Hydrol. Earth Syst. Sci. Discuss., 8, C4510–C4523, 2011
www.hydrol-earth-syst-sci-discuss.net/8/C4510/2011/


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

**K. J. Franz and T. S. Hogue**

kfranz@iastate.edu

Received and published: 19 October 2011

We would like to thank Dr. Brown for the substantial effort he put into this review. His detailed comments and suggestions have helped us improve this paper significantly.

Reviewer Comment: 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 uncer-



tainties, 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:

âĂć 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.

REPLY: In response to the first point: We have added a brief discussion of the history

of verification to the introduction. We have also specified challenges for the hydrologic modeling community based on our experiences from this study. These challenges and issues are discussed in the results section and summarized in the conclusions.

In response to the second point, we note that we are basically utilizing a source-based approach and using metrics to evaluate parameter uncertainty on model performance. We believe there is extensive literature on uncertainty methods and techniques. Given the goal of our paper is to evaluate model performance and our focus on historical simulations rather than forecasts, we have chosen not to include an extensive literature review of uncertainty methods and techniques. We feel this may distract from the paper and the goals of the research, which we have tried to improve upon in the revised manuscript.

The following is the revised introduction:

In the classic definition, forecast verification is the process of assessing the skill of a forecast or set of forecasts (Murphy and Winkler, 1987; Jolliffe and Stephenson, 2003; Wilks, 2006). Verification methods have been well developed in the atmospheric sciences (Jolliffe and Stephenson, 2003; Wilks, 2006) and their application to hydrologic forecasts has been progressing in recent years, particularly for probabilistic verification (Franz et al., 2003; Bradley et al., 2004; Verbunt et al., 2006; Laio and Tamea, 2007; Bartholmes et al., 2009; Renner et al., 2009; Brown et al., 2010; Demargne et al., 2010; Randrianasolo et al., 2010). One of the earliest attempts at verification was published by Finley (1884) who undertook an evaluation of the success of tornadoes forecasts. His early (and controversial) work sparked interest and a range of alternative methods in probabilistic verification, many of which are in use today (Murphy, 1997). Notable early verification papers in atmospheric and meteorological sciences have since included Cooke (1906) who undertook one of the first extensive verification studies, Ramsey (1926) and de Finetti (1937) who undertook early work in subjective probability theory, Murphy (1966) who overviewed probabilistic predictions and decision making, and Murphy and Epstein (1967) where the authors provided an overview



of early development in probabilistic predictions and summarized terminology and definitions in the field. More recent work on probabilistic verification measures includes Wilks (1997; 1998), numerous papers by Murphy (1991; 1995; 1996; 1997) as well as papers by Murphy and colleagues (e.g. Murphy and Winkler, 1987; Murphy and Wilks, 1998). All methods of verification, from early work by Finley (1884) to recent work by Bradley and Schwartz, (2011), involve the comparison of a forecast (or set of forecasts) to the corresponding observation (Wilks, 2006). Murphy and Epstein (1967) lay out simple goals for forecast verification, including: evaluating the value of predictions, evaluating the skill of predictions, performing quality control on the forecast, and finally, investigating the cause(s) of prediction errors.

Model evaluation is not dissimilar from forecast verification, except that the approach is generally aimed at evaluating the reproduction of historical events rather than the prediction of future events. However, the goals of forecast verification and model evaluation (i.e. verification) are analogous. Hydrologists are interested in the value and skill of their simulations, as well as the potential sources of error in their modeling system (Muleta and Nicklow, 2005; Beven, 2006; Gupta et al., 2006; Clark and Kavetski, 2010; Kavetski and Clark, 2010; Schoups et al., 2010. Despite the solid existence of probabilistic verification measures in the atmospheric and meteorological sciences, few metrics have been normally applied by the hydrologic community. Historically, evaluation of hydrologic models ensembles has been undertaken with standard deterministic measures, such as error, correlation, or bias, typically applied to the ensemble mean or median and occasionally application of a containing ratio metric (Xiong and O'Connor, 2008). While creating a deterministic variable simplifies the corresponding model evaluation, deterministic evaluation measures are deficient for fully analyzing probabilistic forecast or model performance (Franz et al., 2003; Bradley et al., 2004; Demargne et al., 2010). The recent growth of probabilistic streamflow estimates in hydrologic modeling, including ensemble data assimilation methods (Kitanidis and Bras, 1980a, 1980b; Evensen, 1994; Margulis et al., 2002; Seo et al., 2003, 2009), multi-modeling platforms (Ajami et al., 2007; Duan et al., 2007; Vrugt and Robinson, 2007; Franz et

al., 2010), Extended Streamflow Prediction (ESP) and other probabilistic forecasting systems (Day, 1985; Krzysztofowicz, 2001; Faber and Stedinger, 2001; Franz et al., 2003; Bradley et al., 2004; Franz et al., 2008; Thirel et al., 2008) and post-processing techniques (Krzysztofowicz and Kelly, 2000; Montanari and Brath, 2004; Coccia and Todini, 2010; Weerts et al., 2011) warrants greater integration of probabilistic model evaluation into the hydrologic community.

There have been few publications on the probabilistic assessment of model performance. Duan et al., (2007) used the ranked probability score to evaluate the outcome of a multi-modeling system. De Lannoy et al. (2006) evaluated model uncertainty for soil moisture using the rank histogram (or Talagrand diagram) and several moments from the probability density functions (such as ensemble spread). Franz et al. (2008) applied probabilistic verification methods to ESP hindcasts produced using two different snow models to assess the impact of the model structure on streamflow predictions. Finally, Shrestha et al. (2009) used the range of the probability interval and number of observations that fell within the interval to assess estimates of model parameter uncertainty in a lumped conceptual model.

The focus of the current study is to provide a succinct overview of a range of available probabilistic verification measures and to demonstrate their application in evaluating and distinguishing model ensemble performance. We utilize two commonly applied parameter estimation methods (Generalized Uncertainty Likelihood Estimator (GLUE; Beven and Binley, 1992) and the Shuffled Complex Evolution Metropolis (SCEM; Vrugt et al., 2003) and an operational rainfall-runoff model (Sacramento Soil Moisture Accounting Model (SAC-SMA; Burnash et al., 1973) for demonstration purposes. We evaluate the uncertainty associated with model ensembles propagated through parameter estimates, although the metrics presented here are readily transferable to evaluate model performance from other probabilistic systems. We are not undertaking explicit evaluation of the "best" parameter estimation method being used, but rather highlighting how the applied metrics can help better inform users on model performance and



behavior when different results (ensemble hydrographs) are apparent. We also highlight unique challenges in applying probabilistic verification to hydrologic model ensembles and provide initial guidance on those measures which may be most suitable to the hydrologic community. The study sites, model, parameter estimation methods and verification metrics are presented in Section 2.0. Results from the application of the verification metrics are discussed in Section 3.0. Concluding statements are provided in Section 4.0.

The following is the revised conclusions:

When evaluating ensembles of simulations, deterministic metrics are often applied to the median or expected value. This practice ultimately removes a significant amount of ensemble information from the evaluation process. We have demonstrated a sampling of metrics that are traditionally applied for verification of forecasts, and have shown these to be informative for evaluation and comparison of ensemble streamflow simulations. A considerable amount of information about the uncertainty estimation methods can be obtained when treating the simulations in a probabilistic manner. A critical skill of a probabilistic simulation is the ability to indicate which flow is most likely, rather than just merely capture the event using large uncertainty bounds. A simulation ensemble can be considered accurate if it contains all the observations within the uncertainty bounds; however if the uncertainty bounds are so large that there is little precision in the ensemble, the ensemble is useless for any meaningful decision-making application. Discrimination and reliability diagrams give information about the accuracy and precision of the uncertainty estimates. The use of flow categories and the joint distribution plots allow analysis of the ensembles for discharge levels of interest.

We have identified some challenges when using forecast verification metrics for model ensemble evaluation. First, most forecast verification metrics were developed for forecasts of a single variable (e.g. rain or no rain, or peak discharge) to occur over some forecast interval, whereas model simulations produce a continuous variable most often evaluated at the model timestep. This means that in the case of evaluating model simu-

lations, the sample size will likely be very large. Furthermore, the number of timesteps with low flow will be very large relative to the higher flows and model skill for low flows will dominate the results. Because low flows are often the range of least interest, approaches to limit the influence of low discharge events on the statistics should be investigated. One possible approach to deal with variations in sample sizes across flow regimes is to evaluate categories of flows as shown. But careful consideration of the influence of the sample size and sampling distribution on the confidence of the verification metric, an issue not addressed in this study, should be taken (Bradley et al., 2003; Wilks, 2006). Because probabilistic statistics rely on significant number of model-observation pairs to obtain meaningful results (Wilks, 2006), evaluation of the model uncertainty associated with flood events will be limited by small sample sizes in most cases. Common problems such as identifying flow and probability thresholds or appropriate distributions exist and, because they may be treated differently in different studies, will limit the ability to compare results across different studies. Finally, we did not test for time-dependent clustering of the ensemble members or independence of the events analyzed, such as described by Christoffersen (1998), to determine statistical correctness. There is significant memory in a sequence of hydrologic model outputs and hydrologic observations, which violates the assumption of sample independence. Investigation of this issue with respect to hydrologic model and forecast verification is a recommended topic for future studies.

Nonetheless, advanced probabilistic verification metrics developed for forecast verification provide a rigorous platform by which modeling methods can be evaluated and compared. The application of these metrics require no information in addition to what is already available as part of the traditional model validation methodology, except that it considers the entire ensemble or uncertainty range in the approach. These measures are much more informative about the nature of model uncertainty estimates than simple deterministic measures. Through our efforts in this and future papers, we hope to advance discussion about evaluation of simulation uncertainty and more robust model verification measures.



Reviewer Comment: âĂć 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).

REPLY:We recognize that the objectives and message of the paper were not made sufficiently clear in the first submission. The original paper was overwhelmed by comparisons between the three parameter uncertainty methods and the results and discussion did not support our stated objectives for this paper. We have removed the modified GLUE method from the paper. We retained the SCEM and GLUE in order to show how the evaluation results vary between two very different ensembles. We have integrated the results and discussion sections to remove redundancies. We have also rewritten

the results such that comparative statements about the GLUE and SCEM are made only to discuss application of the evaluation measures. We have also tried to highlight problems with the application of the selected measures to the model ensembles as mentioned above.

While we find the suggestion to demonstrate bias correction to be valid and would make an interesting paper, we do not think it would support the purpose of this manuscript which is to broach the subject of the currently inadequate approach to model uncertainty evaluation.

Reviewer Comment: âĂć 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).

REPLY:We have reduced the discussion of the distribution properties and, in general, have edited the methods section to remove unnecessary material about the statistics.

Reviewer Comment: âĂć 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.

REPLY: We have removed the term likelihood within the results section. We have also fixed notation where necessary and addressed other specific reviewer comments that follow.

Reviewer Comment: âĂć 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 spacetime dependence is to use a block bootstrap.

REPLY:We have removed this section as it was confusing and detracted from our analysis.

Reviewer Comment: âĂć 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).

REPLY:We have removed the use of symbol N. Ensemble size is z and timestep is n in the revised manuscript.

Reviewer Comment: âĂć Eqn. (1). You should omit the first summation in the denominator. âĂć Eqn. (3). You should omit the first summation in the denominator.

REPLY: Thank you for pointing out these errors. They were incorrectly rewritten during typesetting, we will review the PDFs more carefully in the future before they are sent to reviewers. See equations 5 and 7 in the attached PDF.

Reviewer Comment: âĂć 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?

REPLY: Equation 7 (equation 2 in the revised version) and inconsistencies in the notation have been fixed.

Reviewer Comment: âĂć Use conventional indicator notation for eqns. (9) and (10).

REPLY:Fixed. See equation 8 and 9 in the attached PDF.

Reviewer Comment: âĂć Page 3096, line 16: 10th quantile? I think you mean 10th percentile.

REPLY:This has been fixed.

Reviewer Comment: âĂć 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.

REPLY:We agree with the reviewer's assessment that the CR uses the entire ensemble bounds and does not consider the specific distribution of the ensembles , however we wanted to demonstrate the standard application of the metric and include it for that purpose.

We added a statement in the results section with respect to methods to investigate the ensemble distributions: In its standard application, the CR provides a useful summary of the accuracy of the uncertainty bounds, but does not consider the distribution of the ensemble members. It also cannot reveal whether the ensemble is over- or under-



estimating the observation. More detailed information about the ensemble member distributions and associated performance can be obtained by considering multiple intervals within the ensemble, rather than the ensemble bounds only such as through the application of the rank histogram (Hamill and Collucci, 1997; Hamill, 2001; Wilks, 2006) or spread-bias diagram (Brown et al., 2010).

Reviewer Comment: âĂć 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.

REPLY: The text has been modified to discuss CR, and other measures, with respect to ensemble bounds.

Reviewer Comment: âĂć Page 3098, line 1: probability, not "likelihood."

REPLY: This has been fixed.

Reviewer Comment: âĂć 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).

REPLY:Section 2.4.4 has been revised. The section now first dicusses the conditional distributions as calibration and likelihood distribution and then states that they are displayed on reliability and discrimination diagrams, respectively. We then state specifically how we use the terms reliability and discrimination to discuss the results from these figures. See the revised Section 2.4.4 in the attached PDF.

Reviewer Comment: âĂć 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.

REPLY: This section has been rewritten and the notation changed. The term flow category is used instead of “event”.

Reviewer Comment: âĂć Eqn. 13 is wrong. You cannot condition on an experimental value (probability), you need to define the variable and its experimental value separately.

REPLY: We presented this equation in a manner similar to the way in which it is presented in the literature cited (i.e. Wilks, 2006). We have, however, altered this discussion to present the concept by way of an example rather than the equation used previously given this comment by the reviewer.

Reviewer Comment: âĂć Eqn. 16 is wrong. In your notation, you have subtracted an "event" from an "observation."

REPLY: We have corrected wording and the notation for the BS equation (now equation 15).

Reviewer Comment: âĂć Line 3105, line 20. The reliability diagram provides a measure of Type-I conditional bias, not skill.

REPLY: we have removed the term “skill” with respect to reliability diagram results.

Reviewer Comment: âĂć 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).

REPLY: We have removed the term “calibration” from the reliability discussion in the results. We also make the following statement in the methods section: “To avoid confusion with the model parameter calibration discussion, hereafter we refer to the calibration of the ensemble probability as reliability.”



Reviewer Comment: âĂć 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.

REPLY: We were referring to the fact that CR does not indicate whether the ensemble is over- or under-estimating the observation, and have clarified that in the text. The original statement has been removed.

Reviewer Comment: âĂć 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.

REPLY: We removed this term in the conclusions because we did not evaluate the usefulness of the ensembles for application such as forecasting.

Reviewer Comment: âĂć Page 3112, line 20: what is meant by "commensurate with the dimension of the ensembles themselves"? REPLY: This sentence has been removed in the revised conclusions.

Reviewer Comment: âĂć Page 3112, line 8. Hersbach is misspelled. REPLY: We removed the discussion about the CRPS and this reference.

Reviewer Comment: âĂć Page 3113, line 1: "theses measure" should be these measures REPLY: This has been corrected.

Please also note the supplement to this comment:
http://www.hydrol-earth-syst-sci-discuss.net/8/C4510/2011/hessd-8-C4510-2011-supplement.pdf

---

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