Hydrol. Earth Syst. Sci. Discuss., 7, C2854–C2862, 2010
www.hydrol-earth-syst-sci-discuss.net/7/C2854/2010/

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 reviewer Federico Lombardo for "why hydrological forecasts should be evaluated using information theory"

11 October 2010

We thank Federico Lombardo for his helpful and interesting comments. We will put some references in the introduction to link it somewhat more to the debate on uncertainty analysis. In the following sections we hope to clarify some points in the paper and raise some more topics of discussion on models and verification. We will include the points most relevant to the present paper in the revised version. Furthermore we hope to continue some exchange on these topics, without the pressure of deadlines. To improve readability, we quoted the original reviewer's comments as indented text.

The difficulty of evidence from application to real test cases

> I think the paper by Weijs et al. (2010a) is relevant to the topics covered by HESS, and the research presented is important and innovative for the hydrological community. However, I believe that more appropriate examples and evidence to support the topic presented in the paper should be

needed, as stated (C2244: 8-10) also in the thoughtful Comment by Anonymous Referee #1 (2010), which I consider almost entirely appended to my Comment.

There are a number of difficulties regarding illustration with a practical example. We refer to the response to reviewer 1 for an elaborate discussion on this. We are working on one application as an illustration of the topics in section 4, but we don't think we can use it as evidence unless a more elaborate study is performed.

On consistency

What does the statement in (4660: 7-10) imply? Do you mean that Consistency is an inherent property of forecasts? This seems to be in contradiction with what is correctly affirmed in (4677: 20-22). In fact, as stated by Brier (1950): "(: : :) one of the greatest arguments raised against forecast verification is that forecasts which may be the "best" according to the accepted system of arbitrary scores may not be the most useful forecasts. (...) This may lead the forecaster to forecast something other than what he thinks will occur". Please, clarify.

4660:7-10 gives the authors' interpretation of Murphy's requirement of consistency in a narrower context. One of Murphy's examples was that a forecast should not give a uniform temperature for a whole area, while the forecaster knows that some spatial variability is always present. Another example is a deterministic forecast while the forecaster internally knows there is uncertainty. In that sense, it might often occur that a forecast is not consistent and this is correctly seen as undesirable by Murphy (1993). Our statement is in the context of the forecast of one single probability distribution for e.g. the streamflow tomorrow. In that case, the only reason for a forecaster to forecast

something else than he thinks is that he is rewarded by a score that can be hedged, i.e. a non-proper score. This is undesirable. Consistency is thus not an inherent, but a desirable property of forecasts. Our statement in 4660:7-10 was meant to imply that a forecaster can be consistent without having the best estimate. He can, however, not consciously let his internal estimate diverge from the best estimate, because that would mean he is irrational while at the same time be rational enough to realise this. However, his external estimate can knowingly diverge from his internal best estimate, if he is inconsistent because he is hedging an improper scoring rule. Inconsistency can thus be equated with dishonesty in this context Murphy's association of goodness with consistency seems to concord with the idea that dishonesty can not improve the goodness of a forecast (except by luck, but not in an expected sense). We added the following to the paper "Consistency is therefore a desirable property, which can be interpreted as honesty, because it is about the match between the internal beliefs and the external forecast."

An interpretation of Brier's statement is that forecasts intended to optimize one utility may not yield optimal decisions according to different utility functions. As we argue in section 4, information as score is an exception to this, because it lets the probability estimate use all available past data. In that sense it is not an "arbitrary" score of quality. The resulting probability estimate for the future is therefore optimal given the available data. Bayesian decision theory says that an optimal probability estimate leads to optimal decisions, whatever the utility function may be.

Any non-information-theoretical measure applies some kind of weighting on the information of the different observations. This either implies that (1) prior knowledge is added, i.e. some observations are considered "more representative" of an "underlying" distribution than others (a violation of the likelihood principle, information is magically added that was not a priori in the model nor in the observations); or (2) that the measure reflects a utility and therefore is a measure of value rather then quality. We will try to clarify this in the paper.

On truth, science and verification

> In (4660: 17-19) the authors state: “In meteorology, the evaluation of quality is called verification (Latin: veritas=truthfulness). This term is somewhat misleading, because establishing that a model simulates the truth is impossible (Oreskes et al., 1994)”. I thank the authors for citing that very interesting paper and I totally agree that the term verification could be misleading. But, in my view, this is only because the truth of a model cannot be demonstrated, not because “establishing that a model simulates the truth is impossible”. In fact, Oreskes et al. (1994) also state: “(...) A model, like a novel, may resonate with nature, but it is not a “real” thing”, which means that a model can simulate (Latin: similis=similar) the truth but cannot be true. The simulation capability of a model can be established “if it is consistent with our experience of the natural world”. Otherwise, why do we have to mind about the evaluation of a single forecast using observations?

Indeed, if we regard simulation as being similar instead of an exact match, then we can demonstrate that a model can simulate the truth. However, we still have to resort to some arbitrary definition of similar. I think an important point in Oreskes et al. (1994) is that even though a model perfectly predicts all observations so far, we are never sure about the accuracy of predictions about the future. Maybe we should change line 18-19 to “..., because the truth of a model cannot be demonstrated.

My view on Oreskes et al. (1994) is that they do not pay enough attention to probability. I think a probabilistic view on the philosophy of science (see (Jaynes and Bretthorst, 2003; Solomonoff, 1964)) is far more coherent and precise than Popper’s view of rejection and corroboration of hypothesis, which seems to be the basis for the Oreskes et al. paper. In other words, ultimately science is not about testing and rejecting theories, categorizing them as "false" and "not yet proven false", and more or less “corroborated” it is about making probability estimates of future observations based on past

observations. I think Solomonoff’s theory of algorithmic information theory on inductive inference using Turing machines and attaching probabilities based on program length is more consistent with my view on science, with a central role for probability and information (see Solomonoff (1964)). In other words: indeed we can never establish that a model or theory is consistent with the truth, in the sense that we can expect it to always yield perfect deterministic predictions. We can however attach a certain likelihood to a theory and its predictions, based on all past observations. This is the best we will ever be able to do. The highest ideal in science is thus a well-calibrated probabilistic prediction, that for some physical laws may be so sharp that in practice we regard it as a deterministic prediction. Unfortunately, even a perfectly calibrated probability estimate is unattainable. It think that Solomonoff’s framework shows how it can theoretically be done, but that due to Turings proof that there is no solution to the halting problem, the probability is incomputable, and with finite computational resources can only be approached. This is a comforting thought: Science can in principle be conducted by machines, but they can only be perfect with infinite computational resources. We already have had years of evolution to improve our tricks to do it as well as possible with finite, but significant resources and it will take a while before we will outsource ourselves :) I must admit that I am only just beginning to scratch the surface of these theories, but I wonder why they received relatively little attention outside the artificial intelligence community, compared to the ideas about rejecting theories and finding a single best model, which is basically throwing away information.

Another interesting point regarding the difference between models and truth can be taken from Deutsch (1998). Taking his view, we can think of models as simulators of virtual reality. When a model can produce the exact same input to our senses as reality does, then we cannot distinguish between a model and physical reality. Both can be regarded as algorithmic processes that generate inputs to our brains that have some pattern in them, which represent physical laws that we define as truth. In that sense, the truth of the truth cannot be demonstrated either, since it is also impossible to tell the difference between truth and a virtual reality generated by model simulation.

About the comparing forecasting systems vs. forecasts with observations

> Anyway, this is not a rhetorical question. In fact, what we can find also in the paper by Oreskes et al. (1994) is that "Models can only be evaluated in relative terms, and their predictive value is always open to question. The primary value of models is heuristic". This is a very interesting issue that the authors raised by citing that paper. Do not the authors think that a comparative evaluation of weather forecasting systems (e.g., Ehrendorfer and Murphy, 1988) should be of greater use instead of evaluating single forecasts by comparison with observations?

Indeed, ultimately one would want to compare forecasting systems. An evaluation of forecasts by comparing them with observations is part of such an evaluation of a forecasting system. I think the paper you mention by Ehrendorfer and Murphy (1988) is very interesting and I think it will be interesting to study the concept of sufficiency in the information-theoretical context. The framework that is presented by them also compares forecasting systems by both comparing them to observations first. A similar comparison can be done by comparing the divergence scores resulting from verification of the forecasts from two different systems. This gives an indication about the relative qualities of both systems. Based on the fact that a forecasting system trained to maximally exploit the data, we can conjecture that we can never find a forecasting system that is sufficient for one that is trained on minimizing the divergence score, if all forecasting systems use the same input data.

Another important point in going from comparing series of forecasts to the forecasting systems that generate them is model complexity. Especially when we train a forecasting system for maximally informative forecasts, we run the risk of overfitting if no penalization for model complexity is included in our evaluation.

Jensen-Shannon divergence

> Finally, about the question raised by Anonymous Referee #1 (2010) in (C2246: 17-21), could the Jensen–Shannon divergence (Lin, 1991) in place of the Kullback–Leibler divergence be of use?

I think in the context of forecast verification, the property that the score is unbounded for a forecast that states something with certainty, but turns out to be false, is in fact desirable, as its error in probability is worse than any other can be. The fact that the Jensen-Shannon divergence is bounded thus makes it less desirable. The fact that infinite scores cannot be handled in verification studies should thus be seen as an argument against ever assigning a zero probability to anything unless it is truly impossible, and not as an undesirable artifact of the score. Furthermore any local score that is not unbounded will necessarily not be a proper score and can thus be hedged. See also the reply to reviewer 1 on this point. However, it may be interesting to study the application of this divergence in other contexts, after carefully studying all its properties..

Technical corrections

We performed all the proposed technical corrections.

References

D. Deutsch. *The fabric of reality*. Penguin Books London, 1998.

M. Ehrendorfer and A.H. Murphy. Comparative evaluation of weather forecasting systems: Sufficiency, quality, and accuracy. *Monthly Weather Review*, 116(9):1757–1770, 1988.

E.T. Jaynes and G.L. Bretthorst. *Probability theory: the logic of science*. Cambridge University Press, Cambridge, UK, 2003.

A. H. Murphy. What is a good forecast?: An essay on the nature of goodness in weather forecasting. *Weather and Forecasting*, 8(2):281–293, 1993.

N. Oreskes, K. Shrader-Frechette, and K. Belitz. Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, 263(5147):641, 1994.

RJ Solomonoff. A formal theory of inductive inference. Part I. *Information and control*, 7(1): 1–22, 1964.