Hydrol. Earth Syst. Sci. Discuss.,
doi:10.5194/hess-2016-339-RC1, 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.

Anonymous Referee #1

Received and published: 3 October 2016

It is a well written and interesting paper, which addresses relevant scientific questions within the scope of HESS. The concepts for model comparison are novel. There are substantial conclusions, about the need for including processes (quick flow reservoir/parallel groundwater storage), based on comparison of the different conceptual rainfall-runoff models, tested to for the same catchment / data. Scientific methods are valid and clearly outlined. However, even though scientific significance and quality in my view is excellent, when it comes to presentation quality, I would recommend minor improvements, before the paper is published. 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. 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 modelteams/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. 3) The comparison based on quantitative metrics (NSU and NSUlog) 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 NSU and NSUlog cannot discriminate between different conceptual models. In stead, 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? 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). Specific comments: Abstract. OK! 1.Introduction Well written. Ok. 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. 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. 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. 5. Discussion and 6. Conclusion OK.

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