Hydrol. Earth Syst. Sci. Discuss., 9, C521–C529, 2012
www.hydrol-earth-syst-sci-discuss.net/9/C521/2012/

Hydrology and
Earth System
Sciences
Discussions

Interactive comment on "Ideal point error for model assessment" *by* C. W. Dawson et al.

C. W. Dawson

C.W.Dawson1@lboro.ac.uk

Received and published: 27 March 2012

Author's Response to Anonymous Referee #2

27th March 2012

We are grateful to Anonymous Reviewer #2 for providing a detailed consideration of our paper. This review raises a number of challenging issues and questions, which we are able to respond to fully below. We respectfully acknowledge the reviewer's opinion on many of the issues raised, but do not agree with much of the criticism that is levied at our work. Indeed, we would assert that many of the negative comments on Ideal Point Error (IPE) made by the reviewer are in complete agreement with the arguments that are developed in our paper and actually serve to reinforce the reported analysis. We also argue that this paper provides a necessary and timely investigation of the potential and limitations of a newly emerging integrated hydrological modelling performance

metric – an analysis that was not fully developed or considered in the papers in which IPE was first presented and defined (Elshorbagy et al., 2010a,b). Moreover, the approach and format by which we evaluate IPE is consistent with other papers published in this journal (Beran, 1999), in Hydrological Processes (Seibert, 2001; Criss and Winston, 2008) and in ASCE Journal of Hydrologic Engineering (Jain and Sudheer, 2008).

The reviewer has been helpful in presenting a set of separate, clearly defined criticisms and questions. We respond to each of these in turn below. To aid clarity, we provide both the reviewer's text and our response in a numbered format.

Detailed Responses to Reviewer's Comments

Comment 1: In this manuscript, the authors discussed a particular error measure and its suitability for assessing the performance of hydrological models. The manuscript focuses on a measure named Ideal Point Error (IPE), which was developed and published earlier by others. Apparently, as the authors indicated, the IPE can have many variants (forms) that can decrease the lack of generality of the original one. I am sure that different researchers can suggest different variants and actually different error measures – in fact we can come up with endless variants of IPE and numerous error measures but one important question will remain: What is the significance/importance of this? I strongly believe that the authors failed to address this.

Response 1: In this comment, the reviewer emphasises the endless variants of IPE that are possible – a fact that we make very clear in our paper and provide important technical guidance on in the early sections. The importance of sound, technical understanding of the potential issues associated with applying IPE as an evaluation tool for hydrological models is self-evident. Yet, this guidance was not fully developed in the original papers in which IPE was defined (Elshorbagy et al., 2010a,b); thereby necessitating the further technical considerations presented here.

However, this paper is about far more than the technical issues surrounding different IPE variants. The crux of the paper is about the advantages and limitations of adapting

IPE from a unique, absolute measure, to one that is standardised against a benchmark model so that it can be used as a generic measure of relative model performance. It is this adaptation that, to some extent, also addresses the problem of endless variants of IPE. Indeed, the conclusions (Page 1688) summarise this key argument clearly on lines 6-10:

"IPE equates to a moving target which is dependent on the model combination used. Hence, results and conclusions drawn from the analysis are unique to the set of models used in calculating IPE. A more generic use of IPE has been discussed in which a naive t+n step-ahead model is adopted for benchmarking purposes."

This reviewer's comment therefore fails to adequately recognise the key arguments and ideas presented in our paper. This failure is particularly curious given that Comment 11 (below) engages directly with issues surrounding our use of benchmarks; suggesting that the reviewer is aware of the importance and significance of this aspect of the work.

Comment 2: There is no "meat" at all in this manuscript that warrants its publication in a Hydrology Journal. It is a comment on the so called IPE that can be really summarized in one page and included as part of a real paper about hydrological modelling.

Response 2: We reject the assertion that our paper could be adequately summarised in a single page but respectfully note that it is relatively short. This brevity reflects the succinct discussion that is presented. The "meat" of the paper is twofold:

1) The provision of the 'missing' discussion about the technical limitations of IPE that was absent in Elshorbagy et al.'s original paper and that is essential if hydrologists are to properly understand cases in which IPE may / may not be applied without additional adaptation; 2) The theoretical consideration of the relative merits of using IPE as a stand-alone or standardised (relative to a benchmark) evaluation tool.

We reject the idea that IPE could be dealt with as a 'comment' in a broader paper. We respectfully remind the reviewer that IPE was originally presented as a minor element

in a broader hydrological paper and the result was an insufficient consideration of the issues surrounding its application in hydrology; thus necessitating this follow-up paper.

Comment 3: In literature, there are several papers that talk about model assessment, and they are much more profound than this manuscript. This manuscript does not add any significant knowledge and cannot be stand alone paper.

Response 3: The purpose of this paper is not to be particularly profound but to highlight and discuss important technical and theoretical considerations surrounding IPE. To this end, our paper follows the general format and approach of past stand-alone hydrological publications that consider metrics in this way. These include papers published in this journal (e.g. Beran, 1999); Hydrological Processes (Seibert, 2001; Criss and Winston, 2008) and ASCE Journal of Hydrologic Engineering (Jain and Sudheer, 2008). It is, therefore, representative of current practice in hydrology and has value as a stand-alone publication.

Comment 4: The authors emphasised on making the IPE more general and transferable to other studies, but when I look at all variants proposed in the manuscripts, I noticed that the emphasis was placed on minor issues when more important issues were ignored.

Response 4: We struggle to understand what the reviewer thinks should be more 'important' than the questions we raise about whether IPE should be better applied in a standardised manner using naive model benchmarks. The issue of benchmarking performance metrics to derive more transferrable evaluation products has been of concern for more than a decade (c.f. Sibert, 2001). The paper is structured such that more minor, technical points are developed in the early sections of the manuscript and more substantive discussions are developed in Sections 4 and 5. The reviewer has clearly not given sufficient weight to the importance of these latter sections.

Comment 5: For example, all variants imply equal weight for all error measures included but who said this is right? Is RMSE as important as the bias? Isn't this dependent on the application and what the model is intended to predict?

Response 5: This general comment supports our general statement made on Page 1673, Line 25-29 and, therefore, the need for a more detailed technical consideration of the problems in applying IPE. Moreover, the reviewer's questions about whether a measure used in IPE should be considered more or less 'important' are fundamentally flawed because of the difficulties associated with defining 'importance' in a context of different models and application domains. A far better approach is to consider discriminatory power (see Dominguez et al., 2011). Indeed, a far more pertinent question, which is the one that we consider in the Section 2.4., is the question of which metrics and combinations offer the greatest discriminatory power.

Comment 6: Doesn't this defeat the argument that one IPE is good for all studies? Isn't changing the weights leading to different results?

Response 6: This is exactly our argument about the difficulty in using of IPE as a stand-alone performance measure as made in Lines 6-7 on Page 1688.

Comment 7: This is just one example of issues with the IPE and any other error measures, I don't mean to encourage the authors to go ahead and make a new story about how the weights can make a new error measure because I believe there are more important stuff to be done in hydrological modelling.

Response 7: We would agree that the issue of weights in IPE may not be the most important issue facing hydrologists (in the broadest sense) at present. However, for the reviewer to dismiss the arguments surrounding integrated metric measures such as IPE so easily is to fail to recognise the importance of understanding the limitations of different methods for evaluating model performance. Model performance validation through fit metrics is critical – especially in the large literature on data-driven modelling approaches, which is where IPE originated. Indeed, in data-driven modelling, there is arguably nothing more fundamental than the veracity of the fit metrics used by the modeller to select and validate their solutions as this is often the only evaluation tool

available. In this context we argue that our paper has importance, but accept that such context should perhaps be presented more clearly in the revised manuscript.

Comment 8: Such work might have a little bit of significance or utility if it is extended to first apply it on a real case studies (rather than this artificially engineered errors), then take this IPE and use it as a cost function in an optimization problem to show how this can improve optimization algorithms or model calibration. This was not attempted by the authors.

Response 8: Our paper presents a theoretical approach to understanding IPE and its limitations in a manner that is similar to previous critical papers on metrics for model assessment (see studies cited in Response 3). Whilst case studies can offer detailed and specific insights, consideration of more general theoretical aspects ensures that the results of a case study can be properly interpreted. Theory must, therefore, precede case studies, and this is the motivation behind our paper.

Comment 9: Page 1682, Lines 10-15: misleading argument because ME is meant to measure the bias, so if the RMSE is large but ME is zero, then it is indeed a good model from the bias point of view. This is what ME is intended to measure.

Response 9: We accept that the use of the term 'poor model' on Line 12 is overly simplistic and subjective. We agree that our interpretation of the model performance could be better qualified in the manner suggested by the reviewer.

Comment 10: Page 1682, Lines 16-22: Trivial conclusion and basic knowledge.

Response 10: This is not a conclusion – it is a comment that reiterates the basic justification for an integrated metric measure, which is indicated by our results. It is surely encouraging to note that our results correspond to fundamental principles.

Comment 11: Page 1684: There is an exaggeration in the discussion about the benchmarking issue because such error measures are really intended to compare models against each other; not really against other models applied to other case studies. This is not realistic and such ideas should not be propagated without solid proofs.

Response 11: The reviewer adopts a very ideological stance on the use of benchmarking, which is not consistent with the range of views expressed within the hydrological literature. In physically-based modelling the use of benchmarks is arguably irrelevant because the model is assumed to represent the physics of the hydrological system adequately, and the Physics is the standard. However, in every other approach to hydrological modelling, benchmarks are required due to the lack of a 'physical' standard. Indeed, without any standard benchmarks, every investigation must be considered unique, and it would be impossible to cross-compare models or their performances. This negates the potential for knowledge generation through model comparison studies. On the type of benchmarks that offer the most potential, Seibert (2001) states:

"Obviously, there are more rigorous benchmarks that can be used...... We can also use the observed runoff, shifted backwards by one or more time steps. In this case, we use the observed runoff at time step t as a prediction of the runoff at time step t+n. This type of benchmark is especially suitable for forecast models."

This is clearly supportive of the approach taken in our paper and would be happy to make a stronger justification of this in the manuscript.

Comment 12: Page 1686, Lines 26-29 and also the Conclusions section: There is no way that you can claim the Naïve (t+4) can be transferred to other case studies as a benchmark model. It is study dependent.

Response 12: This is a restatement of the reviewer's argument which was dealt with under Response 11.

Comment 13: And what about other hydrological applications that are not flow forecasting?

Response 13: This is a reasonable point in the respect that our paper does only consider IPE from the perspective of river flow forecasting. Adjustments to the introduction

should be possible so that this is more clearly made. It should, however, also be noted that the original definition and application of IPE (Elshorbagy et al., 2010a,b) was in the context of a range of hydrological modelling problems.

Comment 14: The whole issue of correlation among error measures (reported in Table 3) does not make any sense. Does the strong correlation (0.98) between MARE and ME means that there is redundancy?! They measure completely different things, and both need to be reported. This misleading argument should be removed if this manuscript is published anywhere.

Response 14: This is, in effect, a restatement of the argument made in Comment 5. We argue that the measures used in IPE should be selected on the basis of their discriminatory power (i.e. according to the relative performance difference between individual models as opposed to their absolute performance), not according to the specific components of error that they measure. In this context, the strong correlation between MARE and ME indicates substantial redundancy in the two metrics, in the respect that the statistics offer little basis for discriminating between our models. It is evident from the reviewer's interpretation of our conclusions that we should consider reinforcing our argument to make things even more clear.

Comment 15: Page 1674, Line 15: "Average" should be "Absolute". Page 1675, Eq. 1: "MARE" in the first component should be "RMSE". Page 1677, Eq. 3: A plus sign is missing.

Response 15: The reviewer is correct. These need amending.

References

Beran, M. (1999) Hydrograph prediction – how much skill? Hydrol. Earth Syst. Sci., 3 (2): 305-307.

Criss, R.E. and Winston, W.E. (2008) Do Nash values have value? Discussion and alternate proposals. Hydrol. Proc. 22 (14): 2723-2725.

Dominguez, E., Dawson, C.W., Ramirez, A. And Abrahart, R.J. (2011) The search for orthogonal hydrological modelling metrics: a case study of 20 monitoring stations in Colombia. J. Hydroinform., 13: 429-442.

Elshorbagy, A., Corzo, G., Srinivasulu, S. and Solomatine, D.P. (2010a) Experimental investigations of the predictive capabilities of data-driven modelling techniques in hydrology – Part 1: Concepts and methods. Hydrol. Earth Syst. Sci., 14, 1943-1961, doi:10.5194/hess-14-1931-2010.

Elshorbagy, A., Corzo, G., Srinivasulu, S. and Solomatine, D.P. (2010b) Experimental investigations of the predictive capabilities of data-driven modelling techniques in hydrology – Part 2: Application. Hydrol. Earth Syst. Sci., 14, 1931-1941, doi:10.5194/hess-14-1943-2010.

Jain, S.K. and Sudheer, K.P. (2008) Fitting of hydrologic models: a close look at the Nash Sutcliffe Index. J. Hydrol. Eng., 13 (10): 981-986.

Seibert, J. (2001) On the need for benchmarks in hydrological modelling. Hydrol. Proc., 15 (6): 1063-1064.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 9, 1671, 2012.