Hydrol. Earth Syst. Sci. Discuss., 8, C1941–C1947, 2011
www.hydrol-earth-syst-sci-discuss.net/8/C1941/2011/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Evaluating uncertainty estimates in hydrologic models: borrowing measures from the forecast verification community" *by* K. J. Franz and T. S. Hogue

J.D. Brown (Referee)

james.d.brown@noaa.gov

Received and published: 3 June 2011

This paper provides an overview of techniques for verifying probabilistic forecasts of hydrologic variables and illustrates their application to ensemble simulations of streamflow from three different hydrologic model parameter estimation schemes. As the authors rightly mention, probabilistic verification is underutilized in hydrology, despite the plethora of techniques available to estimate hydrologic uncertainties, while other disciplines (notably, the atmospheric sciences) have a rich history of forecast verification. Thus, applications and extensions of probabilistic verification techniques in hydrology

are important and must be welcomed. The paper is generally well written (from a non-technical standpoint) and is appropriate for publication in HESS. However, I have several major criticisms and suggestions for the authors to consider prior to publication. My overall recommendation is major revision, and I would strongly encourage the authors to resubmit, given the potential for a very useful contribution. The major points follow, with technical corrections listed afterwards:

- The introduction has two major weaknesses. First, there is insufficient coverage of the diversity of verification techniques and measures that originate from outside hydrology and, specifically, from the atmospheric sciences. Given the title of the paper, one would expect to see some evidence of the rich history of probabilistic verification from the atmospheric sciences and some of the challenges associated with "borrowing" measures for hydrologic applications. Are there unique challenges for hydrologic verification? If so, what are these challenges? The introduction need not answer these questions, but should at least pose them for later discussion. It would seem that these questions must be addressed if the paper is going to do more than exemplify the application of existing verification metrics to hydrologic variables (which has been done before). Secondly, while the focus of the paper is primarily on verification, the introduction could better distinguish between the source-based approach to quantifying uncertainty, whereby uncertainties from specific sources (model inputs, parameters, structure etc.) are propagated through a model structure, and purely empirical techniques that aim to capture uncertainties via the joint probability distribution of the observed and forecast variables (of course, there is overlap between these categories). This is relevant not only in the context of uncertainty estimation, but also verification, as the latter problem of "statistical post-processing" is concerned with the same joint probability distribution.
- I recommend that the authors re-think their experimental design and case study. The comparison of (essentially) two different parameter estimation techniques is,

in my view, a distraction to the core aims of the paper. It leads to extensive explanations (of the two parameter estimation techniques), which do not contribute to an 'overview of the available probabilistic verification measures' or a better understanding of the challenges that arise in applying them to hydrologic variables. Also, it inevitably requires careful evaluation of the relative performance of these parameter estimation techniques later on, including explanation of the significant differences in performance identified, which is missing from the results and discussion. Indeed, the results and discussion sections (subsections of Section 3 and Section 4) are all very descriptive, with no explanation of the differences seen. This makes it very difficult to follow and to appreciate the value of the metrics for identifying specific problems with the chosen methods of uncertainty estimation. The use of an adapted version of GLUE only exacerbates this problem and constitutes a further distraction. Instead, the authors should consider a simpler experimental framework, such as forecasts from a single hydrologic model across several locations (using one parameter estimation scheme) or, if they want to provide a comparative evaluation, a set of forecasts before and after bias correction. The latter might tie in nicely to an updated introduction since, as stated before, there is a close connection between verification (bias identification) and statistical post-processing (bias correction). If possible, the case study should illustrate some of the challenges associated with "borrowing" measure from the atmospheric sciences for use in hydrology (once these challenges are identified).

- The choice of metrics and discussion of associated attributes of forecast quality is debatable. I understand that this criticism can always be raised, given that there is finite space and many metrics available. However, the key to proper use of probabilistic forecast verification in hydrology is the selection of metrics that are appropriate to the problem at hand. Thus, it would help to have some guidance on how the choice of metrics relates to type of application or at least a more

careful justification of the choice of metrics for the chosen case study. While skill scores are not used (Page 3094, line 4), relative measures of forecast quality, to which skill scores belong, are arguably some of the most intuitive and useful measures, since they allow for straightforward comparisons (e.g. between locations or between parameter estimation techniques in this case). They are also useful when the original units of error would otherwise make such comparisons difficult (e.g. mean error of the ensemble mean, continuous ranked probability score etc.). By way of another example: there is no mention of Type-II conditional bias. In general, this is just as important as reliability (Type-I conditional bias). For example, an operational forecaster would like to know if their forecasts systematically underestimate observed flood flows. My criticism is not that some metrics/attributes have been omitted but how this decision process was informed. I think many readers would benefit from this.

- I would recommend that the discussion of measures for distribution properties is reduced or dropped completely (Page 3094: 2.4.1). It is normal practice to conduct data exploration, and there are many other useful metrics for data exploration not mentioned here (scatter plots, quantile-quantile plots etc.). The degree and types of data exploration that might be useful are also problem dependent. This is not the main focus of the paper and one could convey the importance of conducting some data exploration more concisely (without providing measures and detailed discussion). Also note that some of this discussion can take place in the context of verification metrics, such as score decompositions, which convey the relative contributions of systematic bias (unconditional bias and Type-I and Type-II conditional bias), as well as uncertainty (of the observed variable) and sharpness (of the forecast variable).

- The mathematical notation is poor in many places and many equations contain errors or lack clarity. There is no single problem to mention here, but there are many minor mistakes and use of irregular or incorrect notation. Terms are also

used incorrectly throughout, including mathematical terms (e.g. event when referring to an outcome and likelihood instead of probability. Note that likelihood is used in the context of the parameters of a statistical model, otherwise probability is the correct term) and verification terms (e.g. 3105 line 20 "reliability diagrams allow evaluation of skill"). These are further identified under the technical corrections, below.

- Page 3100, Section 2.4.6. The discussion of sample size is unclear to me. It seems to imply that confidence intervals were computed for the verification metrics. If so, how? Or were "indicative" confidence intervals somehow computed from the sample size information alone? If so, this is problematic, as the width of a confidence interval depends strongly on the choice of metric. One approach to computing confidence intervals for verification metrics in the presence of space-time dependence is to use a block bootstrap.

Technical corrections:

- A different notation should be used for timestep (N) and ensemble member size (n), using a consistent case (upper case is normally reserved for random variables).
- Eqn. (1). You should omit the first summation in the denominator.
- Eqn. (3). You should omit the first summation in the denominator.
- Some of the mathematical notation is a little irregular and there are frequent mistakes. For example, eqn. (7) is wrong and uses poor notation. Consistent notation should be used to denote a sample mean (e.g. of the range in eqn. (8)). Why is the division by N used in eqn. (9), but a multiplication by 1/N used elsewhere?

- Use conventional indicator notation for eqns. (9) and (10).
- Page 3096, line 16: 10th quantile? I think you mean 10th percentile.
- Note the relationship between the CR defined in eqn. (9) and (10) and the rank histogram (or probability integral transform for probability distributions), especially in the subsequent discussion, where it is mentioned that the "CR....does not consider the distribution of ensembles." Also see Brown et al (2010) in the reference list, where several intervals are defined with respect to the forecast median and the average frequency of observations falling within the intervals are computed. Essentially, the limitation of the CR, as identified, stems from the use of one interval.
- Reference to the minimum and maximum quantiles is made throughout the paper, but it is not clear what precisely is meant by these quantities. For example, in the context of eqn (9) and (10), it would be better to refer to the lowest and highest ensemble members. In general, one uses a plotting position formula to estimate quantiles from data, and the extreme upper and lower limits are undefined.
- Page 3098, line 1: probability, not "likelihood."
- Page 3098, line 7. The conditional distribution is not referred to as "reliability." Reliability is a measure of departure between the estimated conditional probability given the truth and the truth. Indeed, measures of reliability can take several forms (such as a squared deviation). The same applies to discrimination (line 14).
- Page 3098, line 16: I don't understand the notation here. Also, note that an event is a set of outcomes, or a subset of the sample space. The conditioning must take place for a specific experimental value or outcome.

- Eqn. 13 is wrong. You cannot condition on an experimental value (probability), you need to define the variable and its experimental value separately.
- Eqn. 16 is wrong. In your notation, you have subtracted an "event" from an "observation."
- Line 3105, line 20. The reliability diagram provides a measure of Type-I conditional bias, not skill.
- Page 3106, line 10. It is misleading to talk about (statistical) calibration here when a large part of this paper is concerned with evaluating techniques for hydrologic model calibration. You need to define statistical calibration in this context (i.e. reliability).
- Page 3109, line 17 What exactly is meant by: "The CR does not provide information about biases in the ensembles." The CR is indeed sensitive to bias, although there is no separate identification of the bias and spread contributions.
- Page 3112, line 18. Utility is mentioned here, but it is not used elsewhere. Indeed, it would be helpful to distinguish between measures of accuracy and utility in the introduction.
- Page 3112, line 20: what is meant by "commensurate with the dimension of the ensembles themselves"?
- Page 3112, line 8. Hersbach is misspelled.
- Page 3113, line 1: "theses measure" should be these measures

---

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 8, 3085, 2011.