Hydrol. Earth Syst. Sci. Discuss., 7, C2830–C2852, 2010
www.hydrol-earth-syst-sci-discuss.net/7/C2830/2010/

Hydrology and
Earth System
Sciences
Discussions

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

S. V. Weijs et al.

s.v.weijs@tudelft.nl

Received and published: 11 October 2010

Responses to anonymous reviewer 1 for "why hydrological forecasts should be evaluated using information theory"

11 October 2010

We thank the anonymous reviewer for the detailed and constructive comments. The view from a more practical side that the reviewer gives is very helpful to try and make the paper more appealing to practice and to convey the theoretical message. We have tried to address the comments as much as possible to make some things more clear, giving some extra illustrations. We also are working on a practical example to illustrate the points made in section 4. To improve readability, we included the reviewer's original comments as quoted (indented) text.

On illustration with practical examples

> 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 modelers 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.

We agree with the anonymous reviewer that examples would help to support our claims and confirm our arguments. However, indeed the main focus and contribution of this paper lies in the theoretical discussions and illustration of theory with a paradox. This argumentation is not necessarily made clearer by adding an (necessarily) elaborate practical case study, which also distracts attention from the main point, which is offering an information-theoretical perspective. We think, experience from practice and deduction from theory are two complementary routes to arrive at insights and this paper focuses purely on the first. We think that illustration with an example can have two objectives:

1. To test the correctness of the theoretical results

2. To demonstrate the practical significance of the results

The first objective is difficult to achieve in this case due to random effects and would require an elaborate statistical analysis and a careful experimental setup with controlled experiments using artificially generated data sets. To the extent that our results rely on established theory and logical reasoning (both open to scrutiny), we think that such an analysis is not necessary in this paper. To do that correctly, a whole new experimental design is necessary that would best be treated in a separate paper. We added a section with some ideas for future work on this.

The second objective could be achieved by adding examples, but it would require many case-studies to see in which cases the effects described in section 4 are most significant, i.e. which utility functions filter the information in the most destructive way. To perform and describe these experiments in a meaningful way, we think it is best to devote a separate paper to it, after some careful thought has been devoted to design good experiments. We made the theoretical focus of this paper more clear in the introduction and added some ideas for future practical tests in a "future work" section.

With regard to what might be demonstrated with examples, we have the following ideas. In Weijs et al. (2010), a practical example of using the decomposition is already presented. A new application of the divergence score on hydrological ensemble forecasts would be interesting, but would not serve as support for our arguments in section 3 and 4, which are the key message of this paper. The paradox and two possible resolutions treated in section 3 are mainly about interpretation of deterministic forecasts and appealing to logic. Following the suggestions of the reviewers, we elaborated some of the illustration examples a bit to make them clearer.

The theoretical results regarding inference of explicitly probabilistic models based on information measures, as given in section 4, could probably be shown in an experiment. However, this would require a detailed and elaborate study involving different models, artificial and real data sets, truly independent calibration and validation periods and a detailed analysis of the results. In general the comparison of skill scores, inference methods and uncertainty estimation methods is difficult and plagued randomness and unexpected features in the data, which makes it difficult to draw general conclusions from experiment. In some cases the difference may be more significant than in others, but given the theoretical and philosophical considerations, there is no reason to not prefer a theoretically more justified method, given that it is not more difficult to apply than existing methods. This is the case even if in some situations the results may not significantly differ from the alternative methods or other effects disturb the results. The present paper tries to bring an information-theoretical view on a number of points by logical reasoning on the basis of some recent results Weijs et al. (2010); Benedetti (2010) and some well-established older results (data processing inequality, likelihood principle Cover and Thomas (2006); Bernardo (1979); Berger and Wolpert (1988)). Practical examples would indeed be very valuable to support the reasoning and clarify the practical implications of the point made here, but we think it would be complicated and too much to treat in one paper. We certainly plan to devote some future research on this and to find a relevant real world case where it can be tested and give some ideas for this in the last section. We also will try to include one example of inference based on a binary utility function as an illustration to this paper.

On the connection to the previous paper

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

> ever that their previously published 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 emphasized 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 utilityrisk.", but the words "scales for evaluation" and "utility-risk", for instance, are not used anymore in the next sections).

Indeed the review correctly points out that the main points of the paper are the interpretations in sections 3 and 4, which build on the decomposition presented in a previous paper, of which the relevant material is summarized in section 2. Another new point is placing forecast verification in the context of the requirement for testable predictions, mentioned in the introduction.

Following the suggestion by the reviewer, we removed some parts from section 2, which were not essential to understand sections 3 and 4. We also emphasized the objective of the paper more, mentioning explicitly the theoretical character of the paper and the relation to the previous paper. Also, we clarified the link between the utility-risk scale and value on the one hand and the information-uncertainty scale and quality on the other.

> 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?

We will also put the paper somewhat more in the context of ensemble flood forecasting initiatives. For example, the present paper relates quite directly to the questions stated in HEPEX (Thielen et al., 2008):

- How can hydrological ensembles be generated that reflect all known uncertainties?
- How can automatic calibration aid in characterizing uncertainty?
- How can systematic over- and under-prediction of rainfall forecasts from both deterministic systems and EPS be detected and corrected for better flood forecasting.

The statement on page 4678, lines 1-2 was not clearly formulated. A better phrasing would perhaps be: "The difficulties and debate about the evaluation of forecasts can be significantly clarified using an information-theoretical viewpoint." This avoids the confusion between results as referring to outcomes of practical experiments and results

in the sense of mathematical theorems of information theory that were proven from some axioms, which we referred to. The word “difficulties” in this statement refers to e.g. trying to evaluate forecasts that are not completely specified or to discussions on users and utility in contexts were quality is the dimension of evaluation (cf. sensitivity to distance).

On quality and value

> 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.

The statement that the distinction between quality and value is not always explicitly made should be seen in the light of the points made in the conclusions of Weijs et al. (2010), stating that quality can be equated with information, and any other measure somehow reflects a utility, i.e. value. For example, the Brier score is often presented as a measure of quality, but at the same time defended on the basis of reflecting value for users with an uniform cost-loss ratio distribution. Another example is the RPS score, which is used as a score for quality, but required to be sensitive to distance. This distance only makes sense in the context of a decision problem (e.g. the disutility of a forecast being 10 m^3/s off). However, if it really would be a pure measure of quality (correspondence between forecasts and observations), then why could the score change depending on how forecast probability is spread over things that might never be observed? This relates to the discussion about locality. As the reviewer points out, the distinction between value and quality is made in literature, but the we argue that most measures that are used to measure quality can only be meaningfully interpreted when they are seen as measures of value in some special cases (see also the conclusions section in Weijs et al. (2010)).

Indeed, forecasts may be pictured in more ways than just one number, but we claim that quality in the sense defined by Murphy should be expressed in terms of information. As argued in section 4, information is the evaluation measure that should be used when making decisions about the model and the parameters (i.e. learning), while value might be a useful evaluation measure if a decision about allocating funds to different flood-forecasting initiatives must be made (i.e. cost-benefit analysis on investing money in flood warning). While value can be defined in many ways, depending on the specific utility functions of the users, quality can only be defined in terms of information, or we must give up either locality or propriety. We will elaborate the text in sect 2. line 10-15 to make it clearer.

Handling of extreme events

> 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).

When events have a very low probability, the forecasting becomes most challenging, first of all because we have little experience and data for them, making modeling difficult, and secondly because these are the events we are not used to cope with and therefore can have severe consequences. The divergence score is unbounded in case the forecast assigns a zero probability to some event which then turns out to occur. This can be avoided by never assigning zero probabilities to anything the is not truly impossible. For example a probability of 0.001 might be assigned to the category between the highest observed flood and infinity. This is of course very crude, but any better idea for representing the tails leads to a higher expected divergence score and is thus encouraged. The unbounded scores are thus not so much a problem of the score, but related to the general difficulty of estimating probabilities of extreme events using little data. Unfortunately the divergence score's sensitivity shows that apart from being difficult, this task is also very important. We will add a paragraph on this somewhere in the paper. Maybe also in another section, because we are contemplating whether or not to keep the link to the Brier score in the paper, as it might not be essential for what follows.

On the user's guesswork

> 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 this paper, a forecast is a testable prediction about the future. This can come directly from a model or be the result of a combination of algorithms and human expertise. The user is supposed to be less informed than the forecaster, and in principle willing to literally copy the forecast as best estimate. In case this is not true and the user actually has relevant extra information, then it would be helpful if the user shared this with the forecaster, so the forecast be improved (i.e. “participatory forecasting”). We added a section on definitions later in this reply and will make sure that the definitions are clear in the revised paper.

> 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).

In the context of the literal interpretation of forecast information that is intended to show the paradox, there is no exaggeration in our statement: "However, it is not the task of a user to guess what the forecaster wanted to say." The statement is valid in the context of all forecasts that do not completely specify a probability distribution, as later elaborated using the principle of maximimum entropy. We argue that if a user has to interpret i.e. mentally recalibrate a forecast, he is actually adding probabilistic information that the forecaster failed to add, i.e. guessing (a vital part of) what the forecaster wanted to say. Indeed, in practice there are two soothing circumstances, which are two resolutions of the paradox mentioned later in sections 3.1 and 3.2, but the point of raising the paradox is to show that these solutions are implicit and undesirable.

On infinitely surprised reservoir operators

> 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 forecast 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.

Within the literal interpretation of the forecast information used to raise the paradox, we do think the forecaster will be "infinitely surprised". In practice this is indeed never the case because of other information available to the user next to the forecast (e.g. the user’s own observations of the catchment and the sky and his life experience of not believing anything people of computers tell him), but this should not influence how we evaluate the forecasts, because they should be optimal in their own right. In other words, the evaluation is meant to evaluate the forecasts and not the combination of the forecasts and the user’s experience. The only reason why a forecaster would not be infinitely surprised is because he adds information to the forecast by not taking it at face value (either based on his own observations or prior information available to him).

For the case of interval forecasts with an associated probability of e.g. 90%, the literal maximum entropy interpretation is piecewise uniform, with 90% of the probability assigned between the two interval bounds, and the remaining 10% spread out over the two outside intervals, bounded by $-\infty$ and ∞. This results in an infinitesimal density and the resulting score for an observation outside the interval will still get infinite penalty.

About the the two formulations for the infinite penalty:

1. "Even though a reservoir operator might be infinitely surprised if he has taken a deterministic inflow forecast of 200m^3/s at face-value and he finds out the inflow was 210m^3/s, his loss is not infinite."
2. “In contrast, for decision problems like reservoir operation, optimally preparing for 200m^3/s automatically implies also preparing for 210m^3/s to some extent. This makes the loss function non-local (locality is discussed in Sect. 4.1).”

The first formulation is intended to show the contrast between 3.2, which is about loss and utility, and the previous paragraph, which is about information and surprise. The second formulation is in the context of utility, loss and decisions. The statement is intended to show that a decision is often not linked to only one outcome, but is near-optimal for several outcomes, in that way defining a sort of closeness between different outcomes. This in contrast to the horse race, where the utility of the decision betting on one horse on depends only on the winning of that horse. We clarified this by putting emphasis on the words surprised and loss in statement 1. We also rephrased the text around the second statement to make its meaning clearer.

The role of persistence and real-time discharge assimilation

> 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.

Persistence in discharge offers the possibility to use the current the previous discharge as one of the predictors in the forecasting system. This leads to better forecasts with lower entropy probability distributions and lower divergence scores if they are well-calibrated. If the verification score is applied to the final forecasts, the complete forecasting system is evaluated, including data assimilation. Indeed this may not say much about the quality of e.g. a rainfall-runoff model used in the forecasting chain and it could very well be that most of the skill comes from persistence. This, however, can easily be checked by comparing the divergence scores of the whole forecasting system, with a simple timeseries forecasting model that only uses past discharge. The difference between the two scores is the number of bits of information per timestep that the rest of the system adds to the probabilistic persistence forecast.

The statement on 4670:23-26- "This brings back the question who ought to specify these constraints, which in fact constitute information. The fact that the user can reduce the maximum entropy by adding this common sense constraint actually means that the forecaster failed to add this information." - is intended in the light of the preceding part of section 3.1. There is no good reason why the forecaster would not communicate an exponential distribution with its parameters instead of giving a mean value and then hoping that the user interprets this limited information by adding his own information and applying the principle of maximum entropy. The point is that the forecaster is the one who is best equipped to make probability estimates, and once this has been done, these should be summarized in a way that fully captures all his available information and the remaining uncertainty, i.e. be consistent with the forecasters internal beliefs.

Definitions used

> 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 postprocessing, hydrologists also have to handle the calibration of the hydrological model, which is a component of the forecasting system.

The following are the definitions as we use them in the paper. We will make sure

forecaster: The person or institution that issues information about some uncertain future event to a user or group of users, either in a public forum or in a specialized communication.

user: The person or institution that receives the information sent by the forecaster. The user may use the information from the forecaster to support a decision, hopefully increasing his expected utility of that decisions.

model: an algorithmic set of equations and procedures which processes information from observations to make predictions. This includes possible post-processing schemes or other standardized procedures.

calibration: The use of past model predictions and corresponding observations to adjust some parameters of the model, aiming for an improvement of the model's future predictions according to some objective function. In the case of post-processing, the parameters describe the transformation of the predictive distribution from the model to the final forecast.

When the post-processing procedure is seen as a part of the total model, which we think it should, the above definition captures both hydrological model calibration and the statistical post processing. The corrections to the forecast distribution can be described by an algorithm with tunable parameters, exactly like the hydrological model.

> 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 modeling 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?

The difference between forecasting and simulation lies in what is used as input data and is not really fundamental. Where the simulation model for example uses only observed rainfall and potential evaporation as inputs, a forecasting model may have additional inputs, like the observed previous discharge. In both cases, calibration is the adjustment of the algorithm that the model represents (i.e. mathematical relations, parameters) based on past pairs of observed inputs and outputs to improve the extraction of information for those inputs in the future.

We think calibration of stochastic models based on information would be valid for both the forecast and the simulation cases and probably even help to overcome some of the difficulties, or at least to understand them better. For example, a forecasting system using past discharge does very well in recession curves of moderate discharge peaks. This recession behaviour is the result of hydrological processes and should be captured by the hydrological model. In an operational forecasting setting, much of the skill could originate statistically assimilating the past discharge. Using RMSE as calibration objective implies a Gaussian homoschedastic forecast distribution, making the hydrological model relatively insensitive to small errors in the recession. In a probabilistic real time setting, however, forecasts of the recession curve would be sharper, making the model more sensitive to those errors. In other words, an error of $10m^3/s$ in the recession curve would cause more surprise than a $10m^3/s$ error just after a rainfall event. The latter thus contains more information for the model.

Somehow, this information should be fed back into the hydrological model. The challenge is to channel the information to the right part of the model. This channeling is entirely determined by the a priori parameters estimates, the a priori model structure and the a priori stochastic part of the model, which is equivalent to the training measure used in a deterministic model calibration. In other words, the observation data is the only source of information that is not put into the model a priori.

The fact that utility changes from situation to situation is exactly the reason why we want to train models based on information in the observations. In that way the model learn as much from the information as possible. By filtering the information through a utility we have less to learn from. Because this also means less information to fit, the fit will be easier to achieve but have less predictive power. The problem is only exacerbated if we train on different utilities for floods and droughts. We learn from one part of the information in one case and from another part in another case, and switch

from one to the other. This introduces a unnatural discontinuity in the model, which is then composed of two submodels and a switching mechanism. Moreover, if the signal that determines in what situation we are is determined by discharge, rainfall or a combination of both, that information was available in the training signal anyway, and nothing new is to be learned. A possible solution where the switch is gradual removes the discontinuity, but the same could be achieved by a heteroscedastic probability distribution, that changes with the circumstances. This would be a more natural reflection of the fact that in some conditions, we can be more certain than in others.

> Would it be recommended to perform calibration in a "forecasting mode"? What is the impact of real-time data assimilation in this case?

This is a complex question. we think the answer should be “yes”, but this would also mean that given enough real time data, the hydrological model becomes quite unimportant and we are in fact training a completely data based model. If we then attach physical meaning to the hydrological model and parameters and want to use it to see the effect of e.g. future land use changes or add some topological information, we might get wrong predictions. Calibrating in forecasting mode requires careful consideration of available data versus model complexity and not using the same information twice. This is definitely an issue that requires further study.

> 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 principle, a model that is trained to use all information and give a best probability estimate based on that will also lead to good decisions for various utility functions. Given that both high and low flows are generated by the same natural system, one single model seems most appropriate. If the models structure is correct and uncertainties are explicitly acknowledged, in theory we can never do worse by combining the important processes for floods and droughts in one model. Only in case the processes for drought and floods are so disconnected from each other that observations from dry situations cannot help flood prediction and vice versa, we could consider two separate models (or two separate training objectives, which is the same as two separate models in our interpretation).

on non-locality as a violation of scientific logic

> 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., forecast, but "non-observed" events. In this sense, "probabilities assigned to non-observed events" are also important. How do we handle this?

Note that for the binary case, false alarms are already captured by scoring on the probability attached to the event that actually occurred. Any forecast probability given to a non-observed event cannot be given to the observed event, and will therefore lower the score. For the binary case (flood - no flood) this is trivial, but for the multi-category case, one might think it matters how probability is spread between non-observed events. For utilities, i.e. the context of decision making, it does indeed matter. This can make the number of false alarms an interesting quantity for decisions whether to invest in such a system (and might also be interesting to study the psychology of users losing trust in the system etc.).

However, for information, i.e. the context of learning from data and science, the way probability is spread over non-observed events should not matter, because that would imply learning from what could have been observed, but was not (see also the paper of Benedetti). We don’t think there could be any scientific knowledge that doesn’t ultimately stem from observations. Note that we use Murphy’s distinction between value and quality, so a measure of quality cannot have anything to do with utilities and thus belongs to the information/learning/science realm.

Also note that by using something in our evaluation that is not information (i.e. likelihood), then it must some kind of prior knowledge (e.g. a weighting of information or a preference for certain shapes of forecast distributions). Since this information is not part of the likelihood it is prior information and could and should have been included in the forecast prior to confronting it with data in the evaluation.

explanation of figures 2 and 3

> 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.

We will put some more elaborate explanations of the figures in the text.

other points

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

Indeed this was not clear, it was meant to refer to the binary decision kind of utility functions. Fundamentally, it is true for all objective functions that are not information, but the effect of filtering information is strongest for the case of a utility function linked to a binary decision problem.

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

Optimal classification of flood events was meant as a forecast whose only objective is to forecast whether or not a certain flood threshold will be exceeded, e.g. a binary output instead of a real number in m^3/s. A model that is trained for such a task thus gets feedback only in terms of “hit”, “false positive” and “false negative” instead of some real number error measure or likelihood value.

how does ensemble size affect the divergence score?

> 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?

In principle, the divergence score evaluates the final probability estimate of the observed event. The score is therefore not directly affected by the ensemble size. It is, however, affected by the definition of the categories in the final forecast that is derived

from it. If the forecasting system tries to forecast an event subdivided in many climatologically equally likely categories, the climatic uncertainty will be higher than when there are less categories. Also the divergence score will be higher (worse), simply because the correct forecasting in which category the event lies becomes increasingly difficult with more categories. A recent paper by Peirolo (2010) also gives some interesting intuitive discussion on this.

Another issue is the sensitivity for overly certain wrong forecasts. Therefore, an ensemble forecast should not assign a zero probability to events outside the range of scenarios. The perfectly reasonable reaction of the divergence score would be to give an infinite penalty if such an event occurs. Assigning a $1/(n+1)$ probability to the two categories outside the extreme scenarios could be a solution, but due to the sensitivity, some gain could be possible by defining a few more categories outside the extreme scenarios to give a realistic representation of a tail.

A a discussion of information-theoretical scores in the practical context of ensemble forecasting is indeed interesting and important for the forecasting community. However, there is so much to say about this that we think it requires some extra research and that it would also make the most important messages of this paper less clear. The focus of this paper is indeed quite theoretical. We think application to a more practical setting would be a good next step, but is outside the scope of the current paper. We welcome some forecasting data to do these tests and are happy to supply the code for calculating the scores and decomposition to anyone interested in doing practical tests with it.

Technical corrections

We followed all suggestions, except for the last one, because when the disutility is defined as MSE, then the series of forecasts that minimizes MSE per definition has the highest utility.

The word “wife” was taken from the example in Kelly (1956), but I replaced it by the gender-neutral “spouse”, since we’re not in the fifties anymore. The rest of the paragraph was also rewritten in gender-neutral terms.

Final remarks

We hope that these answers have clarified some points in the paper. We will try to reflect the main points of this interesting discussion in the final revised paper and refer to this discussion for background. Furthermore, a clear need for further research on several topics came to light. We will include a section in the conclusions on future challenges and possible practical experiments.

References

R. Benedetti. Scoring Rules for Forecast Verification. *Monthly Weather Review*, 138(1):203–211, 2010.

J.O. Berger and R.L. Wolpert. *The likelihood principle*. Institute of Mathematical Statistics, Hayward, CA, 2nd edition, 1988.

J.M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3):686–690, 1979.

T.M. Cover and J.A. Thomas. *Elements of information theory*. Wiley-Interscience, New York, 2006.

R. Peirolo. Information gain as a score for probabilistic forecasts. *Meteorological Applications*, in print, 2010.

J. Thielen, J. Schaake, R. Hartman, and R. Buizza. Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. *Atmospheric Science Letters*, 9(2):29–35, 2008.

S.V. Weijs, R. Van Nooijen, and N. Van de Giesen. Kullback–Leibler divergence as a forecast skill score with classical reliability–resolution–uncertainty-decomposition. *Monthly Weather Review*, 138(9):3387–3399, September 2010.