probabilities are not directly available in the verification setting, the forecaster may assume that the observations provide information about the true probabilities. A prudent approach is to view the observations as random binary variables, each drawn according to the true event probability. By viewing the observations as random variables, the observations become representative of the true event probability. Since the forecaster must evaluate the similarity between the forecasted and true probabilities, and the observations are assumed to be drawn with the true probability, the problem may be inverted by quantifying the probability that the observations were drawn based on the forecasted probabilities. This will be referred to as the probability of reliability.

Drawing a random binary variable based on a forecasted probability is modeled by the Bernoulli distribution. In order to estimate the probability of reliability, each forecast should be viewed as a Bernoulli trial, with the probability of $p(p_t, O_t)$ according to equation (2).

$$p(p_t, O_t) = \begin{cases} p_t & \textit{if} \quad O_t = 1 \\ 1 - p_t & \textit{if} \quad O_t = 0 \end{cases} \tag{2}$$

Equation (2) provides a means to estimate the probability of a single observation of the event, assuming that the forecasted probability is equal to the true probability. Although equation (2) allows the forecaster to estimate the probability of each observation being drawn with the forecasted probability, the forecaster will be required to estimate the probability of a set of forecasts and observations occurring simultaneously in order to have sufficient information for robust reliability assessment. A first step is estimating the probability of the specific set of forecasted probabilities ($p_{1:T}$) and observations ($O_{1:T}$) occurring, according to equation (3). Note that equation (3) assumes that the forecasts are serially independent.

$$p(p_{1:T}, O_{1:T}) = \prod_{t=1}^{T} p(p_t, O_t) \tag{3}$$

While equation (3) provides the forecaster with the probability of the specific forecast and observation sequence, this probability will become infinitesimal for a large number forecast and observation pairs. It is suggested here that the probability of reliability should be formulated into a probability distribution, which may be achieved by viewing the observations as random variables. When viewing the observations as random variables, all permutations of $O_{1:T}$ must be examined. Therefore, it is necessary to estimate the probability of K events occurring, where K is estimated according to equation (4).

$$K=\sum_{t=1}^{T} O_t \tag{4}$$

This necessitates the summation of $p(p_{1:T}, O_{1:T})$ over each permutation of K observations in T trials, estimating the probability that K events may occur. Within this setting, the Poisson-Binomial distribution [*Hodges and Le Cam*, 1960; *Hong*, 2013] estimates the probability of reliability exactly, and the corresponding Probability Mass Function (PMF) is shown in equation (5).

$$f_{PB}(p_{1:T}, K)=\sum_{A\in S}\left(\prod_{t\in A} p_t \prod_{t\in A^c}(1-p_t)\right) \tag{5}$$

In equation (5), S is the set of all the permutations of K event occurrences in T trials that satisfy equation (4), A represents a specific permutation drawn from S, and A^c is the complement of A ($A^c=(1-A)$). Therefore, f_{PB} $(p_{1:T}, K)$ is the probability that K events will occur if $p_{1:T}$ is equal to the true series of event probabilities. More specifically, equation (5) estimates the probability that K observations would have occurred, assuming the forecast is reliable, which is equal to the probability of reliability. At this point, it is important to note that the Poisson-Binomial distribution will only be sensitive to bias in $p_{1:T}$, and more complex types of unreliability will require additional considerations. This issue will be examined in sections 6.2, where numerical experiments examine the utility of the Poisson-Binomial distribution for reliability assessment, and section 7, where a new approach is developed for reliability assessment.

3. Formal Hypothesis Testing

In this article, it is suggested that reliability assessment should take a rejectionist approach, where a forecaster hypothesizes that the forecast is reliable (null hypothesis), and attempts to disprove that hypothesis. If the forecaster cannot provide sufficient evidence to prove that the true probabilities are different from the forecasted probabilities, then the hypothesis of reliability cannot be rejected. Verification with this methodology is regularly performed for continuous predictands, typically with the use of the chi-squared test [*Joliffe and Primo*, 2008], but is rare among forecasts of dichotomous hydrometeorological events.

Such a hypothesis test may be performed with the Poisson-Binomial distribution, but requires the formulation of the Cumulative Distribution Function (CDF). The CDF of the Poisson-Binomial distribution is estimated according to equation (6).

$$F_{PB}(k\le K)=\sum_{k=0}^{K}\left[\sum_{A\in S_k}\left(\prod_{t\in A} p_t \prod_{t\in A^c}(1-p_t)\right)\right] \tag{6}$$

In order to perform this hypothesis test, a significance level (p-value) will need to be selected to reject the null hypothesis, which will be 0.05 throughout this article. More specifically, if $0.025\le F_{PB}(k\le K)\le 0.975$, then the hypothesis of reliability will not be rejected, and therefore the forecast will be deemed reliable. If multiple different forecast methods are deemed reliable, then the sharpest of the reliable forecast methods will be selected as the “best” forecast, satisfying the paradigm of “maximizing sharpness subject to calibration,” as described in section 1. Within the Poisson-Binomial Distribution, increasing sharpness is identified with a reduction in variance, which is estimated by equation (7).

$$\sigma_{PB}^2=\sum_{t=1}^{T} p_t(1-p_t) \tag{7}$$

Direct estimation of equation (6) is computationally infeasible for any useful sample size due to the large number of permutations of the observed events [*Hong*, 2013]. In order to overcome this issue, it is possible to use the Discrete Fourier Transform and the Characteristic Function, as demonstrated by *Hong* [2013], to solve the Poisson-Binomial CDF at any practically relevant sample size. This provides an exact solution to estimate the Poisson-Binomial CDF, thus allowing for precise hypothesis testing.

4. Conventional Reliability Assessment

The Poisson-Binomial distribution is absent from the hydrometeorological literature, and only approximations are present for probabilistic event forecast verification. All conventional reliability metrics are based on the Binomial Distribution, which is a specific case of the Poisson-Binomial distribution, where all forecasted probabilities are equal. Use of the Binomial distribution is therefore an approximation in the probabilistic verification setting, leading to a loss of statistical power, with the exception of reliance on climatology, where the historical frequency of the event is used for forecasting. The Binomial CDF is much simpler than the Poisson-Binomial CDF, as shown in equation (8), and has therefore been an attractive alternative for general use.

$$F_B(k \le K) = \sum_{k=0}^{K} \binom{T}{k} \overline{p_{1:T}}^{k} \left(1-\overline{p_{1:T}}\right)^{(T-k)} \tag{8}$$

In equation (8), $\overline{p_{1:T}}$ is the average of all forecast probabilities $\left(\overline{p_{1:T}} = \frac{1}{T}\sum_{t=1}^{T} p_t\right)$ and $\binom{T}{k}$ is the binomial coefficient, estimated according to equation (9), which removes the need to sum over all permutations of event occurrences.

$$\binom{T}{k} = \frac{T!}{k!(T-k)!} \tag{9}$$

The Binomial CDF provides a simplified function for estimating the probability of reliability, but this will become increasingly approximate as the variability in forecasted probabilities increases. In order to reduce these errors, it has become common to group similarly valued forecasts, referred to as binning. Although binning is utilized to reduce error in the Binomial Distribution, it has the added benefit of identifying complex types of unreliability, and therefore may also be necessary when using the more appropriate Poisson-Binomial Distribution. This binning approach will divide the possible range of probabilities ($p_t \in [0,1]$) into B groups, which are typically evenly spaced. According to equation (10), each of the probabilities within the bin limits $\left(\left[\left(b\frac{1}{B}-\frac{1}{B}\right), \left(b\frac{1}{B}\right)\right]\right)$ are selected for set b, where b is the selected bin number, which contains n_b forecasts.

$$p_{b,1:n_b} = \left\{ \left(b\frac{1}{B}-\frac{1}{B}\right) \le p_{1:t} < \left(b\frac{1}{B}\right) \right\} \tag{10}$$

Along with the binned probabilities, the observations must be binned as well, which is shown in equation (11), and the total number of observed occurrences within each bin is estimated according to equation (12), which is the application of equation (4) to multiple bins.

$$O_{b,1:n_b} = \{\, O_t \quad if \quad p_t \in p_{b,1:n_b} \tag{11}$$

$$K_b = \sum_{i=1}^{n_b} O_{b,i} \tag{12}$$

In order to evaluate the Binomial distribution at each bin, the bin-averaged forecast probability ($\overline{p_b}$) will be estimated from equation (13).

$$\overline{p_b} = \frac{1}{n_b} \sum_{i=1}^{n_b} p_{b,i} \tag{13}$$

With this bin averaged probability, the Binomial CDF may be evaluated according to equation (14).

$$F_{B,b}(k \le K_b) = \sum_{k=0}^{K_b} \binom{n_b}{k} \overline{p_b}^{k} \left(1-\overline{p_b}\right)^{(n_b-k)} \tag{14}$$

By binning the forecasted probabilities, the forecast verification problem is broken up into multiple separate problems, where the bin-averaged probability becomes increasingly representative of the set of probabilistic forecasts with decreasing bin size.

Rather than directly estimating the Binomial CDF, meteorologists and hydrologist commonly use approximations. The most common verification methods are the BS and the Reliability Diagram. The original form of the BS is presented in equation (15), estimating the mean square error (MSE) of the forecasted probabilities and corresponding observations. As mentioned before, a perfect BS requires both perfect reliability and sharpness.

$$BS=\frac{1}{T}\sum_{t=1}^{T}(p_t-O_t)^2 \tag{15}$$

In order to assess reliability directly, the BS must be decomposed [*Murphy*, 1973]. Decomposition of the BS requires binning forecasted probabilities and observations, allowing for the comparison of bin average forecasted probabilities ($\overline{p_b}$) and bin observation frequencies ($\overline{O_b}$), as is shown in equation (16) [*Stephenson et al.*, 2007].

$$BS=\underbrace{\frac{1}{T}\sum_{b=1}^{B}n_b\left(\overline{p_b}-\overline{O_b}\right)^2}_{\text{Re}\,liability}+\underbrace{\frac{1}{T}\sum_{b=1}^{B}\left[\sum_{i=1}^{n_b}\left[\left(O_{b,i}-\overline{O_b}\right)^2+\left(p_{b,i}-\overline{p_b}\right)^2-2\left(O_{b,i}-\overline{O_b}\right)\left(p_{b,i}-\overline{p_b}\right)\right]\right]}_{Variances\quad and\quad Covar\,iance} \tag{16}$$

In equation (16), $\overline{O_b}$ is the observation frequency within bin b $\left(\overline{O_b}=\frac{K_b}{n_b}\right)$. The first summation on the right-hand side of equation (16) is the reliability estimate based on the BS (BS_R), which is minimized with a perfectly reliable forecast. The second summation in equation (16) is the within bin variance of the observation and forecast, and the within bin covariance of the observation and forecast, which is minimized with perfect sharpness. Through equation (16), the BS_R may be directly estimated as the MSE of the bin averaged forecast probabilities and the bin observation frequencies. For the remainder of the article, the error in the bin averaged forecast probabilities ($\overline{p_b}-\overline{O_b}$), will be referred to as probabilistic residuals. As the number of bins approaches infinity, the probabilistic residuals will approach Gaussianity, making the BS_R approach perfect estimation of the probability of reliability from the Binomial distribution with increasing sample size [*Feller*, 1945], with the exception that the BS_R is inversely proportional to the probability from the Binomial distribution. Although the BS_R will approach the exact solution to the Binomial distribution as B approaches infinite, there will be some error due to this approximation at any practical number of bins.

The Reliability Diagram provides a means for graphical comparison of the probabilistic residuals, allowing for visual assessment of forecast performance. In this diagram, $\overline{O_b}$ is plotted on the vertical axis and $\overline{p_b}$ is plotted on the horizontal axis, and then the corresponding points are compared to the one-to-one line. The proximity of the Reliability Diagram to the one-to-one line indicates a high probability of reliability. Although the Reliability Diagram is very useful for diagnosing errors in different bins, it may be misleading as probabilistic residuals in each bin are not proportional to the Binomial distribution. In order to overcome this problem, *Bröcker and Smith* [2007] translated the Reliability Diagram into probability space using the Binomial CDF. This provides a more accurate assessment of reliability from the Reliability Diagram.

The use of the BS_R and Reliability Diagram provides simple means for assessing the reliability of probabilistic hydrometeorological event forecasts, but these simplifications have drawbacks. First, these methods are approximations of the Binomial distribution, except in the case described in *Bröcker and Smith* [2007]. As approximations, it is not clear the extent to which these methods damage the assessment of forecast reliability. Second, both methods are based on the Binomial distribution, which is limiting. It becomes a balance between having sufficiently small forecast bin variance to reduce errors, and enough observations in each bin to draw meaningful conclusions. A certain number of bins may be necessary to fully assess reliability, but the required number of bins to reduce approximation errors in the Binomial distribution is potentially greater than the number required for reliability assessment with the Poisson-Binomial distribution. Further discussion on necessary bin size for different distributions is provided in sections 6 and 7. Finally, thresholds for hypothesis testing within the BS_R cannot be derived theoretically, and therefore the BS_R cannot precisely distinguish between reliable and unreliable forecasts. Although the BS_R provides a useful

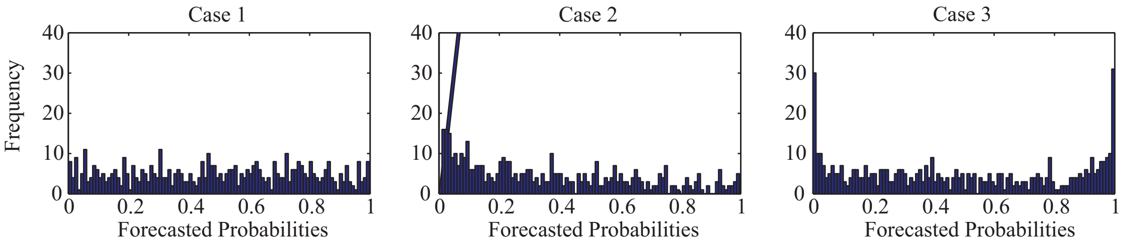

Figure 1. Histograms of the forecasted probabilities for each case, with $x=2$ in equation (18).

method for comparing the probability of reliability, it is restricting from the rejectionist perspective. Due to the problems highlighted above, it is necessary to examine the impacts conventional verification tools have on reliability assessment. Such an examination was performed with numerical experiments, as described in sections 5 and 6.

5. Numerical Experiments

Multiple synthetic probabilistic forecasting experiments were performed to examine the performance of conventional reliability assessment in comparison to the Poisson-Binomial distribution. Within these experiments, three forecast cases were implemented to examine the effects of varying degrees of forecast sharpness. The first case is presented in equation (17), where the forecasts are sampled from the standard uniform distribution. Based on case 1, a second case creates forecasts with probabilities tending toward zero, as shown in equation (18). For the generation of the forecasts for case 2, the exponent x will be set to a value of 2 throughout the experiments presented in section 6.1, but will range from 1 to 1.25 in the experiments presented in section 6.2. A third case is generated according to equation (19), creating to a "U"-shaped distribution. Case three is the sharpest of all the forecasting cases, and is therefore the best case, assuming that all forecasts are reliable. Note that T is set to 500 throughout this study.

$$p_{t,1} \sim U(0,1) \tag{17}$$

$$p_{t,2}=p_{t,1}^{x} \tag{18}$$

$$p_{t,3}=\begin{cases} p_{t,2} & if \quad t<\dfrac{T}{2} \\ 1-p_{t,2} & otherwise \end{cases} \tag{19}$$

Histograms of these forecasts are provided in Figure 1. From Figure 1, it is clear that the case 1 makes every probability equally likely to be forecasted, case 2 has a tendency to forecast towards 0, and case 3 tends towards both 0 and 1.

In section 6.1, the different verification methods will be examined under reliable forecasting conditions. This requires sampling the observations according to the forecasted probabilities, thus ensuring that the forecasted probabilities are the true probabilities. The sampling of observations is shown in equation (20), where $p_{t,case}$ is the probability of forecast t for a given case. $O_{1:T,case}$ is therefore a set of observations which the given case forecasts reliably.

$$O_{t,case}=\begin{cases} 1 & if \quad U(0,1) \leq p_{t,case} \\ 0 & \quad otherwise \end{cases} \tag{20}$$

Further experiments are performed to determine the ability of the verification methods to reject unreliable forecasts. In order to perform this analysis, the exponent (x) in case 2 ranges from 1 to 1.25, and the corresponding values are estimated for case 3. These new cases (case 2 and 3 with x values ranging from 1 to 1.25) are then compared to observations drawn with probabilities according to case 1 ($O_{1:T,1}$). This creates a scenario where the forecasts become increasingly unreliable, due to both skewed probabilities and overly

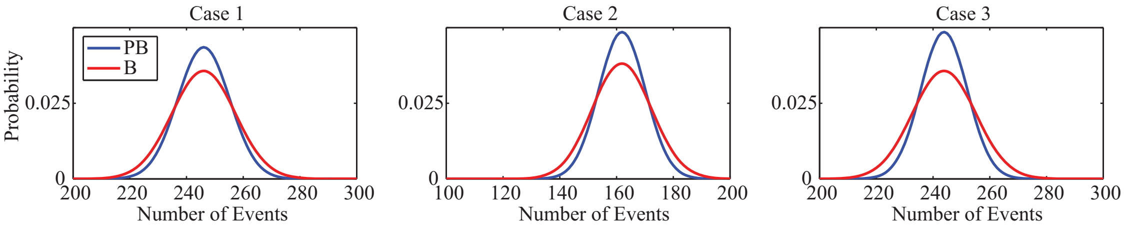

Figure 2. Comparison of the probability distributions of the Poisson-Binomial (PB) and Binomial (B) distributions.

confident probabilities, with respect to the observations. Results for these increasingly unreliable forecasts will be examined in section 6.2.

6. Results

6.1. Reliable Forecasts

A first examination of the errors related to conventional metrics requires a comparison of the Binomial and Poisson-Binomial distribution. This is presented for each forecast case in Figure 2, where the Binomial and Poisson-Binomial probability distributions are presented for each case, with the use of a single bin. A first observation from this figure is that the Binomial distribution is wider than the Poisson-Binomial distribution for every case. This increased width of the Binomial Distribution is expected, as the variance of the Binomial distribution will always be greater than the Poisson-Binomial distribution, except in the case where all probabilities are equal (climatology). This is proven in Appendix A.

Figure 2 also shows that the difference between the Binomial and Poisson-Binomial distribution increases as forecast sharpness increases, which is supported by the presentation in Appendix A. A wider distribution suggests that simplifying the verification problem, through the use of the Binomial Distribution, reduces one's ability to reject the hypothesis of reliability, thus increasing the possibility of type II errors. This error is largest in Case 3, which happens to be the sharpest case. Given that each of the three forecast cases is reliable, Case 3 should be selected as it provides a reliable forecast with the most certainty. In the event that all cases are unreliable, Case 3 is the most probable to be erroneously deemed reliable, as it widens the Binomial CDF, increasing the likelihood of incorrectly selecting Case 3 as the best forecast based on the Binomial CDF. Overall the single bin analysis shows that use of the Binomial distribution reduces statistical power.

Due to the loss of information caused by simplifying the problem with the Binomial distribution, the binning approach may be used to reduce the effects of forecast variability. In order to assess the effects of binning forecasts, Figure 3 shows the width of the 95% confidence interval for each distribution as a function of bin size, where the total width of the confidence interval, summed across all bins, is presented. This figure demonstrates the rapid growth of the 95% confidence interval with an increasing number of bins. Since the grouping process reduces the sample size at each bin, the 95% confidence interval is widened, causing an aggregate effect on the overall determination of reliability. By binning similarly valued forecasts, one vastly reduces the ability to distinguish between reliable and unreliable forecasts, further increasing the chance of Type II errors. This loss of information due to binning is especially concerning in the case of hydrometeorological extremes (i.e., floods, droughts, heat-waves), which are, by definition, low probability events, making it essential to efficiently use information from every observation. Overall it is important for forecasts to be verified with as few bins as possible, increasing the effective sample size, thus maximizing one's ability to reject unreliable forecasts.

A further observation from Figure 3 is that forecast sharpness affects the magnitude of approximation errors in the Binomial distribution, even with a large number of bins. It is expected that errors in the Binomial CDF, in comparison to Poisson-Binomial CDF, will decrease with an increasing number of forecast bins, as each bin becomes more representative of its members. This is evidenced in Case 1, where the Binomial CDF approaches the Poisson-Binomial CDF with decreasing bin size. Alternatively, the Binomial CDF in Case 2

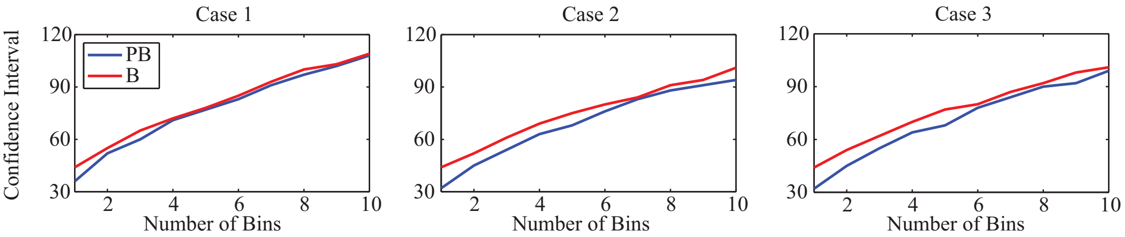

Figure 3. The width of the 95% confidence interval (*K*) of the Poisson-Binomial (PB) and the Binomial (B) CDFs, with respect to the number of bins.

and Case 3 has persistent error even with 10 bins. This result suggests that a large number of bins may be necessary for errors associated with the Binomial CDF to be considered negligible.

Further analysis of the effects of varying bin size is performed with respect to the BS_R in Figure 4. In this figure, the variability in reliability scores between the three cases is compared with increasing numbers of bins, through 100 replicates of each forecast case. In this figure, it is expected that the difference between the distributions of BS_R values will decrease with increasing bin size, due to reduced approximation errors in the Binomial distribution. Since the probabilities within each bin become more homogeneous with an increasing number of bins, the BS_R becomes more consistent across varying levels of sharpness. The results here show that the BS_R requires around six bins to remove these approximation errors. Although Figure 4 indicates the within bin variance is becoming negligible (equation (16)), note that the distribution of reliability values is widening, indicating the loss of information with increasing number of bins. As was found in Figure 3, the increasing number of bins reduces the statistical power of any verification metric.

6.2. Increasingly Unreliable Forecasts

A comparison of the BS_R, Binomial distribution and the Poisson-Binomial distribution for identifying unreliable forecasts is presented in Figure 5, where the observation is drawn from case 1, but the forecast is created with cases 2 and 3 with increasing *x* (equation (18)). The analysis of the Binomial and Poisson-Binomial distributions in Figure 5 uses a single bin approach, whereas the BS_R uses six bins based on the analysis of Figure 4. In Figure 5, the fraction of 100 forecast replicates that are rejected, with a significance of 95%, is shown with respect to *x*, where the threshold for the BS_R was estimated from the results presented in Figure 4 (the threshold for BS_R is set to 0.0043). For case 2, it is clear that the fraction of forecast replicates rejected with the Poisson-Binomial distribution increases more rapidly than with the Binomial distribution or the BS_R. This

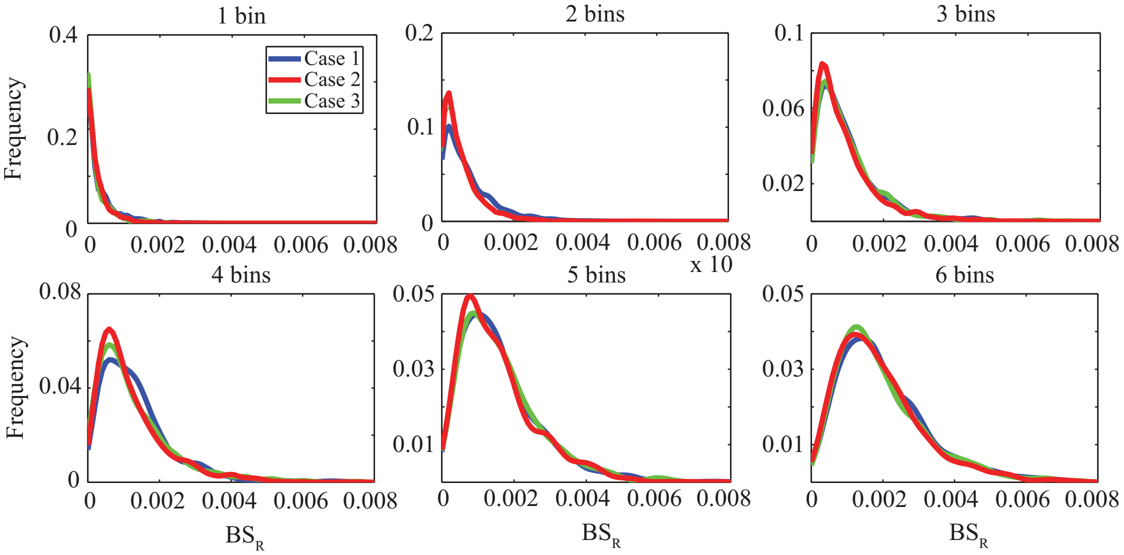

Figure 4. Distribution of reliability scores for case 1, in blue, case 2, in red, and case 3, in green, for multiple numbers of bins, with $x=2$ from equation (18).

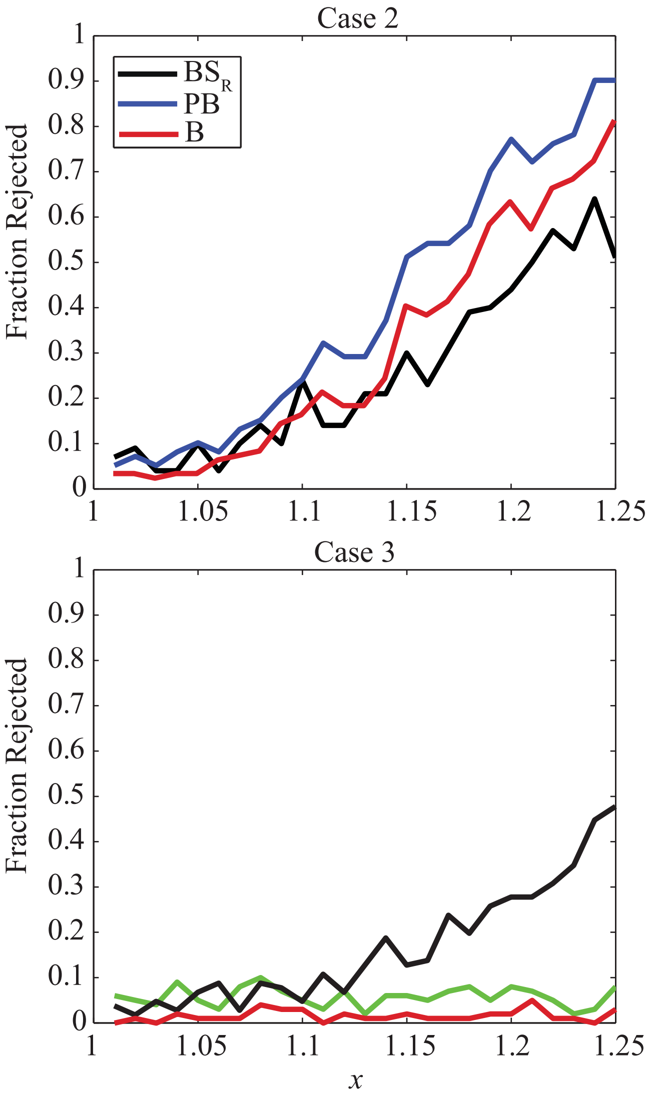

Figure 5. Fraction of forecasts rejected via the reliability component of the Brier Score (BSR), the Poisson-Binomial distribution (PB), and the Binomial distribution (B), from 100 replicates, for varying *x* values (equation (18)) of case 2 and 3.

indicates that Poisson-Binomial distribution has the greatest statistical power, the Binomial distribution has a small loss of information, and the BS$_R$ has greater loss of information than the Binomial distribution. This result shows that the Poisson-Binomial distribution is very effective in rejecting unreliable forecasts that are improperly skewed, and therefore biased, but the results are much different for case 3. In this case, the Poisson-Binomial and Binomial distributions are largely unable to reject case 3 with an *x* value of 1.25. Alternatively, the BS$_R$ approaches a rejection rate of 0.5 with and *x* value of 1.25. This indicates that a multibin approach is required to reject some unreliable forecasts. Although a single bin verification framework minimizes the width of the 95% significance interval, this will only be useful if the forecast is significantly biased, as in case 2. Alternatively, if the forecast is unbiased, yet still unreliable, as in case 3, the errors will go unnoticed without examining separate bins. Further exploration of this scenario is performed with the Reliability Diagram.

The reliability diagram for the scenarios explored in Figures 5 is presented in Figure 6. In this figure, the top row shows the median Reliability Diagram of all 100 replicates for increasing skew (case 2), with associated 95% significance intervals from the Binomial distribution, and the second row shows the rejection rate for each bin. Likewise, the bottom two rows show the same information for case 3. Each reliability diagram uses six bins, following the analysis presented in Figure 4. With respect to case 2, the median reliability diagram steadily approaches the upper limit of the significance interval at the lower bins, with increasing *x* values. This translates into increasingly frequent exceedance of the significance interval for these bins, as shown in the second row of Figure 6. Note that this frequency increases at a similar rate to the BS$_R$, which indicates a similar level of statistical power. With respect to case 3, the Reliability Diagram shows increasing deviations at the outer probabilities, but remains reliable at the medial probabilities, with increasing *x*. These deviations at the outer probabilities occur at a similar rate, keeping the forecast unbiased. Although it is clear that this forecast is unreliable from the multibin perspective, single bin analysis is unable to diagnose these errors. Therefore, it is necessary to use a multibin approach when examining the reliability of event forecasts. This motivates the development of a new framework for testing the hypothesis of reliability.

7. Proposed Verification Framework

In order to overcome the inability of the single bin analysis to effectively reject forecasts with unbiased, yet unreliable probabilistic residuals, a multibin verification framework must be developed. Since the multibin

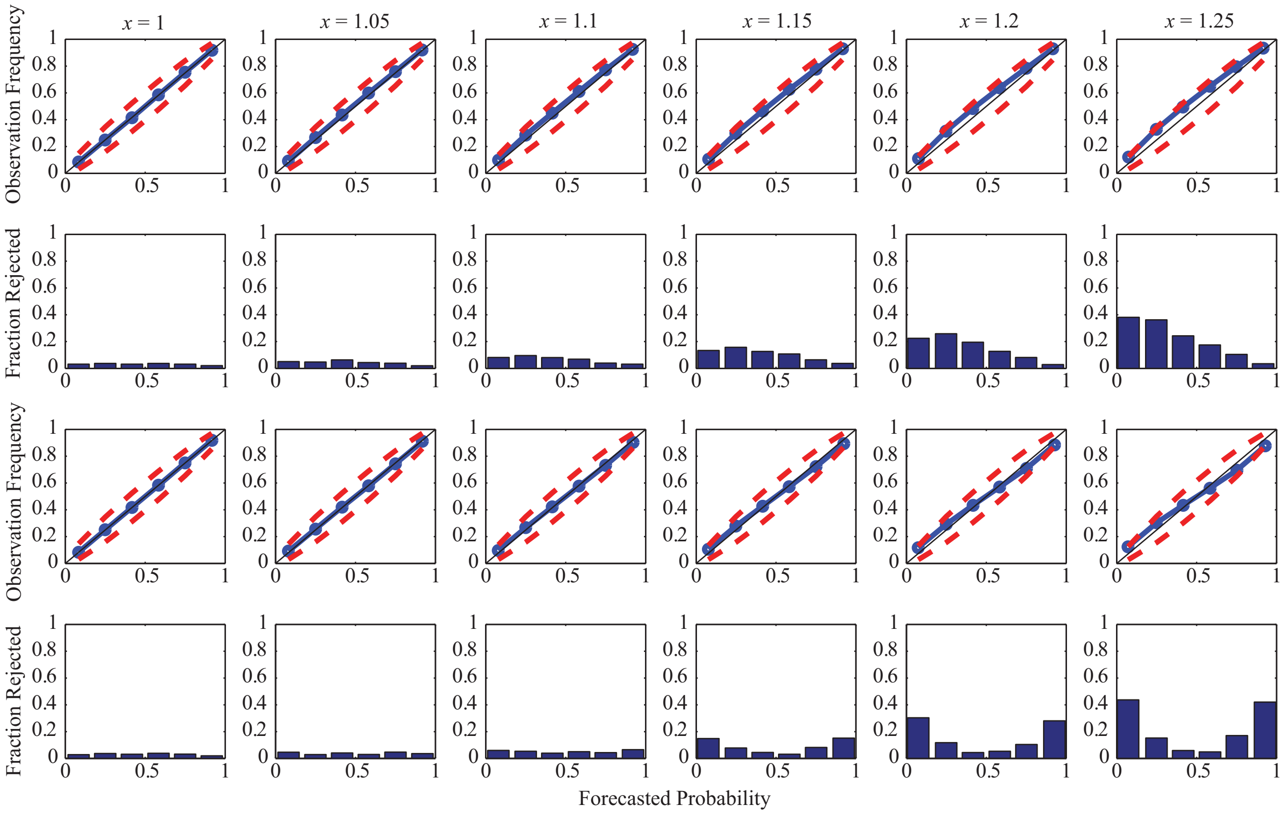

Figure 6. The top row is the median Reliability Diagram (blue line with circles) with the associated 95% significance interval from the Binomial distribution (dotted red line) for 100 replicates of case 2 and varying *x* values (equation (18)). The second row is the fraction of rejected forecasts for each bin. The third row is the same as the top row, but for case 3, and the bottom row is the same as the second row, but for case 3.

approach was shown to reduce statistical power, a two stage approach is proposed: (1) use a single bin analysis to maximize the ability to reject biased probabilistic forecasts, and (2) use a multibin approach to assess unbiased, yet unreliable, probabilistic forecasts. Within this framework, a few considerations must be made. First, the significance level (α) will become complex. Since multiple hypothesis tests will be performed, the forecaster will need to adjust the significance level. For this study, the Šidák correction is selected, which is presented in equations (21) and (22).

$$\alpha_1 = 1 - \sqrt{1-\alpha} \tag{21}$$

$$\alpha_B = 1 - \sqrt[B]{1-\alpha_1} \tag{22}$$

In the above equations, α is the significance level (set to 0.05 in this study), α_1 is the significance level for the single bin stage of the analysis, and α_B is the significance level for each bin of the multibin stage of the analysis. If any of the $B+1$ hypothesis tests reject the null hypothesis, then the hypothesis of reliability is rejected with a minimum significance level of $1-\alpha$.

The multibin stage of the analysis will require the forecaster to determine the appropriate number of bins for verification. A first note is that only even numbers of bins should be considered, as an odd number of bins will have a bin centered around 0.5, which will be sensitive to bias in the forecasted probabilities, and therefore will be unlikely to provide additional information beyond the single bin analysis. In addition, the forecaster should consider the nature of the probabilistic forecast errors when performing the analysis, which requires a discussion of the generation of probabilistic event forecasts.

Probabilistic event forecasts will typically be created with probabilistic forecasts of continuous variables (e.g., precipitation, streamflow, soil moisture). This necessitates forecasting of a continuous probability density. From this density, the forecasted event probability will be the portion of the continuous forecast

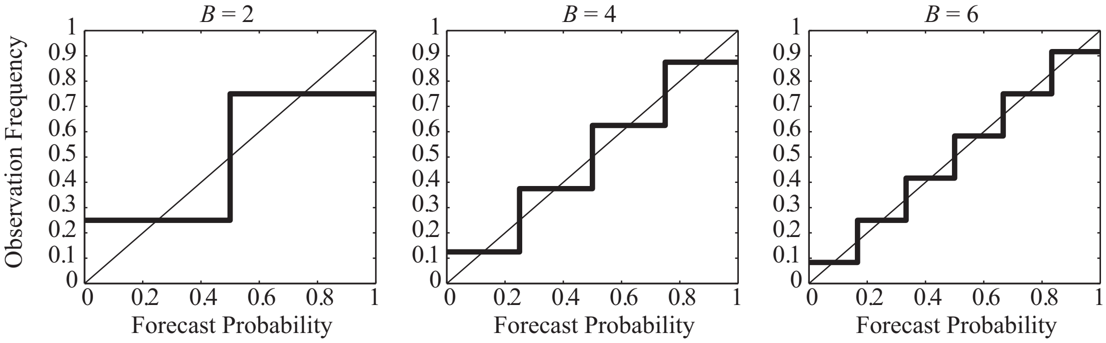

Figure 7. Comparison of the forecasted probabilities and associated observation frequencies for the step-wise function (equation (23)), which is a representation of the worst forecast that is unbiased with observation frequencies that are monotonically nondecreasing with increasing forecast probability.

density exceeding some predefined threshold. Given that the forecast is unbiased, yet unreliable, the most common problem will be continuous forecast densities that have improper variance, leading to an event forecast that is overly certain or uncertain. Such a scenario can be assessed with only two bins, centered at probabilities of 0.25 and 0.75. In the event that the underlying continuous forecast density is unbiased and has proper variance, yet has improperly set higher moments (e.g., skew and kurtosis), the two-bin analysis will be unable reject the hypothesis of reliability. Although this situation poses a potential problem for two-bin analysis, the combination of unbiased forecasts with properly set variance, in conjunction with improper higher-order moments, is expected to be rare. Beyond this assumption of rarity, identifying unreliable forecasts with errors in higher-order moments will require a greater number of bins to identify unreliable forecasts. With this increase in the required number of bins, the necessary number of observations to reject the null hypothesis will grow rapidly. Due to this increase in the required number of observations, an analysis was performed to determine the minimum number of observations that must be available to warrant analysis with different numbers of bins.

In this analysis, the minimum number of observations necessary to reject the hypothesis of reliability for different numbers of bins was estimated. A function was developed that calculates the observation frequency for each bin ($\bar{O}_b$), which creates the maximum possible probabilistic residuals for each bin, and therefore requires the minimum number of observations to identify as unreliable. This function was created under the assumption that the observed frequency is monotonically nondecreasing with increasing forecasted probability, based on the expectation that the forecasted probabilities and observation frequencies are positively correlated, and the probabilistic residuals are unbiased. Therefore, the function must be symmetrical about the bin edges, which will be ensured if it has passed the single bin analysis. Under these assumptions, equation (23) estimates the observation frequencies that create the maximum probabilistic residuals, given an even numbers of bins. Figure 7 shows the function for bin sizes of 2, 4, and 6, for illustrative purposes. Following this equation, the maximum probabilistic residual is given by equation (24).

$$\bar{O}_b = \frac{2b-1}{2B} \tag{23}$$

$$\max\left[\bar{p}_b - \bar{O}_b\right] = \frac{b}{B} - \frac{2b-1}{2B} = \frac{1}{2B} \tag{24}$$

With the maximum probabilistic residuals provided by equation (24), the required number of observations in each bin, N_b, to reject the hypothesis of reliability may be estimated with the inverse Binomial distribution. In order to determine the required number of observations to reject the probabilistic residuals given by equation (23), one may solve equation (25).

$$\frac{1}{2B} = \bar{p}_b - \bar{O}_b \geq \frac{F_B^{-1}\left(0.5, N_b, 0.5 - \frac{1}{2B}\right)}{N_b} - \frac{F_B^{-1}\left(\frac{\alpha_B}{2}, N_b, 0.5 - \frac{1}{2B}\right)}{N_b} \tag{25}$$

In this equation, $F_B^{-1}\left(0.5, N_b, 0.5 - \frac{1}{2B}\right)$ is the inverse of the cumulative Binomial distribution, which solves for the number of event occurrences at the median of the distribution, over N forecasts, with a probability of $0.5 - \frac{1}{2B}$ (center of the bin located immediately below 0.5). Therefore, $\frac{F_B^{-1}\left(0.5, N_b, 0.5 - \frac{1}{2B}\right)}{N_b}$ approximates $\bar{p}_b$, and

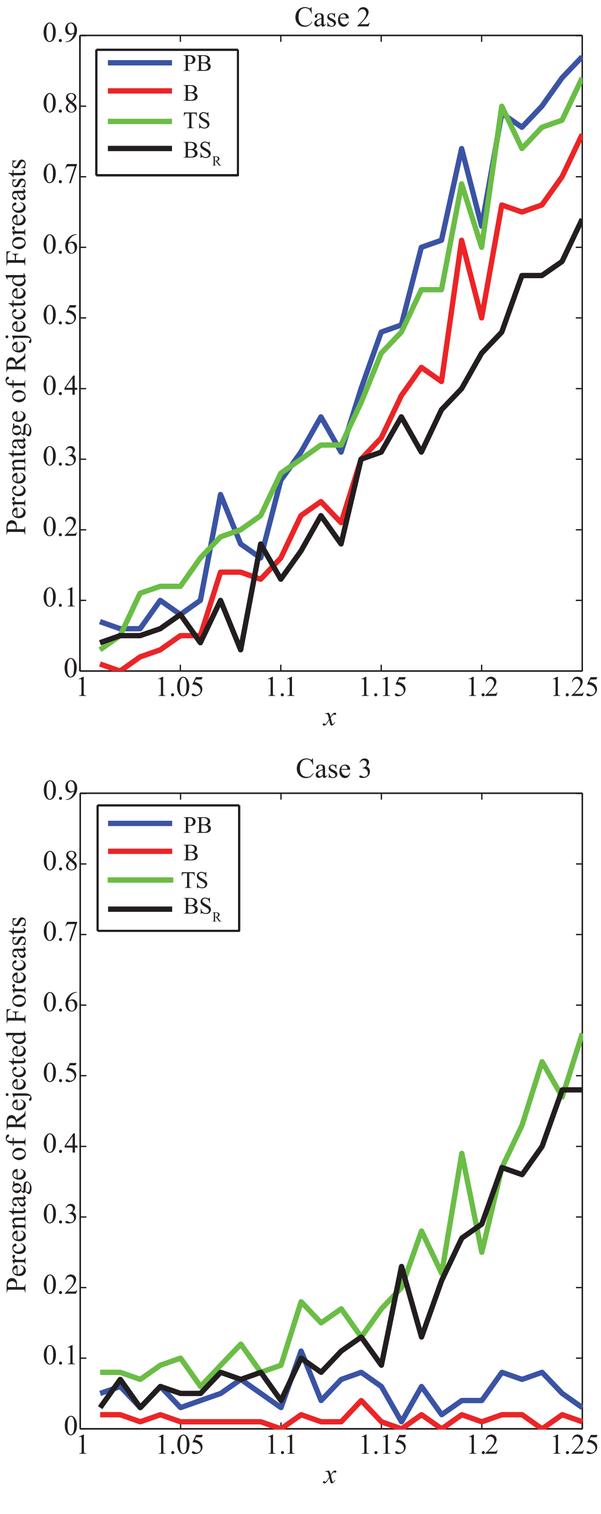

Figure 8. Percentage of rejected forecasts as a function of *x* (equation (18)). PB is the single bin Poisson-Binomial distribution, B is the single bin Binomial distribution, TS is the proposed two-stage verification framework, and BSR is the reliability component of the decomposed Brier Score.

$\frac{F_B^{-1}\left(\frac{\alpha_B}{2},N_b,0.5-\frac{1}{2B}\right)}{N_b}$ is equal to the threshold for $\bar{O}_b$, based on the lower bound of the confidence interval $(\frac{\alpha_B}{2})$ defined in equation (22). This equation was solved numerically for N_b, starting at $N_b=1$ and increasing N by increments of one until the equation is satisfied. For $B=2$, N_b would need to be greater than 12, for $B=4$, N_b would need to be greater than 97, and for $B=6$, N_b would need to be greater than 265. Assuming that the observations are evenly distributed in all bins, the minimum number of observations, summed across all bins, required to reject the hypothesis of reliability would be 24, 388 and 1590, for $B=2$, $B=4$, and $B=6$, respectively. Due to the rarity of scenarios in which more than two bins is warranted, and the rapid growth in minimum required number of observations to reject the hypothesis of reliability, this study proposes that two bins are prudent for the majority of cases.

8. Results With Proposed Verification Framework

The proposed verification framework is compared to the BS_R (with six bins), the Poisson-Binomial distribution and the Binomial distribution in Figure 8. This Figure presents similar results to Figure 5, to ensure consistency in the analysis. From Figure 8, it is clear that the proposed methodology (green line) is comparable to the single bin analysis of the Poisson-Binomial distribution (blue line) for case 2 (solid lines), indicating minimal loss of information when adding a second verification stage. There is a minor loss of information, and this is due to the requirement of decreasing the significance level in the single bin case. The proposed technique still outperforms both the BS_R and Binomial distribution, indicating that this is an effective means to reject biased probabilistic residuals.

With respect to case 3, the proposed method shows the ability to reject the unreliable forecasts. As is expected, the rejection rate increases as the forecasts become increasingly unreliable. Further, the rate at which the unreliable forecasts are rejected with the proposed method increases at a faster rate than the BS_R, which indicates that this method provides more statistical power than the BS_R. As the BS_R and Reliability Diagram were found to reject unreliable forecasts at a similar rate, it can be concluded that the proposed methodology is more effective than the Reliability Diagram in rejecting unreliable forecasts as well. One caveat is that visualization with a Reliability Diagram is useful in diagnosing the form of forecast errors (i.e., overly sharp or insufficiently sharp forecasts), and therefore this methodology will never entirely replace the Reliability Diagram for examining the cause of forecast errors.

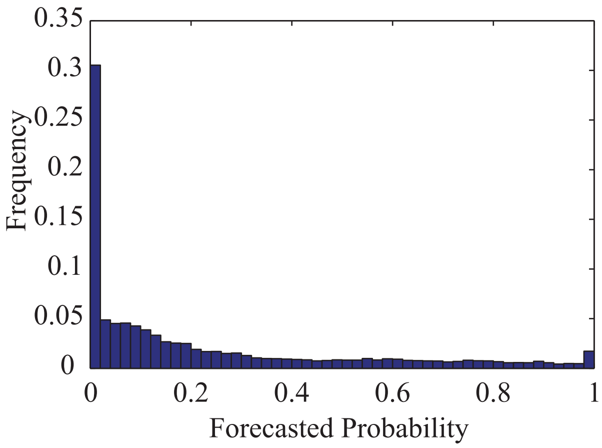

Figure 9. Histogram of the National Weather Service 12 h probability of precipitation forecasts.

9. Case Study: Probability of Precipitation Forecasts

In order to assess the utility of the proposed verification framework on real forecasts, a case study with National Weather Service (NWS) 12 h probability of precipitation forecasts was performed. Probability of precipitation is regularly forecasted by the National Weather Service throughout the United States. This data are archived in the National Digital Forecast Database (NDFD), which may be accessed through the National Operational Model Archive & Distribution System (NOMADS) (http://nomads.ncdc.noaa.gov/data.php#ndfd). For this experiment, forecasts from 1 January 2009 through 31 December 2009 were gathered. For verification, all hourly precipitation gages available through the National Climate Data Center (NCDC), that are located the state of Oregon, USA, were gathered from the NCDC ftp site (ftp://ftp.ncdc.noaa.gov/pub/data/hourly_precip-3240/35/). Within Oregon, there are 108 gages with hourly precipitation observations during the year 2009. At each location, the forecasts were separated between daytime (5 A.M. to 5 P.M.) and nighttime (5 P.M. to 5 A.M), creating two individual sets of forecasts for each location (216 total sets of forecast and observation pairs). This separation was performed to remove any potential inconsistencies between the day and night forecasts/observations. Forecasts at each of the 108 locations, throughout the study period, are shown in a histogram in Figure 9. From this figure, it is clear that

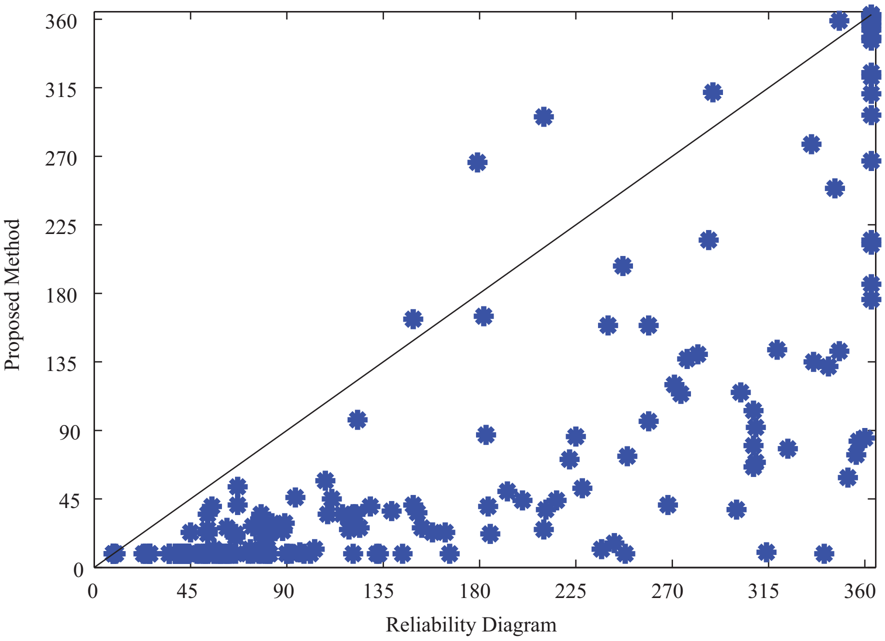

Figure 10. Comparison of the necessary number of observations to reject the forecasts from each of the of the observation rain gages, for the proposed two-stage approach and the Reliability Diagram.

the forecasts assign zero, or near zero, probabilities at a higher rate than any other value. This is reflective of the nature of precipitation throughout Oregon, where the majority of 12 h periods in a given year will not experience precipitation. These forecasts also have an increased frequency at lower probabilities than high probabilities, with the exception of forecasts equal to 100%.

The NWS probability of precipitation forecasts have been well studied [*Bickel et al.*, 2011], and were found to be unreliable, as a whole, throughout the US. Since the forecasts are known to be unreliable, the aim in this section is to compare the ability of the proposed two-stage verification method, the Reliability Diagram, and the BS_R in rejecting the forecasts. In order to compare the statistical power of these techniques, the number of observations required (ranging from 10 to 365) to reject the hypothesis of reliability, for both the proposed two-stage approach and the Reliability Diagram, is compared in Figure 10. In this figure, the horizontal axis shows the number of observations required by the Reliability Diagram to reject the hypothesis of reliability, with 95% significance, with the vertical axis showing this information for the proposed approach, and the black line is the one-to-one line. Note that all but five points lie below the one-to-one line, indicating that for 211 locations, the Reliability Diagram requires more verifying observations than the proposed two-stage approach, to determine that the forecast is unreliable. This indicates that the proposed approach has more statistical power than the Reliability Diagram, allowing for rejection of unreliable forecasts with fewer forecast and observation pairs. Further, this suggests that the assumption of two bins being sufficient in the multibin stage of the proposed approach is valid for this application.

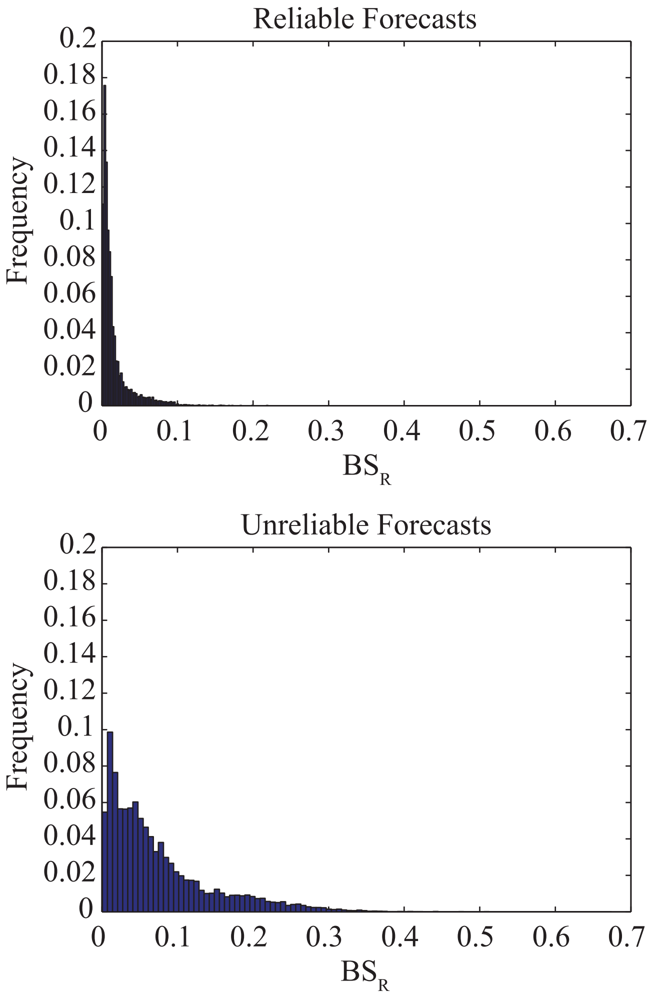

Figure 11. Histograms of the BSR values corresponding to the reliable (Top) and unreliable (Bottom) forecast lengths, as determined by the proposed two-stage method.

In order to compare the BS_R and the proposed two-stage approach, Figure 11 shows the histogram of BS_R values for reliable forecasts (Figure 11, top plot) and for unreliable forecasts (Figure 11, bottom plot). The reliable forecasts in Figure 11 are sampled from each of the 216 forecast sets for which the proposed two-stage approach is unable reject the hypothesis of reliability. For unreliable forecasts, all forecasts for which the proposed two-stage approach was capable of rejecting the hypothesis of reliability were examined. From the two histograms in Figure 11, it is observed that unreliable forecasts have a higher occurrence of large BS_R values, which is expected. Although the unreliable forecasts tend to display larger BS_R values than those of the reliable forecasts, many of the unreliable forecasts have very low BS_R values, indicating that the BS_R may not always be capable of distinguishing between reliable and unreliable forecasts. Due to the knowledge that the BS_R is an approximation of the six bin approach used in the Reliability Diagram, it is expected that the BS_R will be less powerful than the Reliability Diagram, and therefore less powerful than the proposed two-stage approach. Overall this real forecast verification experiment suggests that the proposed two-stage approach is the strictest criteria for determining forecast reliability, supporting the findings from the numerical experiments presented in section 8.

10. Conclusions

Probabilistic forecasting of events has become an important tool for forecasters to represent uncertainty in hydrometeorological applications, allowing forecasters to communicate the certainty of an event occurring. Assuming that these forecasted probabilities are reliable, the end user of that forecast can effectively manage the risk of that event occurring. This necessitates verification that the forecast is reliable, to ensure that event mitigation measures are made on correct information. This has motivated the exploration of reliability assessment in this study.

From a theoretical standpoint, this article showed that the Poisson-Binomial distribution is an exact model of the probabilistic verification setting. Although the Poisson-Binomial distribution is ideal for assessing reliability, it is absent from the hydrometeorological forecast verification literature. Conventional verification tools are based on the Binomial distribution, as an approximation of the Poisson-Binomial distribution. Beyond the Binomial approximation, these tools make further approximations to develop single valued scores (BS_R) and diagrams (Reliability Diagram). This creates two layers of approximations, which have the potential to create errors in reliability assessment. Quantifying the errors resulting from these approximations is a central focus in this article.

The approximation of the Poisson-Binomial distribution, via the Binomial distribution, was found to be a balance between bin size and forecast variability. As forecast variability increases, the necessary number of bins increases, but this increasing number of bins leads to a loss of information. By breaking up the verification problem into multiple different bins, the sample size in each bin is reduced, leading to a loss of statistical power in rejecting unreliable forecasts. Beyond the underlying Binomial approximation, the BS_R was found to further reduce the ability to reject unreliable forecasts. Being based on the binning approach, the BS_R has an upper limit of accuracy equal to the Binomial distribution, but imposes a normal approximation of the Binomial distribution, which will further reduce the statistical power at any practical number of bins. In addition, thresholds of acceptability (significance level) for the BS_R have no analytical solution, and therefore require sampling to estimate for any number of bins and sample size. Accurate estimation of BS_R thresholds are possible in the numerical experiments, but will be difficult for real forecasts. Similarly, the Reliability Diagram is an approximation of the Binomial distribution, except in the case discussed in *Bröcker and Smith* [2007]. These approximations generally reduce the ability to differentiate between reliable and unreliable forecasts.

This article presented experiments that support the hypothesis that the Poisson-Binomial distribution maximizes the forecaster's ability to reject unreliable forecasts. The exception to this conclusion was a forecast that is unreliable, yet unbiased. Although the single bin Poisson-Binomial distribution maximizes the ability to reject biased forecasts, a single bin is insufficient when the unreliable distribution is unbiased. Solving this problem requires a multibin approach, motivating the development of a new verification framework. A two-stage verification framework was proposed, where a single bin analysis is used to maximize the ability to reject biased forecasts, followed by a two-bin approach to reject unbiased, yet unreliable forecasts. Results in section 8 suggest that the proposed framework is effective in identifying both biased and unbiased unreliable forecasts. Further, an examination of a real probabilistic forecast, the NWS 12 h probability of precipitation forecasts, supported the finding that the two-stage approach to reliability assessment, via the Poisson-Binomial distribution, is more powerful in determining reliability than the BS_R and the Reliability Diagram. One caveat is that this method could benefit from further testing in more real data experiments, as the singular real case study examined may not be representative of all forecasts. Although more testing is suggested to confirm these findings, the two-stage approach, via the Poisson-Binomial distribution, was found to be the most statistically powerful of all verification methodologies examined, and is therefore suggested for use when assessing the reliability of probabilistic event forecasts.

Appendix A: Proof That the Variance of the Binomial Distribution Is Greater than or Equal to the Variance of the Poisson-Binomial Distribution

The variance of the Poisson-Binomial distribution is provided in equation (7), and the variance of the Binomial distribution is provided in equation (A1).

$$\sigma_B^2 = T\overline{p_{1:T}}\left(1-\overline{p_{1:T}}\right) \tag{A1}$$

This study suggested that the Poisson-Binomial distribution will have more statistical power than the Binomial distribution, except when all forecasted probabilities are equal, and therefore the inequality in equation (A2) must be proven.

$$\sigma_B^2 = T\overline{p_{1:T}}\left(1-\overline{p_{1:T}}\right) \geq \sum_{t=1}^{T} p_t(1-p_t) = \sigma_{PB}^2 \tag{A2}$$

Equation (A2) may then be expanded to equation (A3).

$$T\overline{p_{1:T}} - T\overline{p_{1:T}}^2 \geq \sum_{t=1}^{T} p_t - \sum_{t=1}^{T} {p_t}^2 \tag{A3}$$

By definition, $\sum_{t=1}^{T} p_t = T\,\overline{p_{1:T}}$, and therefore equation (A3) simplifies to equation (A4).

$$\sum_{t=1}^{T} {p_t}^2 \geq T\,\overline{p_{1:T}}^2 \tag{A4}$$

At this point, the left-hand side of this equation may be expanded according to equation (A5), as there will be a set of $\Delta_{1:T}$ that satisfy both $p_t = \overline{p_{1:T}} + \Delta_t$ and $\sum_{t=1}^{T} \Delta_t = 0$.

$$\begin{aligned} &\sum_{t=1}^{T} {p_t}^2 = \sum_{t=1}^{T} \left(\overline{p_{1:T}} + \Delta_t\right)^2 = \sum_{t=1}^{T} \left(\overline{p_{1:T}}^2 + 2\overline{p_{1:T}}\Delta_t + {\Delta_t}^2\right) = \\ &\sum_{t=1}^{T} \left(\overline{p_{1:T}}^2 + {\Delta_t}^2\right) = \sum_{t=1}^{T} \overline{p_{1:T}}^2 + \sum_{t=1}^{T} {\Delta_t}^2 = T\overline{p_{1:T}}^2 + \sum_{t=1}^{T} {\Delta_t}^2 \end{aligned} \tag{A5}$$

Equation (A6) can be found by substituting the right-hand side of equation (A5) into equation (A4) and subtracting $T\overline{p_{1:T}}^2$ from both sides.

$$\sum_{t=1}^{T} {\Delta_t}^2 \geq 0 \tag{A6}$$

Equation (A6) will only reach equality in the event that all Δ_t are 0, and therefore the variance of the Binomial distribution will always be greater than that of the Poisson-Binomial distribution, except in the scenario where all forecasted probabilities are equal. In addition, equation (A6) shows that the difference between the variance of the Binomial and Poisson-Binomial distribution will grow as the forecasts increases in sharpness (tendency toward forecasting either 0 or 1), and therefore the $\sum_{t=1}^{T} {\Delta_t}^2$ increases.

Acknowledgments
The authors would like to acknowledge the financial support provided by NOAA-MAPP grant NA11OAR4310140 and also NOAA-CSTAR grant NA11NWS4680002. All data were gathered from the National Digital Forecast Database (NDFD), which was accessed through the National Operational Model Archive & Distribution System (NOMADS) (http://nomads.ncdc.noaa.gov/data.php#ndfd).

References

Annan, J. D., and J. C. Hargreaves (2010), Reliability of the CMIP3 ensemble, *Geophys. Res. Lett.*, *37*, L02703, doi:10.1029/2009GL041994.
Bickel, J. E. (2007), Some comparisons among quadratic, spherical, and logarithmic scoring rules, *Decision Analysis*, *4*(2), 49–65.
Bickel, J. E., E. Floehr, and S. D. Kim (2011), Comparing NWS PoP forecasts to third-party providers, *Mon. Weather Rev.*, *139*(11), 3304–3321.
Bradley, A. A., S. S. Schwartz, and T. Hashino (2004), Distributions-oriented verification of ensemble streamflow predictions, *J. Hydrometeorol.*, *5*(3), 532–545.
Brier, G. W. (1950), Verification of forecasts expressed in terms of probability, *Mon. Wea. Rev.*, *78*(1), 1–3.
Bröcker, J. (2009), Reliability, sufficiency, and the decomposition of proper scores, *Q. J. R. Meteorol. Soc.*, *135*(643), 1512–1519.
Bröcker, J., and L. A. Smith (2007), Increasing the reliability of reliability diagrams, *Weather Forecast.*, *22*(3), 651–661.
Buizza, R., A. Hollingsworth, F. Lalaurette, and A. Ghelli (1999), Probabilistic predictions of precipitation using the ECMWF ensemble prediction system, *Weather Forecast.*, *14*(2), 168–189.
Carriquiry, M. A., and D. E. Osgood (2012), Index insurance, probabilistic climate forecasts, and production, *J. Risk Insur.*, *79*(1), 287–300.
Christensen, H. M., I. M. Moroz, and T. N. Palmer (2015), Evaluation of ensemble forecast uncertainty using a new proper score: Application to medium-range and seasonal forecasts, *Q.J.R. Meteorol. Soc.*, *141*, 538–549, doi:10.1002/qj.2375.
DeChant, C. M., and H. Moradkhani (2014), Hydrologic prediction and uncertainty quantification, in *Handbook of Engineering Hydrology*, edited by S. Eslamian, pp. 387–414, CRC press, Taylor & Francis, Fla.
Demargne, J., et al. (2014), The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, *Bull. Amer. Meteor. Soc.*, *95*, 79–98.
Feller, W. (1945), On the normal approximation to the binomial distribution, *Ann. Math. Stat.*, *16*(4), 319–329.
Gigerenzer, G., R. Hertwig, E. Van Den Broek, B. Fasolo, and K. V. Katsikopoulos (2005), ''A 30% chance of rain tomorrow'': How does the public understand probabilistic weather forecasts?, *Risk Anal.*, *25*(3), 623–629.
Gneiting, T., F. Balabdaoui, and A. E. Raftery (2007), Probabilistic forecasts, calibration and sharpness, *J. R. Stat. Soc. Ser. B*, *69*(2), 243–268.
Gneiting, T., and A. E. Raftery (2007), Strictly proper scoring rules, prediction, and estimation, *J. Am. Statist. Assoc.*, *102*(477), 359–378.
Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *J. Hydrol.*, *377*(1), 80–91.
Hamill, T. M. (2001), Interpretation of rank histograms for verifying ensemble forecasts, *Mon. Weather Rev.*, *129*(3), 550–560.
Hamill, T. M. (2012), Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States*, *Mon. Wea. Rev.*, *140*(7), 2232–2252.

Hodges, J. L., Jr., and L. Le Cam (1960), The Poisson approximation to the Poisson binomial distribution, *Ann. Math. Stat.*, *31*(3), 737–740.

Hong, Y. (2013), On computing the distribution function for the Poisson binomial distribution, *Comput. Stat. Data Anal.*, *59*(0), 41–51.

Jolliffe, I. T., and C. Primo (2008), Evaluating rank histograms using decompositions of the chi-square test statistic, *Mon. Wea. Rev.*, *136*(6), 2133–2139.

Joslyn, S., and S. Savelli (2010), Communicating forecast uncertainty: Public perception of weather forecast uncertainty, *Meteorol. Appl.*, *17*(2), 180–195.

Julie, D., et al. (2014), The Science of NOAA's Operational Hydrologic Ensemble Forecast Service, *Bull. Amer. Meteor. Soc.*, *95*, 79–98.

Laio, F., and S. Tamea (2007), Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrol. Earth Syst. Sci.*, *11*(4), 1267–1277.

Machete, R. L. (2013), Early warning with calibrated and sharper probabilistic forecasts, *J. Forecast.*, *32*(5), 452–468.

Madadgar, S., and H. Moradkhani (2013), A Bayesian framework for probabilistic seasonal drought forecasting, *J. Hydrometeorol.*, *14*(6), 1685–1705.

Mason, S. J. (2004), On using "climatology" as a reference strategy in the Brier and ranked probability skill scores, *Mon. Weather Rev.*, *132*(7), 1891–1895.

Mitchell, J., and K. F. Wallis (2011), Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness, *J. Appl. Econ.*, *26*(6), 1023–1040.

Murphy, A. H. (1973), A new vector partition of the probability score, *J. Appl. Meteorol.*, *12*(4), 595–600.

Murphy, A. H., S. Lichtenstein, B. Fischhoff, and R. L. Winkler (1980), Misinterpretations of precipitation probability forecasts, *Bull. Amer. Meteorol. Soc.*, *61*(7), 695–701.

Pappenberger, F., J. Thielen, and M. Del Medico (2011), The impact of weather forecast improvements on large scale hydrology: Analysing a decade of forecasts of the European Flood Alert System, *Hydrol. Processes*, *25*(7), 1091–1113.

Park, Y.-Y., R. Buizza, and M. Leutbecher (2008), TIGGE: Preliminary results on comparing and combining ensembles, *Q. J. R. Meteorol. Soc.*, *134*(637), 2029–2050.

Saha, S., et al. (2006), The NCEP climate forecast system, *J. Clim.*, *19*(15), 3483–3517.

Schaake, J. C., T. M. Hamill, R. Buizza, and M. Clark (2007), HEPEX: The hydrological ensemble prediction experiment, *Bull. Am. Meteorol. Soc.*, *88*(10), 1541–1547.

Slingo, J., and T. Palmer (2011), Uncertainty in weather and climate prediction, *Philos. Trans. R. Soc. A*, *369*(1956), 4751–4767.

Stephenson, D. B., A. S. C. Caio, and I. T. Jolliffe (2008), Two extra components in the Brier score decomposition, *Weather Forecast.*, *23*(4), 752–757.

Wetterhall, F., et al. (2013), HESS opinions "Forecaster priorities for improving probabilistic flood forecasts", *Hydrol. Earth Syst. Sci.*, *17*(11), 4389–4399.

Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences: An Introduction*, 2nd ed., xvii, 627 p., Academic, Oxford.

Yan, H., and H. Moradkhani (2015), A regional Bayesian hierarchical model for flood frequency analysis, *Stoch. Environ. Res. Risk Assess.*, *29*(3), 1019–1036, doi:10.1007/s00477-014-0975-3.