Reply

SIMON J. MASON AND MICHAEL K. TIPPETT

International Research Institute for Climate and Society, The Earth Institute of Columbia University, Palisades, New York

ANDREAS P. WEIGEL

Federal Office of Meteorology and Climatology, MeteoSwiss, Zurich, Switzerland

LISA GODDARD

International Research Institute for Climate and Society, The Earth Institute of Columbia University, Palisades, New York

BALAKANAPATHY RAJARATNAM

Department of Statistics, Department of Environmental Earth System Science, The Woods Institute for the Environment, Stanford University, Stanford, California

(Manuscript received and in final form 23 February 2011)

The rank histogram (Anderson et al. 1996) is a widely used procedure for evaluating the reliability of an ensemble forecast system (Jolliffe and Stephenson 2003). It indicates the probability that the observed value exceeds the kth of the K-ordered ensemble members, and is less than the $k + 1$th-ordered ensemble member, with additional bins to indicate the probabilities that the observed value is less than the smallest and greater than the largest ensemble member values, respectively. If the forecasts are reliable, then a graph of these probabilities will show equal values for all bins. The probability integral transform (Dawid 1984) is a similar procedure suitable for situations in which the forecast is presented as a continuous probability distribution function. A similar histogram is drawn, but the bins are based on quantiles of the distribution rather than on the ordered ensemble members. Both the ranked histogram and the probability integral transform can be presented equivalently in terms of exceedance probabilities [i.e., the probability that an observation exceeds the kth of K-ordered ensemble members (or quantile, in the case of the probability integral transform)]. In a reliable forecast system the observed values should exceed the kth of K-ordered ensemble members $[1 - k/(K + 1)] \times 100\%$ of the time. A graph of these exceedance probabilities will step downward evenly to the right.

It is generally recognized that, a uniform rank histogram or probability integral transform, provides no guarantee that the ensemble system being tested indicates a conditionally unbiased forecasting system (Hamill 2001; Gneiting et al. 2007). For example, imagine a forecast system whose ensemble mean is negatively correlated with a normally distributed random variable. These forecasts are conditionally biased, but if the ensemble has the ''right'' spread they will, on average, seem as if they have come from the same distribution as the observations and will have a uniform rank histogram Specifically, if the observed values have a variance of a, the ensemble mean has a variance of b and the covariance between the observations and the ensemble mean is c, then with an ensemble variance of $a + b - 2c$ the ranked histogram will be approximately uniform. Conditional exceedance probabilities (CEPs) were suggested by Mason et al. (2007) as a procedure for identifying those cases when a uniform ranked histogram was obtained from a conditionally biased forecast system. Bröcker et al. (2011) point out a problem with this test, and our purpose in responding is twofold: first, to emphasize that regardless of this problem CEPs remain valuable diagnostics of cases in which uniform rank histograms may be derived from conditionally biased forecast systems; and second, to propose a simple correction to the way in

Corresponding author address: Dr. Simon J. Mason, IRI, 61 Route 9W, P.O. Box 1000, Palisades, NY 10964-8000.
E-mail: simon@iri.columbia.edu

DOI: 10.1175/2011MWR3659.1

which the CEPs are calculated, which eliminates the problem anyway.

One of the aims in calculating CEPs is to test for ''complete calibration'' (or ''complete reliability'')—do subsets of the forecasts asymptote to uniform rank histograms (Seillier-Moiseiwitsch and Dawid 1993)? Complete calibration requires that the forecast system is both conditionally and unconditionally unbiased. In an incompletely calibrated (or ''incompletely reliable'') forecast system it can be expected that the probability of exceeding the highest-ranked ensemble member, for example, increases as the value of that forecast decreases and decreases as the value increases. As a result, rank histograms calculated only for forecasts when the central tendency of the ensemble is above the climatological median are likely to slope downward to the left, or slope downward to the right when the central tendency is below the median. While there are unquestionably multiple reasons for verifying forecasts (Jolliffe and Stephenson 2003), one of the most important reasons is to tell us something about how to interpret the current forecast. If a forecast system demonstrates reliability on average, but cannot be assumed to be well calibrated for the current forecast, its reliability would seem to be of limited interest. Knowledge about its complete calibration would be more helpful (Held et al. 2010). For example, in a nine-member ensemble system, if the rank histogram indicates that the probability of exceeding the highest-ranked ensemble member is 10%, can we then assume that the probability of exceeding the current highest-ranked ensemble member is also 10%? The concepts of complete calibration and of CEPs are designed to answer questions like this one.

What the CEP curves, as described by Mason et al. (2007), will do successfully, is to diagnose cases in which an unreliable forecast system generates a uniform-ranked histogram or probability transform integral. In a forecast system in which variability in the median has no association with variability in the observed values, the CEPs will closely follow the climatological exceedance probabilities. Furthermore, in a forecast system in which the ensemble median is negatively associated with the observed values, the slope of the CEP curves will exceed that of the climatological exceedance probability curve. While the curves flatten as conditional biases are reduced, as Bröcker et al. (2011) demonstrate, the problem is that the curves do not become flat if complete reliability is achieved. We agree with their results, and confirm that not only will CEPs fitted following Mason et al. (2007) not be constant (and hence the curves will not be flat) when calculated on ranked ensemble members, but they will also not even be constant on quantiles from a distribution fitted to the ensemble members. Thus, even in a system that is completely reliable by design, the CEP curves will still slope downward. However, even if this problem could not be addressed, CEP curves would still be very useful for comparison of improvements, or deteriorations, of forecast systems.

Bröcker et al. (2011) explain that a positive sampling error in estimating a quantile of the forecast distribution will decrease the exceedance probability, while any negative sampling error will increase the exceedance probability. Because the exceedance probability is a function of the sampling error, the curves are therefore not flat even in a completely calibrated forecast system. However, if the CEPs could be calculated so that they are independent of the sampling errors in the quantiles, then the curves can become flat. Following the notation of Bröcker et al. (2011), Mason et al. (2007) define the CEP for the kth quantile as

$$P(Y > \xi_k | \xi_k) = \frac{\exp(\beta_{0,k} + \beta_{1,k}\xi_k)}{1 + \exp(\beta_{0,k} + \beta_{1,k}\xi_k)}, \qquad (1)$$

where $\beta_{0,k}$ and $\beta_{1,k}$ are parameters to be estimated, and ξ_k is the kth quantile estimate (whether obtained from the values of the ranked ensemble members or from a fitted distribution). If the ensemble is divided randomly into two halves, A and B, and then independent quantile estimates are obtained from these two halves, the CEP can then be calculated as

$$P(Y > \xi_{k/2,A} | \xi_{k/2,B}) = \frac{\exp(\beta_{0,k/2} + \beta_{1,k/2}\xi_{k/2,B})}{1 + \exp(\beta_{0,k/2} + \beta_{1,k/2}\xi_{k/2,B})}, \qquad (2)$$

where $\xi_{k/2,A}$ is the quantile estimate from the first division, while $\xi_{k/2,B}$ is the corresponding estimate from the second division. Since it is arbitrary which of the two divisions is A and which is B, separate parameter estimates of the CEP curves could be made. Repeated random divisions of the ensemble could also be conducted to obtain additional estimates as a check for sampling uncertainty. The reduction of the size of the ensemble by half is clearly undesirable, and will result in an increase in the sampling errors in the quantile estimates. It may be possible to offset this effect, at least partially, by repeating the data so that for each of the original observed values there are two quantile estimates and definitions of an exceedance event. Discussion of what the resulting sample size would be is beyond the scope of this short reply, as are more detailed discussions of the best way to minimize the additional sampling errors introduced by dividing the ensemble. Instead, we illustrate in Fig. 1 that the CEP curves defined

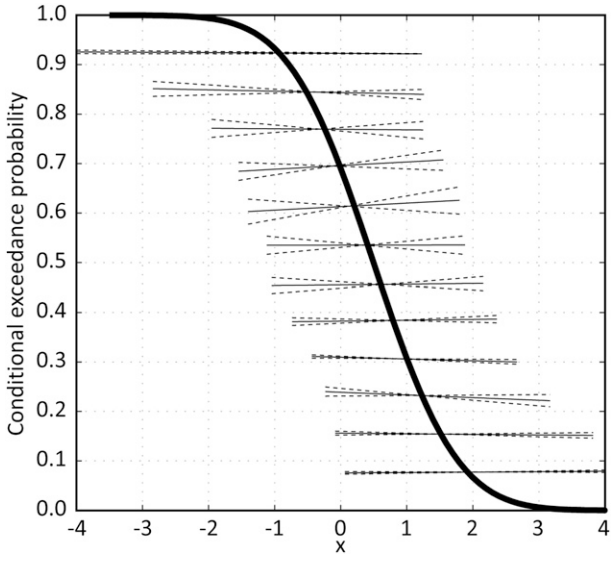

FIG. 1. CEP curves for all 12 of the perfect Monte Carlo ensembles, after division of the ensembles, when the curves are calculated using independent estimates of the quantile and the exceedance events. The bold line indicates the climatological probability of exceedance. The dashed lines indicate the CEPs if ensemble division *A* and *B* are swapped. The solid lines indicate the CEPs if the data are repeated and the ensemble divisions are swapped for the repeat.

using Eq. (2) for the completely calibrated forecast system described by Bröcker et al. (2011) do become flat. The definition of the CEP provided in Eq. (2) raises questions about whether flat curves indicate completely calibrated forecasts, or only forecasts that would be completely calibrated if there were no sampling uncertainty in estimating the quantiles. However, our conclusion is that it is possible to obtain independent estimates of the quantiles and the exceedance events, so that the CEP test for reliability can be applied.

Acknowledgments. This work was funded by grant/cooperative agreements from the National Oceanic and Atmospheric Administration (NA10OAR4310210 and NA05OAR4311004), by the Swiss National Science Foundation through the National Centre for Competence in Research (NCCR) Climate, and by NSF Grants DMS-0906392, DMS-CMG-1025465, AGS-1003823, and SU-WI-EVP2010-04. The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies.

REFERENCES

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

Bröcker, J., S. Siegert, and H. Kantz, 2011: Comments on "Conditional exceedance probabilities." *Mon. Wea. Rev.,* **139,** 3322–3324.

Dawid, A. P., 1984: Statistical theory: The prequential approach. *J. Roy. Stat. Soc.,* **147A,** 278–292.

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

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

Held, L., K. Rufibach, and F. Balabdaoui, 2010: A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics,* **66,** 1295–1305.

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, 240 pp.

Mason, S. J., J. S. Galpin, L. Goddard, N. E. Graham, and B. Rajaratnam, 2007: Conditional exceedance probabilities. *Mon. Wea. Rev.,* **135,** 363–372.

Seillier-Moiseiwitsch, F., and A. P. Dawid, 1993: On testing the validity of sequential probability forecasts. *J. Amer. Stat. Assoc.,* **88,** 355–359.