Hydrol. Earth Syst. Sci. Discuss., 7, C2243–C2250, 2010
www.hydrol-earth-syst-sci-discuss.net/7/C2243/2010/

Hydrology and Earth System Sciences Discussions

Interactive comment on "Why hydrological forecasts should be evaluated using information theory" *by* S. V. Weijs et al.

Anonymous Referee #1

Received and published: 10 September 2010

General comments:

The paper discusses some topics on the application of information theory to forecast evaluation in hydrology. Based on a statistical measure of forecast quality developed and presented in detail by the authors in a previous paper, the Kullback-Leibler divergence score (Weijs et al., 2010; Mon. Weather Rev.), the authors develop their arguments on how the information-theory point of view can throw light on some current topics discussed in hydrological forecasting, namely, the transition from deterministic to probabilistic forecasts and the calibration of forecasting systems. The topic is of great interest to the hydrometeorological community, particularly because of the increasing use of probabilistic forecasts in operational hydrology and the common agreement that

there is a need of developing evaluation measures more focused on the problems handled by hydrologists.

The paper is well written and, in general, the authors explain clearly their point of view, although the ideas the authors want to convey are not always structured in an easy-to-follow way. It is clear that the paper raises and discusses points that are doubtlessly very important for the hydrological community, and I much appreciated that the authors supply a theoretical framework for their claims, which are defended in a clear, and sometimes very passionate, way. However, their arguments are not illustrated by practical applications, which would enhance the paper and give experimental support to their claims. For instance, to what extent real situations confirm the theoretical arguments of the authors?

Experiments may not always hold the truth, but testing models or illustrating arguments against real data sets available is a natural step in hydrological sciences. Thus, my main concern is that much of the paper's arguments lays at the very theoretical point of view, while forecasting is, in great part, supported by practical experience (for instance, forecasting modellers or experts usually acknowledge that much is learned with the practice of real-time forecasting). In my opinion, this point does not shadow at all the importance of the paper, but it raises the question whether the paper is not more suitable to appear under the umbrella of "HESS Opinions" and, consequently, be in a larger (and maybe longer lasting) discussion forum.

In this context, some specific comments are listed below. They mainly point out some suggestions to improve the structure of presentation of ideas, by focusing more directly on the presentation of the authors' arguments and of the links to real situations encountered in operational forecasting.

Specific comments:

In the abstract, the authors state that "We propose a Kullback-Leibler divergence as the appropriate measure for forecast quality". It seems however that their previously

published paper (Weijs et al., 2010; Mon. Weather Rev.) already presents this score and its decomposition in details. By the way, Table 1 reproduces Fig. 2, 3 and 4 of this previous paper, and Figure 1 is very similar to Fig. 1 of the already published paper. The novel (and central) aspect of the paper proposed at HESSD seems thus to be the interpretation made under the shed of the divergence score (concerning deterministic forecasts) and the implications to the calibration process in hydrological forecasting. This should be emphasised in the abstract and linked to what is stated in the outline (page 4662, lines 18-21). The objective of the paper should, in this sense, be more clearly stated. I also suggest that some terminology should be clearly defined in the beginning and kept all over the paper to help the reader in better following the reasoning behind the ideas conveyed (for instance, in the abstract, it is nicely stated that "In this paper we distinguish two scales for evaluation: information-uncertainty and utility-risk.", but the words "scales for evaluation" and "utility-risk", for instance, are not used anymore in the next sections).

Thus, in general, the paper needs some rewording (and re-structuring) in order to make it more focused to the message the authors want to convey. In this sense, Section 2 could be much simplified, especially because it is already presented in the paper Weijs et al., 2010 (Mon. Weather Rev). Only those points that are necessary to the understanding of Sections 3 and 4 should be kept. For instance, the DS could be introduced more straightforwardly; one does not need to know that the measure of uncertainty was derived by Shannon (1948) from three basic requirements (Page 4665, lines 10-11), if the requirements are not explicitly mentioned and are not useful to understand the text that follows, etc.

The introduction could be enriched by mentioning the main initiatives in the hydrological community to develop probabilistic or ensemble flood forecasting (eg. HEPEX, EFAS, MAP-DPHASE) and by commenting on some published papers on the evaluation of hydrological forecasts. This would allow the authors to link their analyses/arguments to the main challenges and/or questions raised by these initiatives and scientific papers,

and, consequently, provide more solid basis for some of the authors' statements, like the ones in Section 1.2 and the one in the beginning of the Conclusions section (Page 4678, lines 1-2): "The difficulties and debate about the evaluation of forecasts can be significantly clarified using results from information theory." Since the authors are not presenting a practical example (with real data set of hydrological forecasts), how can it effectively be true?

In Section 1.2, can you give an example in the literature that illustrates the statement in lines 10-12? Also, the sentence in lines 12-15 needs clarification. In fact, I think that several recent studies in hydrological forecasting do consider the separation into quality and value. It is maybe true that they usually focus more often on the evaluation of forecast quality (maybe because quantifying value in flood forecasting is not straightforward, especially when the aim of a forecasting system is basically the protection of human lives). Besides, I think that the purposes of an evaluation framework can be quality and value, even if these should be expressed by different measures or scores. A forecasting system and its forecasts can be pictured by something more than just a number given by a chosen statistical measure.

In the interpretation of the divergence score (links to the Brier score and decomposition; Sections 2.3 and 2.4), can you add a paragraph on how the score can handle extreme (very rare) events (when probability tends to zero), regarding the terms of the divergence score that are not bounded? (As these are those events that most interest operational hydrologists in flood forecasting).

In Section 3.1, I think that the role of human expertise in the forecasting chain is unclear and sometimes underestimated. Maybe this is because some definitions are missing. For instance, a "forecast" is here understood as a raw model output or as an output that was submitted to human expertise/post-processing? Is the "user", as considered in the paper, a hydrologist that receives a forecast from the model, a stakeholder/engineer with (or without) technical knowledge, the public, etc.?

In practice, forecasts are usually issued with something more than just a number (the case of deterministic forecasts) displayed, and sometimes with an interval associated with some quantitative (or qualitative) probabilistic information. Thus, I think that the representation of a user having to guess what the forecaster wanted to say (Page 4669, lines 26-27) is probably a bit exaggerated (although surely not completely unrealistic). Also, I do not think that reservoir operators will be "infinitely surprised", as stated on page 4672, line 18, if they have to face 210 m3/s when it was forecasted 200 m3/s. They can be angry because they took it as at "face-value", if they do, but not surprised, especially if this is huge amount of flow to their catchment (i.e., a rare event). However, they will be really "infinitely surprised" if for the same situation the forecast indicated clear sky and not a single amount of rain! In my opinion, this (and some other statements; see below) need to be moderated in a revised version by some more realistic considerations of the practice of hydrological forecasting at operational conditions. Or, otherwise, at least, they should be illustrated by concrete examples of past events (case-studies) that support the strength of such statements. Also, the fact that the same sentence pointed out above is a bit rewritten by the authors later on in the text (page 4673, lines 1-4) shows that some re-structuring of the text could benefit the better reading and understanding of the authors' arguments.

In the same context, the role of persistence and real-time discharge assimilation in streamflow forecasting models is also, in my opinion, not clearly assigned. A user (or a forecaster) usually uses this information to forecast or make a decision. In my opinion, the statement on page 4670, lines 23-26, is another example of a statement that needs to be moderated. In the forecasting process of a forecasting system there is not only one actor behind the steps that leads to the production and issue (communication) of a forecast or an alert. So, information is added gradually. But I agree that which part of information/uncertainty should be considered by what (the model, the input data) or by whom (the forecaster, the user – but which user?), etc. is probably not yet clear in several operational forecasting centres.

Also in general terms, the term of "calibration" needs to be more clearly defined. This is a very interesting and important part of the authors' arguments and needs some revising. In hydrological simulation, calibration is clearly understood (parameter calibration of a model using a given objective function). In forecasting, however, sometimes calibration refers to "calibrated forecasts" (in opposition to "raw forecasts" directly taken from model output), meaning usually that some post-processing was performed to adjust forecast probabilities. This is usually done from the help of archives of forecast data and the corresponding observed time series (the availability of such an archive is another matter, largely debated within the community, eg., HEPEX). Besides this post-processing, hydrologists also have to handle the calibration of the hydrological model, which is a component of the forecasting system. A forecasting model is, in several forecasting systems, different from a simulation model, as it uses updating procedures (changes in parameters, states or outputs according to the last observed discharges) to better predict future states. The aim is not reproducing the "average behaviour" or "low and high flows" in a continuous long-term modelling framework anymore (as it usually is in the simulation exercise), but to issue the best streamflow prediction to the next hours or days. How does it affect the calibration issues raised by the authors? Would it be recommended to perform calibration in a "forecasting mode"? What is the impact of real-time data assimilation in this case? Do you think the framework for calibration proposed is compatible with operational constraints (what if the system is multi-purposes or the "utility" changes according to the situation that is being forecasted – e.g., the same system is used for floods and low flows forecasting, or flood security and reservoir inflow, etc.?)

In the statement "It is therefore a violation of scientific logic if the score that is intended to evaluate the quality of forecasts depends on what is stated about things that are not observed" (Page 4675, lines 11-13), what about the fact that forecast evaluation is also interested in the part of false alarms, i.e., forecasted, but "non-observed" events. In this sense, "probabilities assigned to non-observed events" are also important. How do we handle this?

Figures 2 and 3 should be explained in more details in the text. In Fig. 3, for instance, it is said that it shows "three routes" of information, but they are not explained in an organized way. Figure caption is very confusing and, in the text of the paper, only the "third route of information" appears clearly (page 4676, line 23). I think that this figure deserves more attention and links to the text.

Page 4676, lines 10: "hydrological models that are trained on this kind of utility functions. . ." Please, make it more clear or give references/examples.

Page 4676, lines 17-18: what do you mean by "Training for optimal classification of flood events"? Please, explain.

Finally, a lot of probabilistic scores for forecast evaluation are used in ensemble forecasting. How does the size of the ensemble system (number of ensemble members) affect the divergence score?

Technical corrections:

- Page 4659, line 2: change to ". . . the lack of methods for the evaluation of. . ." - Page 4666, line 20: change to ". . . the shapes of the resolution components is visible." - Page 4668, lines 6-7: change to ". . . information, which can be subtracted from the climatological uncertainty, or the missing information." - Pages 4668/4669, lines 28/1-: change to "If the forecast is right, the perfect score of 0 will be attained. If the forecast is wrong, however, a penalty of infinity will be give." - Page 4671, lines 7-9: the sentence is confusing. You mean ", that is. . ." in the sense of "in other words" or you wanted to say ". . ., which is. . ."? (some other sentences in the text lack the same clarification) - Page 4671, line 11: change to "A penalty (objective) function. . ." - Page 4671, line 14: change pdf for "probability distribution function" - Page 4671, line 16: change to ". . .. is actually to evaluate the probabilistic part of the model." - Page 4671, lines 21-24: confusing. Please rephrase it. - Page 4673, line 11: wife/husband (?) Maybe gender selection should be avoided in such sentences. - Page 4673, line 27: ". . . has potentially more utility or value for the user".

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 7, 4657, 2010.