Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-339-AC2, 2016

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Looking beyond general metrics for model comparison – lessons from an international model intercomparison study" *by* Tanja de Boer-Euser et al.

Tanja de Boer-Euser et al.

t.euser@tudelft.nl

Received and published: 7 October 2016

Dear reviewer,

Thank you for your positive and encouraging review of our manuscript. We agree with you that this study is a nice opportunity to extract more knowledge about different processes controlling runoff at different stages. Therefore, we really appreciate your detailed comments and suggestions, which can help us to improve the manuscript. We would like to respond to your comments below.

Abstract

I think the abstract can be a bit longer. I miss this sentence from the Methods section

to describe the analysis made in the experiment "Three types of statistical analyses and comparisons of simulation results and observations were conducted: cumulative discharges, empirical extreme value distribution of the peak flows and flow duration curves for low flows." I would also recommend highlighting in the abstract more findings when it comes to understanding controls of these parts of the hydrograph, related to differences in model performance, as well as, the fact that some parts of the hydrograph is poorly understood and here we can thus identify present knowledge gaps.

We are a bit hesitant to really increase the length of the abstract; however, we agree with you on the missing elements. Therefore, we will restructure the abstract in the revised version of the manuscript.

Introduction

The first part is a bit too general and has many references to the literature, which are not necessary for the understanding of the paper and as a reader I miss guidance to how the citied papers actually contributed to the topics discussed. Often the references are lumped and support general statements that doesn't help the reader much in deeper understanding of the literature and previous work, see example below.

Page 2, row 12: reading "Hydrological studies at different scales and under different climates have shown a large variety of hypotheses on hydrological functioning (e.g., McDonnell, 2013; Zehe et al., 2013; Fenicia et al., 2013; Clark et al., 2016; Seibert et al., 2016)." This is an example of a very general statement with no guidance to what hypothesis the references refer to. Please, give examples and divide the lumped chain of references to explain to the author why these references are chosen – what variety did they show in hydrological understanding?

Page 2 row 22: I suggest starting the Introduction here – the paragraphs above don't contribute much but are common knowledge. This is where the study is motivated and from this point the text is much more interesting and straightforward. (Avoid reference dropping!)

Replying to the three comments above, we agree that the first part of the introduction is rather long and general. However, we prefer to keep some of it to give guidance to readers with a smaller modelling background. So, in the revised version of the manuscript we will reduce the length of the first part of the introduction and make it more specific for the current study. In line with this we shall elaborate a bit further on the selected references.

Page 2 row 31: This sentence is difficult to understand: Ceola et al. (2015) concluded that deriving the causes of performance differences between various model structures is not trivial, mainly due to the considerable differences in model structures which disturbs the identification of model features that increase model performance. Are you trying to explain the problem of equifinality here? (i.e. that many different parameter settings may result in similar model performance, due to compensating processes in the model description?) Please, rephrase!

Thank you for pointing this out; however, equifinality is not what we are aiming at in this sentence. Rather, we wanted to point out that because model structures are rather complex, it is difficult to compare them in an intercomparison study and to derive conclusions on model performance based on the presence or absence of certain processes in the model schematisation. In our study, we tried to overcome this problem by presenting schematic pictures of the model structures in Figure 2 of the paper.

We therefore suggest rephrasing the sentence as follows: Ceola et al. (2015) pointed out that previous intercomparison studies have contributed little to deriving the causes of performance differences between various model structures. This could be attributed to the complexity and the large differences of model structures, and to the difficulty to link the presence of a model feature to a better or worse performance.

Page 4 row 15: I guess the notation of 'local time' can be removed in this decadal context.

Thank you for pointing this out, we will remove the notation of 'local time'.

Method

The experiment is very straight forward and described with relevant level of details. I like the schematic pictures of model structures of Fig 2 and the descriptions in Table 2. Very useful for understanding!

Thank you! Nice to hear that the figure is very useful.

Results

I suggest including statistical metrics for model performance seen in Fig 3-9. Each graph can be assigned with a (few) metric(s) for the data it is showing, to support the analysis.

See also our reply to the third comment of the first reviewer. We agree that adding metrics to the figures will indeed support the analysis. For the revised version of the manuscript we will try to find metrics that reflect the patterns the human eye observes as well.

Section 4.2 Modelling the highest peaks: please, elaborate on potential causes why some models are more successful than other to capture the peaks. Are there processes they did describe that others did not? Where they more carefully calibrated? (did they apply other considerations for parameter choices?)

It is a good point to have a more thorough look at whether the differences in capturing the highest peaks can be explained by differences in model structures. However as the results observed in the Lesse are not consistently observed in the other catchments (supplement, Section 4), it seems difficult to draw sound conclusions on this topic with regards to model structure components. Calibration certainly plays a role as shown in Fig 9 of the paper for the VHM model, where the same model performs better when another calibration objective is used. However, in Figure 4, all models were calibrated using the same objective functions, so we do not expect that calibration can really explain the differences between the models.

Fig 3: it would be more interesting to see relative error of annual flow instead of absolute; at least for a reader who is not familiar with the catchment but with general model performances.

We agree and will change the figure.

Section 4.5 Transition from low to high flows – what can be learnt from these results of poor model performance? Interesting that a snow routine was not necessary, but could be compensated for. . . how?

Regarding the transition from low to high flows, most models show an overestimation of flow for most catchments. This indicates that the rewetting of catchments works differently from what is currently assumed in the models. The variability in performance between catchments further indicates that the models are probably missing a process that is important during this stage of the hydrograph. We will elaborate on this in the revised manuscript.

Regarding the snow module: the influence of snowmelt on the discharge is small; however some snow cover does occur every winter. (generally snow accumulation is scattered in time and in the order of 20-50 mm SWE, and there are around 50 days with snow cover each year). Thus, by calibrating on NSE, many models were able to model the winter discharge right, even without a snow module. The effect of a snow module can better be shown and investigated when the same model is run with and without a snow module.

Discussion

Section 5.1: With the title 'Findings about the Meuse basin' I suggest you add some findings about which processes that seem to control flow of high peaks, low flow, etc. based on your results from model performance using different process descriptions/structures. Alternatively, you should change the title to "Model performance in the Meuse basin" but this will make the paper less appealing in my view. I would rather like to see the hydrological interpretation from the model results (i.e. change perspective

to describe nature instead of models).

We agree with you, Section 5.1 will benefit from a more detailed hydrological interpretation. In the revised manuscript we will link the findings more to modelled runoff processes.

Page 11, row 10: I suggest to rephrase "We therefore hypothesize. . . " to We therefore suggest. . .or rather: " The results thus indicate. . . ". Even if results may raise new questions, I think you should take the opportunity to make a statement from your study here.

Thank you for the suggestion, we will rephrase the sentence.

Section 5.2: Benefits of an intercomparison study: Please, specify more clearly what direct benefit you could draw from using a model ensemble in this catchment. It is a costly and time-consuming process, so what are the proofs from your results claiming that it is worthwhile?

This study provided some clear conclusions for model elements (or runoff processes) that are important during drying conditions, although this stage of the hydrograph is often difficult to model. Following on this, the study reflects the importance of model structures (or choices) for different parts of the hydrograph. A model ensemble and the background on individual performance can help in operational forecast for, among others, uncertainty estimation.

Page 11, row 28: I suggest acknowledging the importance of following a firm protocol (as you seem to have done here) in the collaboration, to ensure transparency for the result analysis and reproducibility of computational experiments. See for instance: Hutton, C., Wagener, T., Freer, J., Han, D., Duffy, C. and Arheimer, B., 2016. Most computational hydrology is not reproducible, so is it really science?. Water Resour. Res.. Accepted Author Manuscript. doi:10.1002/2016WR019285

Thank you for sharing this reference, which we will include in the paper, but maybe it

would be more appropriate in the introduction or method section.

Page 12 row 6: Note that it is quite common that catchment models underestimate extremes even though they do reasonable good on mean values, also in a multi-basin analysis. See for instance: Donnelly, C, Andersson, J.C.M. and Arheimer, B., 2016. Using flow signatures and catchment similarities to evaluate a multi-basin model (E-HYPE) across Europe. Hydr. Sciences Journal 61(2):255-273, doi: 10.1080/02626667.2015.1027710

Thank you for sharing this reference, which we will include in the paper.

Conclusion

I would recommend highlighting your findings in understanding controls of the various parts of the hydrograph, related to differences in model performance. Was it only events during dry conditions you could refer to improved processes understanding? Check if there is more from your work to extract here! For 'Transition from low to high flows' as well as extreme flows I think you should acknowledge that we see knowledge gaps in process understanding and identify weaknesses in the model descriptions.

For transition and wet conditions, the results were not very consistent between the different catchments, revealing the presence of knowledge gaps in understanding model features that causes these differences. We will acknowledge these gaps in process understanding briefly in the conclusions.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., doi:10.5194/hess-2016-339, 2016.