Review of manuscript no.: hess-2016-546 by Sandra Pool et al: Streamflow characteristics from modelled runoff time series - importance of calibration criteria selection

This paper analyses the informational value of different streamflow characteristics (SFCs) used in calibration of a hydrological model, as alternative and/or supplement to traditional calibration criteria like Nash-Sutcliffe efficiency criterion etc. The motivating application is estimation of ecologically relevant SFCs by precipitation-runoff modelling, ultimately in ungauged catchments. The current paper, however, does not address the regionalisation aspects of this challenge (as is clearly stated on page 3, line 28). It is nevertheless an interesting paper bringing together traditions within classical hydrological model calibration and eco-hydrology.

The paper is well written and easy to perceive. Still, I do have some concerns about some aspects of the manuscript, in particular as focus shifts from calibration evaluation to estimation of SFCs. I will elaborate on this below.

The paper contains no clear definition or distinction between a goodness-of fit measure based on an SFC, and a 'traditional calibration criterion' (TCC). In my view, a comparison of SFCs to TCCs should acknowledge that there may be a transition zone between the two, but still provide a clear distinction which makes the comparison meaningful. SFCs are (citing): 'often used to refer to <such> specific aspects of the flow regime', however, this is also true for R_{eff}, R_{eff}(ln), slope of FDC etc. In addition, the analysis use R_{eff} alone as TCC benchmark, although the necessity of combining different TCCs in calibration is well recognised. This constructs a comparison in favour of SFCs.

A definition of 'traditional criterion' could be used, for instance along the lines of 'A goodness of fit measure computed from all sim-obs residuals within the calibration period, possibly transformed'. Then distinct groups of SFCs could be recognised by being 'selected from specific seasons or situations', or 'based on duration of events or conditions' etc.

With no rules for how SFCs are constructed, statements like 'High flow SFCs tend to be under-estimated' are meaningless. Flashiness index and concentration time both characterise high-flow behaviour, but underestimating one would mean overestimating the other. From table 1 the reader may verify that each SFC used in this paper is scaled so its value is positively related to flow magnitude for the relevant section of the FDC, but this should be asserted in the text when used in conclusions.

Most serious objection: The paper draws conclusions based on subjective evaluations of results. There is no reference to statistical significance or confidence intervals, and very few references to thresholds for 'good simulation', 'small error' etc. In section 2.4.1 on page 5, it is stated that an ensemble of 100 calibrated parameter sets were available for each objective function, allowing analysis of parameter uncertainty. I have not investigated the genetic algorithm used for calibration, but could the variability within this ensemble be used to assess which differences extend beyond mere noise?

Details and specifics:

Page 3, lines 22-24. This sentence is unclear and possibly erroneous. What is meant?

Page 3 line 29 and onwards. Please specify also what is **not** used. For instance in (1) make it clear that this SFC is used alone, in (2) whether or not the multi-SFC vector includes the SFC being evaluated, and (3) if the TCCs are kept in or excluded when the SFCs are included.

Page 4 line 13. The reference to NAVD 88 is unnecessary.

Page 6 line 10: Not yet knowing that the SFCs are normalised to a common scale, the reader may wonder how R_{eff} and an arbitrary SFC can be equally weighted. Just put in a 'normalised (see below)'.

Page 6 line 20: Can you specify how small error are required for a SFC to be robust, and how 'relatively good simulations for other SFC' are required for being informative? These limits have the impact of restricting which SFCs enter the Multi alternative.

Sections 3.1.1 and 3.1.2: Would these come more naturally in opposite order? I perceive the I_{single} experiment as the simplest and most obvious, the one-to-any SFC as more involved.

Would it be informative to summarise in a table for all SFCs what is illustrated in fig 3 for TA1?

Page 8 line 6. What is required to deserve a 'well simulated' mark? Is there an a priori defined threshold?

Page 8 line 8. Se discussion above about SFC estimates being high or low.

Section 3.3, page 5: This is one of the weaker parts. See the above point on subjective conclusions.

The categorisation in lines 19 and 20 are well defined, but then not used for anything. The lowest error class collects everything from perfect match to 10% error, capturing the result for all the mean-flow SFCs, 3 out of 4 low-flow SFCs and 4 out of 5 high-flow SFCs. Still, this paragraph states that all high-flow SFCs and three of four mid-flow SFCs are under-estimated, whereas all except one low-flow SFC were over-estimated' (lines 21-22). Such conclusions need to refer to at least a clearly stated threshold, but preferably to statistical significance.

The SFCs listed as having small vs medium absolute errors in line 23 does not correspond to a sorted grouping of the errors in fig 8. The FH6 error is characterised as 'small', but is larger than the DH16, MA41 and TA1 errors listed as medium, as well as the MA26 and FL2 errors not listed in these two lowest groups. Likewise it is difficult to see from figure 8 why MA26, DH13 and FH7 are identified in line 25 as having large error ranges, while MA41, TA1 or DH16 are not. One gets to suspect that the text is referring to another version of figure 8.

The statement in lines 15-17 suggests an identification between goodness of fit and process representation, which this paper neither investigates nor justifies.

Figure 9: The term 'absolute normalised SFC errors' are used. The figure seems to display signed errors.

Any calibration criterion as used in this experiment is a compression of the entire vector of simulation residuals into a scalar goodness-of-fit (GOF) measure. An SFC can provide a narrow-band, highly specialised GOF, whereas the traditional TCCs are 'broadband' GOFs aiming to minimise the expected error in any situation. I question the practical relevance of I_{single} calibration, but the idea that a general, multi-purpose GOF can be constructed from combining several specialised SFCs, in my opinion deserves investigation. With the mentioned weaknesses improved, this paper is a valuable piece in that puzzle. A continuation along this path should investigate possible conflicts between SFCs, and elaborate more thoroughly on uncertainty and identifiability.