CORRESPONDENCE

Comments on “Conditional Exceedance Probabilities”

JOCHEN BRÖCKER, STEFAN SIEGERT, AND HOLGER KANTZ

Max-Planck-Institut für Physik komplexer Systeme, Dresden, Germany

(Manuscript received 21 December 2010, in final form 6 March 2011)

ABSTRACT

In a recent paper, Mason et al. propose a reliability test of ensemble forecasts for a continuous, scalar verification. As noted in the paper, the test relies on a very specific interpretation of ensembles, namely, that the ensemble members represent quantiles of some underlying distribution. This quantile interpretation is not the only interpretation of ensembles, another popular one being the Monte Carlo interpretation. Mason et al. suggest estimating the quantiles in this situation; however, this approach is fundamentally flawed. Errors in the quantile estimates are *not* independent of the exceedance events, and consequently the conditional exceedance probabilities (CEP) curves are not constant, which is a fundamental assumption of the test. The test would reject reliable forecasts with probability much higher than the test size.

In their paper “Conditional exceedance probabilities,” Mason et al. (2007) propose a new approach to testing the reliability of ensemble forecasts for a continuous, scalar verification. Roughly speaking, the null hypothesis of the test is that the k th ensemble member is the $k/(K+1)$ quantile of the distribution of the verification, conditioned on the ensemble. Here, K is the total number of ensemble members. To be more specific, let $X_1 \dots X_K$ denote the ensemble members, which are random variables with values in $\mathbb{R}$ and so that $X_1 < \dots < X_K$ with probability 1. The verification is a random variable Y with values in $\mathbb{R}$. The null hypothesis states that

$$\mathbb{P}(Y > X_k | X_k) = 1 - \frac{k}{K+1}. \quad (1)$$

The left-hand side is what Mason et al. call the conditional exceedance probability (CEP).

The null hypothesis is a consequence of reliability if we impose what could be termed the *quantile* interpretation of ensembles. This states that there is an underlying random distribution function Γ , the *probability forecast*, so that the k th ensemble member is the $k/(K+1)$ quantile of Γ , that is, the unique solution of the equation $\Gamma(x) =$

$k/(K+1)$ for x . It needs to be assumed that Γ is strictly monotonically increasing for the ensemble members to be well defined. The probability forecast Γ is *reliable* if

$$\mathbb{P}(Y < x | \Gamma) = \Gamma(x). \quad (2)$$

Together with a distribution for Γ (which is unimportant in the present context) these determinations completely specify a probabilistic model for $(Y, X_1, \dots, X_K, \Gamma)$. The reliability condition (2) together with the quantile interpretation imply the null hypothesis (1), as we now show. Since the ensemble members are functions of Γ , we have

$$\mathbb{P}(Y > X_k | \Gamma, X_k) = \mathbb{P}(Y > X_k | \Gamma) = \mathbb{P}\left[Y > \Gamma^{-1}\left(\frac{k}{K+1}\right) \middle| \Gamma\right]. \quad (3)$$

If reliability holds though, we can write

$$\mathbb{P}\left[Y > \Gamma^{-1}\left(\frac{k}{K+1}\right) \middle| \Gamma\right] = 1 - \Gamma\left[\Gamma^{-1}\left(\frac{k}{K+1}\right)\right] = 1 - \frac{k}{K+1}.$$

Substituting with this in (3) and taking the expectation conditioned on X_k , we obtain

$$\mathbb{P}(Y > X_k | X_k) = 1 - \frac{k}{K+1},$$

Corresponding author address: Jochen Bröcker, MPIP-KS, Noethnitzer Strasse 38, Dresden, Germany 01187.
E-mail: broecker@pks.mpg.de

which is the null hypothesis (1).

The interpretation of ensembles as quantiles is not the only possible interpretation. An alternative to the quantile interpretation is what we will call the Monte Carlo interpretation, which states that the ensemble members are independent draws (i.e., a sample) from some underlying distribution. The quantile and Monte Carlo interpretation are not dissimilar, and for many applications, the difference between the two interpretations is unimportant. (We are considering the case of verifications in one dimension only; in higher dimensions, the quantile interpretation of course ceases to apply.) For a reliability test through CEPs, though, the difference causes problems. Although the Monte Carlo interpretation was mentioned in Mason et al. we feel that the ensuing problems were not sufficiently highlighted.

Equation (1) cannot be applied directly if the quantiles are unknown (i.e., if the quantile interpretation does not apply). In this case, Mason et al. suggest to replace the (unknown) quantiles with suitable estimates. More precisely, Mason et al. state that

There are two sources of sampling errors in estimating the regression parameters: sampling errors arising from an insufficient number of forecasts, and inaccuracies in estimating the quantiles of the ensemble distribution. The first source of error is common to all verification methods, but better estimates of the quantiles could be obtained by increasing the ensemble size or by fitting a distribution to the ensemble members and calculating the quantiles of the fitted distribution (provided that a distribution can be found that estimates the quantiles well). The CEPs are therefore likely to be estimated most accurately given large ensemble sizes, and for those ensemble members close to the median.

The problem with this approach results from the fact that the quantile estimates would appear both in the conditioning of the CEPs as well as in the definition of the exceedance event itself. The null hypothesis (1) is equivalent to stating that the exceedance events $E_k := (Y > X_k)$ and the exact quantile X_k are independent. [Two random variables Y_1, Y_2 are independent if and only if $\mathbb{P}(Y_1 \in A | Y_2)$ is independent of Y_2 .] As we will show presently, *estimates* of the quantiles will in general *not* be independent of the exceedance events, even if the reliability condition (2) holds. Therefore, variations in the quantile estimates will influence the CEP estimates in a systematic way. Thus, the CEP curves are not expected to be flat even if the forecast is reliable.

Let for all $k = 1 \dots K$ denote ξ_k estimates of the quantiles. If the ensemble is interpreted as Monte Carlo, we can assume the ξ_k to be functions of the order statistics, since they are sufficient for any continuous distribution

FIG. 1. CEPs for all 24 members of the perfect Monte Carlo ensemble (black lines). The topmost curve corresponds to the smallest member ξ_1 , the bottommost corresponds to the largest member ξ_K . The bold line indicates the climatological probability of exceedance. The support of the CEP curves provides a rough guidance as to the variability of ξ_n .

function. Therefore, assuming¹ that the reliability condition (2) holds, we have

$$\mathbb{P}(Y > \xi_k | \Gamma, \xi_k) = 1 - \Gamma(\xi_k). \quad (4)$$

Taking expectation conditioned on ξ_k , this gives

$$\mathbb{P}(Y > \xi_k | \xi_k) = 1 - \mathbb{E}[\Gamma(\xi_k) | \xi_k]. \quad (5)$$

The right-hand side is constant only if the function ξ_k is exactly equal to a quantile of Γ , or more precisely, to a quantile of $\mathbb{E}[\Gamma(x) | \xi_k]$. Any variability of $\mathbb{P}(Y > \xi_k | \xi_k)$ however will give systematically sloped CEP curves, different from what Mason et al. claim.

We illustrate our arguments by constructing an artificial, perfectly reliable, Monte Carlo ensemble as follows: at each time instance, we draw a number v from a uniform distribution on $[0, 1]$. We construct the ensemble $\xi_1 \dots \xi_K$ and the verification Y at each instance by adding independent and identically distributed (iid) random variables to v , drawn from a Gaussian distribution with zero mean and unit variance. Our dataset comprised 10 000 instances, and we used ensembles of size $K = 24$. This ensemble is reliable per construction.

A reasonable estimator for the k th quantile is in fact the k th order statistic itself. (The quotation above suggests that Mason et al. are primarily thinking of this case.) For more quantitative information about the order statistics as quantile estimators see Theorem 14 of chapter VI in Mood et al. (1974). For each individual ensemble member ξ_k we

¹ A further assumption here is that Γ is sufficient for ξ_k , that is, ξ_k are given by Γ and possibly some further randomization independent of Y .

FIG. 2. CEPs for the median of ensembles of different sizes ($K = 10, 20, \dots, 100$). There is significant dependence on the size of the median for all values of K .

obtain exceedance events $E_k = (Y > \xi_k)$. Using the same methodology as Mason et al. we use logistic models with two parameters (intercept and slope) to fit the CEPs using maximum likelihood. Finally, we evaluate the resulting model on all instances in the dataset.

The results are presented in Fig. 1. Visually, these CEP curves are far from flat. As suggested in Mason et al. the logistic model can be tested for zero slope using a χ^2 test. These tests produced p values very close to zero for all curves displayed in Fig. 1. The ensemble would be interpreted as unreliable according to Mason et al.

In Fig. 2 we present the CEP curves for the median of ensembles of different sizes. Again, there is significant dependence on the value of the median for all K . We confirm that the CEP depends less and less on the value of the median as the ensemble gets larger. For the ensemble sizes shown in Fig. 2 though, the p values (testing for zero slope) are still very close to zero. Beyond ensemble sizes of approximately 200, the p value for zero slope becomes greater than 0.1. For comparison, we also analyzed an ensemble comprising exact quantiles. The resulting CEPs (estimated with logistic regression) are

FIG. 3. CEPs for an exact quantile ensemble of 24 members. Here, the CEPs are indeed flat (up to random fluctuations), as expected. The p values are essentially uniformly distributed in this case.

shown in Fig. 3. These CEPs are indeed flat, as expected from the analysis. The p values for zero slope are essentially uniformly distributed in this situation. This demonstrates that the general conclusion of this paper is not a mere artifact of the numerical example.

The conclusion of this work is that the CEP test for reliability cannot be applied if the quantiles are not known, or more precisely, if the ensembles are not interpreted as quantiles, and the latter have to be estimated. The reason is that errors in the quantile estimates are not independent of the exceedance events, and the CEP curves are therefore not constant, which is a fundamental assumption of the test. In a simple numerical example, the test would reject the null hypothesis even though the ensemble was by construction reliable.

REFERENCES

- Mason, S. J., J. S. Galpin, L. Goddard, N. E. Graham, and B. Rajartnam, 2007: Conditional exceedance probabilities. *Mon. Wea. Rev.*, **135**, 363–372.
- Mood, A. M., F. A. Graybill, and D. C. Boes, 1974: *Introduction to the Theory of Statistics*. 3rd ed. McGraw-Hill, 480 pp.