Hydrol. Earth Syst. Sci. Discuss., 5, S2526–S2534, 2009
www.hydrol-earth-syst-sci-discuss.net/5/S2526/2009/

Hydrology and
Earth System
Sciences
Discussions

Interactive comment on "Analysing the temporal dynamics of model performance for hydrological models" *by* D. E. Reusser et al.

D. E. Reusser et al.

Received and published: 3 February 2009

We would like to thank E.J. Pebesma for his review. We think that he raises some important questions which will help to improve the readability and understandability of our manuscript.

E.J. Pebesma states that our approach might "create more obfuscation about what the results exactly mean than enlightenment". This stands in contrast to the comments by the other reviewers who write that we offer "a novel methodology that can be used to identify model weaknesses. Many of the model problems that Reusser et al. identify would not be evident by simply looking at summary skill metrics." (Clark 2009). Bernardara (2009) states that he "appreciates the use of a very wide number of error measures to check the whole range of model errors".

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

From all the specific comments given by Pebesma (2009), we think that comment number 2 is the one that explains best why he reached the general conclusion quoted above:

"The authors deliberately decide not to choose a single model performance measure, but rather decide to analyse every measure they can find, even time varying, and then they complain that this creates a lot of information that is hard to interpret - hence they need SOMs. I would have found it stronger when the authors had chosen one (or a few) measures that served a particular goal, and had concentrated on that."

We would like to thank the reviewer for his comment. It shows, that we did not succeed to make the underlying idea as clear as we would have liked to. We will try to make this clearer in the revised version.

Our idea starts with the fact that a single measure is not able to catch all the features that should be reproduced by a hydrological model. Multi objective approaches use a (large) number of performance measures to capture various aspects of the difference between model and observation. The combination of these performance measures acts as "finger print" for the difference. Such finger prints can be calculated for the entire time series or for time windows covering only a part of the time series. The finger print will be similar for time windows where the difference between model and observation has similar characteristics. Identifying and characterizing periods with comparable finger prints gives a tool to:

- objectively separate periods of differing model performance
- identify characteristics that are not easily found by visual inspection
- find recurrent patterns of differences between model and observation in longer time series.

We would like to point out, that we propose a different perspective on model performance with the goal to answer what goes wrong and when. This provides valuable information for error identification. We do not claim our method is superior but complementary to the usual approach.

The approach suggested by E. Pebesma ("to chose one (or a few) measures that served a particular goal") has a different underlying basic idea and has been demonstrated in Pebesma et. al 2005. Every model calibration exercise serves a particular goal. From this perspective, the selection of a particular performance measure seems a priori straight forward. For a typical engineering application such as simulation of extreme flows, such a performance measure could be the correlation between observed high flows and modelled high flows. The situation is, however, slightly different if during model development, the general goal is to mimic the natural process as closely as possible under the given data constraints. If we try to select only one or very few performance measures, this would imply that we already know beforehand what is actually "going wrong", i.e. which errors are likely to occur and when.

We are aware that the method includes multiple steps which may be difficult to follow for a reader. We tried to make the structure of the manuscript as clear as possible. The comments from E.J. Pebesma and the other reviewers will certainly help to further improve the structure and thus its readability. We are also aware that understanding the results requires some effort and preferably some practice. However this is the case for most novel approaches and has been specifically addressed by Cloke et. al. (2008) with respect to novel performance measures. Thanks to the specific comment number 5 (see below) we are currently testing whether removing highly correlated performance measures is really necessary. Hopefully, we will be able to simplify the procedure by leaving this step out. Thanks to comment 3 we reconsidered the transformations applied to the performance measures again. We are currently testing whether for the sake of consistency one single transformation can be used for all performance measures. These planned changes will also make the procedure easier to understand.

Specific comments

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

1. The authors mention on page 3172 that there are no studies on high resolution temporal dynamics of model performance. One paper that is out there is Pebesma (2005). This paper looks at structure in error time series, e.g. by analyzing its autocorrelation, and looks at how well errors can be predicted, e.g. by a lagged and/or smoothed versions of model input (rainfall). The paper gives also some relevant references to earlier literature.

The paper shows the temporal dynamics of the model error defined as the difference between the observed and predicted time series. Linear models are then used to predict the model error from the prediction and/or the rainfall. As this paper focusses on single events instead of longer time periods, it was not among our cited references. In our opinion, evaluation of single events compared to longer simulation periods are quite different problems. We are not aware of any subsequent publication demonstrating that the approach is also useful for the evaluation of longer periods. However, we will consider to cite Pebesma et. al 2005 in a revised version.

2. I find the problem the authors introduce on page 3173, namely that "The large amount of data produced in such an analysis quickly becomes overwhelming and even confusing" a consequence of the author's decision rather than a fact of life. The authors deliberately decide not to choose a single model performance measure, but rather decide to analyse every measure they can find, even time varying, and then they complain that this creates a lot of information that is hard to interpret - hence they need SOMs. I would have found it stronger when the authors had chosen one (or a few) measures that served a particular goal, and had concentrated on that. Which particular goal serves the full collection of performance measures they choose?

See the introductory answer were we discuss this comment. We will try to make our idea clearer in the revised version.

3. as the Nash-Sutcliffe measure is widely used and the authors decide to use a very wide range of measures, it is unclear to me why they used a transformed version of the

NS measure, and not both the original one and the transformed one.

As we write on p3180 L9 the SOM is sensitive to very distorted distribution shapes. Therefore we transform some of the performance measures and normalize to the range [0,1].

On page 3175 L17 we write, that the NS results in large negative values during periods with little dynamics. If the original Nash-Suttcliffe efficiency was used, the SOM would mainly distinguish between "very large negative values" and "values close to 0 (or 1)". This is not very useful information. We would also like to point out the transformed version of the NS measure is equivalent to the original version in the range [-1, 1] for which NS values are usually reported.

We are aware that the transformation of the performance measures currently is a weakness in the method as it makes the method more complicated. In answer to this comment, we are testing whether the method works if all performance measures are consistently transformed to uniform distributions. If this works, both versions of the NS measure are equivalent for the subsequent steps.

4. Table 1 with the list of performance measures raised a few questions;

We are aware that the performance measures are discussed very briefly. A longer discussion is not possible as the manuscript is quite long already. Also, the two sources we cite define the measures well.

how does NSC measure model error?

The number of sign changes of the residuals is low if there is a bias. It was introduced by Gupta et.al 1998. We will update the manuscript to include this information

With t-test, is a paired or a two-sample t-test indicated?

A paired test is used as defined by Dawson on the hydrotest homepage. We will update the manuscript to make this clear.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

MALE - how can you take a log of a negative error?

On page 3176 line 16 we write that performance measures for log transformed date are indicated with an ”L”. Therefore, MALE is the mean absolut error for the log-transformed data (not error).

For each of them: are they computed over a time window, and if yes over which window?

Page 3174 L23: ”evaluation of the set of performance measures for a moving time window; this yields a vector of performance measures for each time step;”

Page 3178 L1: ”A time window of 10 days and 5 days was chosen as a compromise between looking for the local properties in the time series and having enough data to actually compute the values for the first and second case study, respectively.”

How does the window size influence the research findings?

p3190 L12ff: ”The entire case study was repeated two more times, once with a time window size of 5 days and once with a 15 day window, in order to test the sensitivity of the method for this choice. Detailed results from this comparison are available from the corresponding author. In short, the alternative window sizes resulted in 3 and 4 clusters for the 5 and 15 day window, respectively. Clusters A and E were found in both cases with equivalent descriptions of the error types and temporal occurrence of the error clusters.” We will provide further material on the authors homepage.

5. Table 2: R>0.85 what does that mean? R-squared? or absolute correlation? Or was correlation never strong negative? The 0.85 seems arbitrary (was it important for the analysis?), and weakens the approach. Multivariate statistical techniques are designed to deal with highly correlated data sets, so why the need to preselect?

The absolute correlation was used. We agree that 0.85 is an arbitrary choice. As an answer to this comment, we are testing the method with the entire set of the performance measures and we are confident, that the results are not strongly influenced.

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

However, for reporting the results, we will present only selected measures (probably selecting based on correlation). In addition, by identifying strongly correlated performance measures, we can find groups of performance measures that point on similar structural deficits. Knowledge about these groups may be of use for the selection of an appropriate measure for a certain situation.

6. Page 3179: how is the layout of the SOM (xmax, ymax) chosen, and does it have a consequence for the analysis? Does it matter if the map is square or elongated?

We would like to thank the reviewer for pointing out that this has to be checked. We will report on the influence of the layout of the SOM in the revised manuscript. We do not expect a large influence.

7. The need to apply fuzzy clustering after a SOM makes the whole interpretation very ”soft” and hard to follow.

After creating the SOM, we are interested to group SOM-cells with similar properties in terms of performance measures for the subsequent interpretation. The limit between these groups will of course not be very sharp and we thus think that fuzzy clustering, where the SOM-cells will have a degree of membership to the different groups, is a valid method to do this.

8. The authors acknowledged the R community; a better way of thanking them would be to add the literature references to R and the packages used to the text and references list.

We do cite Cottrell and Bodt (1996), Jachner et al (2007)

As suggested by the reviewer, we will mention the packages used for this study in a revised version version and also include a reference to

R Development Core Team. 2008. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing: Vienna.

9. page 3194, l 21/21: please point out explicitly a couple of examples of less obvious error types that were overlooked first - the real benefits of this approach.

The benefits from this method are: 1) Identification and separation of time periods with different model performance characteristics are achieved in an objective way.

2) Long simulation periods for which analysis of single events becomes almost impossible can be processed. Recurrent patterns become apparent.

3) Subtle, but important differences between observation and model can be detected.

- Weisseritz
  - timing problems mainly during the winter. These were hard to detect when looking at the data from the entire simulation period.
  - consistency of underestimation during summer.
- Malacahuello: Identification of time periods with subtle differences between measured and simulated time series as a result of
  - noise on measured data (time scales of minutes)
  - daily fluctuations in the measured date (possible temperature effects on sensor)

We will update the manuscript to list these benefits more explicitly.

10. When creating "types of errors", errors in both ways were created. Eg. Fig 2, subfig 1 shows peak over- and under-shoots. Is it with this method still possible to classify ("find") a behaviour that is a "peak overshoot", or is a classification only possible as "peak over- or under-shoot"?

As the case studies show, it is still possible to classify a behavior that is a ”peak overshoot”:

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Discussion Paper

For the Weisseritz case, cluster B and D correspond to overestimations and cluster C and E to underestimations (Table 5, Sect. 5.4).

For the Malalcauello case, cluster F and C correspond to underestimations and cluster A to overestimation (Table 5, Sect. 6.2).

However, we do also see cluster for example cluster B in the Malalcahuello case study for which both, over- and underestimations are observed. This is due to the cluster analysis. In this example, the absolute maximum error has more weight during the clustering as shown in Fig7b in Reusser et.al (2009).

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., 5, 3169, 2008.