Discussion comment 1

OVERALL RATING: The paper presented by Beck et al is concerned with the tedious but important task of model evaluation. Overall the paper is interesting in scope, well written and the results are clearly presented. Consequently, I do definitely support the publication of the presented work.

Nevertheless, I do have several comments/suggestions which the authors may wish to consider prior to the publication of the manuscript in a final form. For the sake of clarity, I do list “specific comments” and “small comments” below.

We would like to thank Dr. Gudmundsson for his positive remarks, thorough review, and useful remarks. Below we respond to each of his comments in green font.

SPECIFIC COMMENT:

** Specific Comment 1: ** While the paper is generally clearly written and most of the conclusions are supported by quantitative evidence there is a tendency for value statements (e.g. p. 2, l. 26: “NSE . . . is [a] . . . flawed metric”), claims (e.g. p. 1, l. 11: “. . . more effort should be devoted to calibration. . .”) or speculations (e.g. p. 9, l. 32: “. . .performed well. . . due to the lack of baseflow. . .”), which are not clearly highlighted as the authors interpretations or opinions. Although I do value if researchers defend their views on specific topics, I do also belief that it is important to clearly separate “hard facts” (either theoretical or quantitative) from soft interpretation and opinions in a scientific text. Therefore, I would like to encourage the author team to carefully revise the text of the manuscript, aiming at separating opinions from facts that are supported by either theory or quantitative analysis.

We agree and have therefore re-read the text with this in mind. The sentence “the Nash and Sutcliffe (1970) efficiency (NSE), which is increasingly considered to be a flawed metric for model performance“ was changed to “the Nash and Sutcliffe (1970) efficiency (NSE), which has been criticized in several previous studies“. The sentence “SWBM performed well only in arid catchments, at least partly due to the lack of baseflow under these conditions” was changed to “SWBM performed well only in arid catchments, probably at least partly due to the lack of baseflow under these conditions”.

** Specific Comment 2: ** I do highly value the analysis presented in Figure 3, as this is a compelling way to investigate the physical consistency of the considered models with respect to the coupled water and energy balance. Unfortunately, the authors did only conduct this analysis for four models that output potential evapotranspiration, Ep, which is used to compute the aridity index.

Thank you for the compliment. For Tier-2 of the eartH2Observe project we hope that all modeling groups release their *E*p and *R*n data.

An alternative approach, which was actually used by Budyko (1974), is to compute the aridity index as the ratio λRn/P where P is precipitation, Rn net-radiation and λ the latent heat of vaporization. This has the advantage that the results are strictly interpretable in the context of the coupled energy and water balance. In addition, this would allow to evaluate the output of all considered models.

This is a very good suggestion. We have re-done Figure 3 using the ratio λ*R*n/*P* for the models without *E*p data. We have also changed the text to discuss the results of the added models.

In addition, I would like to question the authors conclusion that the models are most realistic if they scatter around the Budyko-curve (ref. p. 12, l. 20). In fact, Budyko (1974) developed the curve on the basis of a limited number of catchment observations and it is a-priory not clear whether models need to be close to this empirical rule. I do, however, fully support the authors conclusions that data-points that are outside the energy and water supply limits are a strong indication for issues in the model physics.

So far nearly all (large-scale) studies have shown that observations scatter around the Budyko curve or similar curves. In light of this, we argue that, for a model to be considered an accurate representation of reality, it should exhibit a similar pattern. Note that we are not saying or implying that the models should scatter *closely* to the Budyko curve.

Finally, the similarity Figure 3 with the work of Greve et al (2014, doi: 10.1038/ngeo2247), who used the budyko framework to evaluate the credibility of reconstructions of E, P and Ep and Greve et al (2015, doi: 10.1002/2015GL063449, and sorry for citing myself), who introduced a formal way to account for scatter in the Budyko-space caught my attention. Although I am not sure if this is beneficial in the context of the presented paper, I could imagine that the tools provided in both mentioned studies might be helpful do develop additional quantitative insights to model performance.

Thank you for bringing these two papers to our attention. We could indeed use the approach of Greve et al. (2015) to quantify the deviation of the data points from the Budyko curve. However, we feel that summary statistics are not necessary since even a quick glance at the density plot already provides a wealth of (qualitative) information about the model behavior. Moreover, it is not our objective to quantify exactly how well each model "follows" the Budyko curve, since the Budyko curve is, after all, an empirically fitted equation (as recognized by the reviewer in the preceding comment).

SMALL COMMENTS:

Page 1, line 8-9: Repeated use of "(uncalibrated)". I assume one should be calibrated

To wanted to underscore that we refer in both cases to the uncalibrated models. We have removed the parentheses to improve the readability.

Table 1: The way the authors formulate the description of Table 1 reads as this would be a comprehensive review of model validation studies. I am, however, aware of at least two further studies – again, sorry for citing myself - (Gudmundsson et al 2012, doi: 10.1029/2011WR010911, Gudmundsson & Seneviratne 2015, doi: 10.5194/hess-19-2859-2015), that conduct similar assessments. Therefore, I would encourage the authors to either emphasise that the list of studies mentioned in Table 1 is not comprehensive, or to provide a more systematic summary of previous assessments.

Thanks for mentioning these two studies, we have added them to Table 1. In addition, we have added "to the best of our knowledge" to the Introduction and the caption of Table 1, to highlight that, while we made our best effort to find all studies, some may still be missing.

Page 4, line 10: ". . . the combination of Penman-Monthie equations. . ." reads strange.

Changed to "the Penman-Monteith combination equation", the more common name used in the hydrological literature.

Page 6, line 10: To me it is not clear why the square-root transform is necessary, please explain.

The square-root transformation was necessary to give more weight to small values of the signatures (see page 6 lines 12–13 of the original m/s), i.e., to make the *D* values from catchments in different climatic zones to be more similar in magnitude and thus more intercomparable. Without the square-root transformation of, for example, the mean annual runoff (MAR) values, the corresponding average deviation (*D*) values would be dominated by tropical catchments, which tend to exhibit very high runoff amounts.

Page 6, line 10: Is Q1 (Q99) a very high or a very low value. In hydrology both definitions are used. Please specify Equation (1) and associated text: Why do you not use the observations to determine σo? To me this would be much more intuitive and would help to avoid the usage of another dataset which is prone to estimation uncertainty.

Table 3 specifies that Q1 and Q99 are *exceedance* percentiles related to peak and low flows, respectively. We have added "exceedance" also to the main text to avoid confusion.

Our reason for using the fully global signature maps from the Global Streamflow Characteristics Dataset (GSCD) is because it provides a much more representative global picture than a sparse, unevenly distributed set of observations would. Note that the GSCD dataset is completely observation-driven.

Table 3, line 1: Which P dataset was used? Is it the same that was used to drive the models or another one?

For calculating the RC we used *P* data from the WFDEI dataset, which has also been used to drive each of the models.

Tables 4 & 5: I do like the detailed information, but it would be much more accessible if it could be presented in figures (e.g. bar-plots)

We appreciate the suggestion, but the drawback of bar plots is that they make it difficult to deduce the actual values. On the other hand, the current table provides the actual values and allows readers to quickly interpret the results using the colors. However, if the editor feels that bar plots are more appropriate we can certainly make this change.

Figure 3: Colour scale for the density is missing.

Thank you for the comment, we have added a color scale.