Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-339-AC1, 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: 5 October 2016

Dear reviewer,

Thank you for your positive evaluation of our manuscript. We appreciate the comments and suggestions you have made and would like to respond to them below.

1) It is a bit unclear if this is a comparative study or a model comparison (or intercomparison) of different process descriptions. The paper uses different terminology of comparative in different Places. Please check and make clear that it is consistent throughout the paper.

Thank you for this comment, we fully agree with you that the used terminology should

be consistent throughout the paper. The paper is meant as a model intercomparison and as such investigates whether certain processes are more suitable to apply in models than others. We will make sure this is consistent in the revised version of the manuscript.

2) How has the problem definition (setting objectives) for the rainfall-runoff model been 'stated'. When reading the paper, it seems like an overall objective has been to test different conceptual models overall performance for both high flows and low flows at the same time, as defined by the metrics (NSU + NSUlog transformed) and the qualitative high, low and transition flow performance. However, in the mission statement, how was this described for the different model teams/universities participating in the study (the paper is a bit vague on this issue)? I miss a bit on this, was the approach for qualitative testing of the different models know, or not know by the 'modellers'? I guess the model study objectives and the described objectives of the paper (paper 3, last 5 lines of section 1) may not have been fully the same, so I suggest that this 'mission statement' for carrying out the calibration and validation of each conceptual rainfall runoff model, should be briefly mentioned with a few lines, before the objectives of the paper is described.

Thank you for this comment, we realise that the current phrasing may be a bit confusing. The objective of the study was to force a set of models, used by different teams in the Meuse basin, with the same meteorological input and to compare the model outputs. To do this, we needed to specify two elements in more detail: 1) how should the models be calibrated; 2) how should the model outputs be compared (e.g., model evaluation tools). For the first one we decided that modellers probably know best how to calibrate their own model, so we only set the objective functions (NSE and NSElog) and the number of desired model realisations (20). For the second, we decided to base the comparison not only on the calibration criteria, but also on aspects that were not specifically taken into account during the calibration procedure, so as to investigate the full range of a model's capabilities. As the data was of good quality and the models

had already proven their usefulness, we were not surprised that the NSE and NSElog values did not discriminate much between the models. However, this makes the other findings even more interesting. Figure 1 shows again the difference between the calibration objectives and the model evaluation tools. We will make this distinction more clear in the revised version of the manuscript.

3) The comparison based on quantitative metrics (NSE and NSElog) and qualitative criteria (low flow duration curves, extreme value distribution plots, simulated timeseries for low flow, timeseries for transition from low to high flow and timeseries for winter periods), basically reveals that NSE and NSElog cannot discriminate between different conceptual models. Instead, events is introduced, which then is used for arguing that the need for including a very quick flow reservoir preceding the root zone storage, and a slow reservoir in parallel with the fast reservoir to model the recession for the Ourthe catchment. This is fine, but, it is a very subjective evaluation, since there is no metrics for how to evaluate which model gives the best results in the selected events (eg. Summer 2008 etc.). Why have you not calculated some statistics for these events, in order to measure in operational way, the differences in performance of the different models?

In the different steps of comparing the model output we did calculate a range of statistics and hydrological signatures. However, the main problem was that a certain metric only quantifies a specific element of the performance of a model. So, to make a fair comparison between the models (each models excels on other elements), a whole range of metrics was required. Following on this, the more metrics are used, the more difficult it is to present the results in a clear and easy to interpret way to the reader. Therefore, we used the human eye as evaluator in the first version of the manuscript. However, we agree with you that it makes the evaluation subjective. Thus, in the revised version of the manuscript we will try again to find some metrics that can quantify our observations for the different events.

1-3) As a general suggestion for improving the paper I would recommend a Figure as

part of the section on Methods or in the discussion which clearly illustrate the 'modelling protocol' (or suggestion for such after the study) and how the study went through the various steps in a rainfall-runoff modelling process (e.g. a workflow of the modelling process from define purpose, conceptual model building, setup of numerical models to split-sample/proxy basin tests).

Thank you for the suggestion. See our reply to your second comment.

Specific comments:
2.Study areas and data Since the forestry part is high (46 %), it could be relevant to unfold a bit more how interception storage is addressed in the various conceptual models.

We agree and we will briefly discuss the differences in interception module between the models in the revised version of the manuscript.

3.Methods Please add 1-2 references on page 5/line 6 to NSElog. Would it be possible to add to the description of each model the number of parameters which requires calibration? (did any modelling teams do sensitivity analysis before selection calibration parameters, if not calibrating on all parameters?) Either in the text or in Table 2.

We will add some references here in which NSElog is used as a performance indicator for low flows.

Regarding the number of parameters that were calibrated for each model, we will add them to Table 2. Regarding the sensitivity analysis, most of the calibration methods that used pre-filtering of the parameter space implicitly used a sensitivity analysis.

4. Results Please add a table with the results of NSE and NSElog for each calibration and validation period for each station (not the min/max but the average NSE and NSElog for the period). I don't think that it is fully satisfying for the reader to have to consult the background, supplementary material.

Here we do not fully agree: the suggested table would be too large (2 metrics * 5

stations * 3 periods * 11 models) and would mainly contain very similar numbers. As we do not think that such a table is informative to the reader, we prefer to keep the table(s) in the supplement. However, at the beginning of the result section we can add a sentence summarising the NSE/NSElog performance for the other catchments.

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

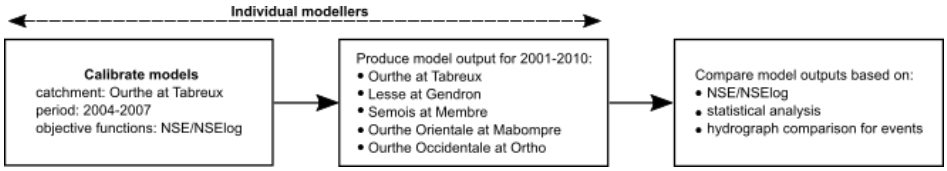

Fig. 1. Modelling protocol, highlighting the calibration criteria (left) and the evaluation tools (right)