Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2019-161-RC2, 2019

Hydrology and Earth System Sciences
Discussions

Interactive comment on "Calibration of hydrological models for ecologically-relevant streamflow predictions: a trade-off between performance and consistency" *by* Thibault Hallouin et al.

Anonymous Referee #2

Received and published: 11 June 2019

Review of the manuscript "Calibration of hydrological models for ecologicallyrelevant streamflow predictions: a trade-off between performance and consistency" by Hallouin et al.

In this paper, Hallouin et al. describe trade-offs in the choice and formulation of objective functions used to predict streamflow characteristics (SFCs) from rainfall-runoff models. Using simulations from 33 catchments in Ireland, the authors evaluated the overall performance, stability, and consistency of 6 objective functions, consisting of 3 "traditional" functions (variants of Kling-Gupta efficiency) and 3 "bespoke" functions

Printer-friendly version

Discussion paper

(based on SFCs identified in two sets of studies from the southeastern U.S. and Germany).

Overall, the paper is well-written, structured in a logical format, and addresses relevant questions for the task of calibrating process-based (e.g. rainfall-runoff) models to predict suites of multiple SFCs. However, I agree with reviewer #1 that a more in-depth assessment of the reasons for variability in objective function performance, stability, and consistency would improve the paper. Currently, SFCs are organized into "fish" and "invertebrate" categories, which I feel is potentially misleading (as described below). It could be more instructive, perhaps in future work, to organize SFCs instead based on the types of information they represent (e.g. timing and duration SFCs versus magnitude SFCs, or SFCs that describe extremes such as high-flows and low-flows versus those that describe central tendency). Such an organization could provide more useful information as to the processes by which certain groups of SFCs are better or more poorly predicted by different objective functions. Absent such a reorganization in this paper, the authors could still probe a little more deeply into the types of SFCs in each grouping to explore how/why the reported patterns arise.

General comments:

The second paragraph of the introduction is a bit generic in referring to "a range of streamflow characteristics". Some examples could help: low-flow versus high flow periods and their associated SFCs? Timing and duration statistics as opposed to magnitude statistics? It would help at this point in the introduction to give the reader a better sense of the types and categories of SFCs that are commonly used, and how and why different types of SFCs might be optimized using different functions. Different aspects of the flow regime are specifically named in methods (P4 L9-10), but would be better introduced and explained in the introduction.

The adjective "bespoke" may be unfamiliar to some audiences, e.g. in the USA. Is there another term that could be more globally well understood? Perhaps "customized"?

Also, can you provide a succinct definition of “behavioural” parameter and how parameters are identified as “behavioural”?

Fig. 2 caption indicates a substantial number of complete hydrological years were discarded from analysis (light blue) but this is not explained in methods. Why was this done? Was it so as to have an equal number of years of available data across catchments, and the minimum available was 14 years? Or was this to facilitate the split-sample tests? Were there trade-offs in this decision, i.e. loss of precision due to using less calibration or validation data than was actually available?

Throughout the text (e.g. P4 L8 and in Table 2), sets of SFCs are presented as belonging to “fish” or “invertebrates” when really they represent SFCs drawn from two sets of studiesâĂŤincluding Knight et al (2014) (fish) and Kakouei et al (2017) (invertebrates)âĂŤthat happened to use these taxonomic groups. As currently presented, this gives the impression that these sets of SFCs have empirically been demonstrated to be more important to fish than to inverts, or vice versa. Similarly, labeling the vectors and corresponding Euclidean distances as “Dfsh” and “Dinv” gives this impression. The cited studies from which these SFCs were derived were not focused on the relative importance of SFCs to one taxonomic group over another. To truly represent such broad taxonomic groups, studies across a wider set of geographic areas (not just the SE United States and Germany) and more diverse sets of species would be required. Regardless, this is not a goal of this paper. This could be presented differently, with the same SFCs and same analysis, in a way that would not give false impressions about overall taxonomic relevance for particular groups of SFCs.

There are some problems of disagreement between results text and results figures. For example, the results text in section 4.2 does not match the results shown in Fig. 7 (see specific comments below). Please ensure correct labeling of figure panels and revise text or figure as needed to ensure agreement. There is also disagreement between results text in section 4.4. and results shown in Fig. 9 (see specific comments below).

Printer-friendly version

Discussion paper

In the results as presented in section 4.3 and Fig. 8, stability of performance across the split-sample tests is apparently assessed only visually. Even for this study with a relatively small number of catchments (33) this becomes a bit unwieldy (and the reader is forced to search through 33 separate plots in each of 3 figure panels to find the examples mentioned in the text). If these methods were to be reproduced or adapted for another study area with a larger number of catchments (in the hundreds, as is common), then a figure such as Fig. 8âĂŤupon which the conclusions about stability relyâĂŤwould become untenable. This paper would benefit from a more quantitative assessment of stability, presented perhaps as means and standard deviations across catchments instead of Fig. 8. Such a quantitative stability metric would (a) be much easier to understand and interpret, while presumably yielding the same results, namely that performance stability was not markedly different between traditional and bespoke objective functions, and (b) would be more repeatable and transferable to other studies in other regions, especially using large numbers of catchments.

In conclusions, I'm not sure that hypothesis 3 is strongly supported by results in Fig 9 (see specific comments below).

Table 2 is cited before Table 1 (section 2.1 and 2.3, respectively). Likewise, Fig. 1 is cited after Figs 2 and 3. Please ensure all tables and figures are cited in proper order.

Specific comments:

P2 L5: "The prediction of streamflow conditions" [need to add or specify "at ungauged locations"]

P2 L7 "hydrological models that produce streamflow hydrographs" [may want to specify these are simulated hydrographs as opposed to observed hydrographs]

P2 L15 "across a range of streamflow characteristics"... does this particularly involve high-flow vs low-flow periods?

P2 L16-17: sentence fragment. Delete "although"?

P3 L9: is there a missing comma? “captured in the effective parameter values of the model, could be compromised. . .”

P3 L12-12: “. . .relating to its own preferences for living conditions” this was also demonstrated by Knight et al (2014).

P3 L13-14: “when several species are considered simultaneously, the number of SFCs to predict will increase accordingly”. . . yes, in general, but that depends on the habitat requirements of the species in question. Species with similar behavior (eg foraging strategies), reproductive timing, and physical (eg thermal) niches may actually share a common set of SFCs as being most relevant to them.

P4 L3-7: Any evidence that these SFCs are ecologically relevant in the Irish catchments?

P4 L8: “Only two hydrological indices are common between the two communities’ respective streamflow preferences”. I don’t thinkâĂŤbased solely on the cited studies in the Southeastern US and GermanyâĂŤthat such sweeping statements can be made about streamflow preferences across such broad taxonomic groups (fish and invertebrates). More accurately, this is a comparison between two sets of studies on two different continents that happened to use different taxonomic groups in their analysis. It does not support conclusions about which SFCs matter most to which taxonomic groups outside the respective study areas of the cited studies. I am familiar with the research in the Southeastern US (Knight et al 2014) which did not use any invertebrate data at all, and so should not be used to suggest that certain SFCs are more (or less) important to fish relative to inverts in that study region.

P4 L13: “calculations were vectorised”. . . can you explain a bit more what this means? I’m familiar with EflowStats but not Python. Perhaps Python users know exactly what you mean, but a brief explanation could help readers with less familiarity. In Table 2, the column header “target species” (fish or invertebrates) might give a casual reader the impression that these SFCs were empirically deemed important in the study-area

catchments. Rather, they were gleaned from Knight et al (2014) (fish) or from Kakouei et al (2017) (invertebrates). More importantly, it could give a false impression that these SFCs are generally more important to the taxonomic group listed which is not necessarily the case based on the previous studies cited. This could be clarified in the table caption or the accompanying text, or by changing "target species" to "citation" and listing the corresponding paper from which the SFC was obtained. Also in Table 2, table caption says "SFC" but column header says "indicator". Would be better to use one term consistently for clarity. This goes for the main text as well, where "SFCs" and "hydrologic indices" are both used.

P4 L20: "in order to have at least seven years for calibration and seven years for evaluation. . ." Can you justify (perhaps with citation) why seven years is an acceptable POR for calibration/validation?

P4 L27-28: "hence representing a good sample of the diversity of Irish soils and geology" Except that from Fig 3 it appears that higher-elevation or mountainous catchments were not well represented. . .? Fig 1 and Fig S1 are essentially redundant to each other, and Table 1 and Table S3 are redundant, except for slight differences in parameter presentation. Parameter representations in Table S3 appear to match Fig S1 but not Fig 1. Suggest retaining only one version of this figure/table combination, or if repeating in the supplement then ensure consistent parameter representations throughout.

P5 L14 and L19: How are energy-limited and water-limited periods defined?

P5 L27: What formulation of ET is this? E.g. Penman-Monteith?

P6 L23: do these KGE variants really encompass "the entirety of the observed and simulated hydrographs"? It seems the first KGE emphasizes high flows, the second (inverted) emphasizes low flows, and the third "equally consider[s] high flow and low flow conditions". This suggests that these 3 KGE variants collectively emphasize high and low flows. . . is this at the expense of moderate flows?

Printer-friendly version

Discussion paper

P8 L20-25: This section reads more like methods than results. It would be helpful to begin the results section by presenting an overview of the most compelling findings. In Figure 6 caption, specify that Ehi is Kling-Gupta efficiency

P9 L6-7: “indicating that these combinations of SFCs are good candidates for general purpose hydrological studies. . .” True for these 33 watersheds, not necessarily in other locations. Specify that this conclusion is only for the study catchments. Also, while you’re at it, you might point out that since all the calibration and evaluation scores were fairly high, this suggests that any of these objective functions (possibly except Elo) could do a reasonable job of predicting the overall hydrograph for these study catchments.

P9 L11-13: I don’t see this in fig. 7. In (a) for Dfsh, it appears that Dinv produces the lowest Euclidean distance. In (b) for Dinv, it appears that Dfsh and Dall are nearly tied for the lowest Euclidean distance. In (c), yes it does appear that Dall has the lowest Euclidean distance for predicting for Dall.

P9 L16-17: Again, I don’t see this in fig. 7. The two largest combinations of SFCs are Dfsh in (a) and Dall in (c). As stated, Ehi performs best for Dall in (c), but for Dfsh in (a), it appears that Eav performs better than Ehi. For the smallest combination of SFCs, Dinv, fig. 7 shows that Ehi performs best but the text states that Eav performs best. Please check that all panels of fig. 7 are labelled correctly and revise the text (or figure if needed) so that the figure and text are in agreement.

P9 L21 “measured by means of the standard deviation” would be better stated “measured by the standard deviation” so as not to confuse two different uses of the word “means”.

P9 L22-23: “in addition to be predict well” Fix grammar.

P9 L32: There is no catchment 24003 presented in Fig 8a. Please correct.

P10 L13, definition of consistency: I found this definition a bit confusing and wonder if it

could be rephrased. Typically a ratio involves two variables to be compared, presented as “the ratio of X to Y”. From the Fig. 9 caption it appears that values have been averaged across the 14 split-sample tests but it is unclear to me (and could be unclear to some readers) what the ratio represents. Ratio of “behavioral parameter sets that remain behavioural”... to what?

P10 L20-21: Text does not match fig. 9. Text states that consistency for Dfsh is 13%, whereas fig. 9 lists that value for Dinv and lists consistency for Dfsh as 0.309.

P10 L22: This does not seem “remarkable”, and perhaps not even meaningful, given that there are only 3 data points (only 3 objective functions). With only 3 to compare, the association between greater consistency and greater numbers of SFCs could be attributable to chance. This also pertains to the conclusion you draw in discussion, P11 L9-10.

P10 L28: “model structure is not adequate for these catchments”... any ideas as to why? Do these catchments share any common attributes?

P11 L6: They’re not really “three different sets of SFCs” because Dall contains the same SFCs as Dfsh and Dinv combined.

P11 L12: Arguably, SFCs can be hydrologically relevant (important) without necessarily being hydrologically representative or indicative “signatures” of the overall hydrograph.

P11 L19-20: “because they may not be key descriptors of the emergent hydrological processes at the catchment scale”... how does this square with results in Fig 6, in which you showed that performance of Dall and Dfsh was nearly equivalent to that of Ehi?

P12 L6-7: This doesn’t fully fit with your findings, as you did not find that consistency was uniformly better for traditional objective functions than bespoke functions. Rather, Ehi had good consistency, whereas Elo had consistency poorer than two of the bespoke functions, and Eav had consistency that was roughly equivalent to two of the

bespoke functions. From your results (Fig 9) there is considerable variability in consistency values across traditional functions and also across bespoke functions. This also applies to your conclusions section P13 L19-20: “traditional objective functions select more consistently the same parameter sets as behavioural across the split-sample tests. . .” this is definitely true for Ehi (much greater consistency than all 3 bespoke functions), but not for Elo (lower consistency than 2/3 of bespoke functions), and is only marginally true for Eav (roughly equivalent to 2/3 bespoke functions, and more consistent than the third). My understanding of your results in Fig 9 are that they produce a somewhat mixed picture, with only (at best) weak support for your third hypothesis.

P12 L20-22: Can you clarify this section a bit? “may not be realistic for practical applications” [why not?] . . . “might not be as critical for the stream ecology” [wouldn’t that need to be determined empirically?] It’s not clear to me what you’re meaning to say here.

P12 L24-26: “Unless, . . . both at once” Remove comma after “unless” and specify that “both” means both “ecologically relevant” and “hydrologically relevant” (although I dislike the latter term as explained above).

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2019-161, 2019.

Printer-friendly version

Discussion paper