Review 2 of “Technical note: Diagnostic efficiency – specific evaluation of model performance” by Schwemmle et al

Summary

I believe the authors have responded well to most review comments. They have made significant changes to the manuscript. The metric has been rewritten in the form DE = X, which in my opinion is a lot less vulnerable to misinterpretation, and the connection between error values and error sources, which in my and other reviewers’ opinions was somewhat speculative, has been removed.

There are a few points that I think deserve a bit more attention. I have added my own responses to those provided by the authors in a new document and uploaded that as a reviewer attachment. Line numbers refer to those in the track-changes manuscript. Author comments are kept in blue. Most importantly, I think the manuscript would be strengthened if the authors would add a cautionary discussion note about potential pitfalls in interpreting the mismatch between a simulated and observed flow duration curve and the associated $Bre\dot{t}$ value. More explanation below.

I hadn’t mentioned this in the last review, but making the code to create the polar plots publicly available is a great idea and very helpful.

Kind regards,

Wouter Knoben

General comments

< majority of text removed for brevity >

For example, in case the high flow part of the simulated FDC is greater than the high flow part of the observed FDC indicates that the simulations overestimate the peak flows. < …>

I don’t think this is quite correct. To me, stating that simulations overestimate peak flows implies that the simulations give us peaks at the same time as the observations have peaks, but that the simulations are too high. I would say that because the observed and simulated hydrographs are sorted independently and contain no temporal information, all we can say from such a case is that the distribution of simulated flows contains higher flows than observed but we cannot conclude when these model failures occur.

Two helpful thought experiments may be (1) to take any hydrograph and create synthetic simulations by increasing the low flows of this hydrograph until they are higher than the highest observed flows, and (2) to take a strongly seasonal hydrograph and create synthetic simulations by shifting this hydrograph in time until the “simulations” overestimate the low flows and underestimate the peaks. In the first case, the simulated FDC has higher flows than the observed one, but stating that the simulations overestimate the peaks would be incorrect (they overestimate the low flows). In the second case, the model overestimates the lows and underestimates the peaks but the FDCs for both hydrographs are identical, meaning that there is no clear 1:1 connection between model error and FDC mismatch and hence care is needed to interpret FDC (mis)match. Hopefully this clarifies why I think that conclusions such as the one in blue above cannot be derived from comparing two FDCs. I would strongly encourage the authors to reflect this

line of reasoning in their manuscript and update the text where appropriate. See also comment about L94 below.

Specific comments

L82. How would this equation deal with catchments where the observations drop to zero, but the simulations do not? This indicates a model error that should show during evaluation but equation 3 will break in such a scenario.

We are fully aware about this shortcoming. Therefore, the metric is only valid for catchments with perennial streamflow (see line 268).

It may be good to make this mention slightly more prominent, perhaps by moving it directly below Eq. 3. In my experience discussion items get overlooked more easily and this is a fairly critical aspect of the proposed metric (i.e. it being only applicable to regions with perennial flow) that deserves visibility.

L94. The conceptual novelty of DE seems to be that it uses observed and simulated FDC's for two of its three metrics. For $\overline{Bret}$ and $|Barea|$, the series of observed and simulated Q values are thus not used as a time series but ordered into a flow duration curve. A major consequence of this is that the temporal connection between observations and simulations is mostly lost, because the two series are not compared on a per-time step basis. With the data ordered as FDCs, it's no longer clear at which point in the simulation certain errors were generated and thus where model deficiencies may be found. An extreme example would be a case where a model matches all observations perfectly, but for some reason returns zero flow on those time steps where the observations are highest. When the simulations are shown as a FDC, those zero flows will have exceedance probabilities of 100% and thus we might think that the model underestimates the low flows, even though the model in fact simulates the low flows just fine but massively underestimates the high flows. It is therefore unclear to me why quantifying the errors between both flow duration curves leads to increased understanding of model errors. From my point of view, it seems equally possible that the temporal disconnect between observations and simulations will mask certain model errors instead. I realize that neither NSE nor KGE would be of much use in this example either, but the almost complete temporal disconnect of the simulations and observations in DE is worthy of discussion. It could also be helpful to define a few more extreme cases of model errors and see to what extent DE can be used to trace those errors.

Again the metric is only valid for catchments with perennial streamflow (see line 268). Of course, comparing the observed FDC and simulated FDC disconnects the time steps. However, the timing error term is not related to the FDC. We used Pearson's correlation coefficient (see Eq. 6) to compare the simulated time series and observed time series on per-time step basis. We want to emphasize that *DE* is not the perfect metric. Instead *DE* represents an alternative tool which can be used in addition for model evaluation.

I agree with this response that the timing error is captured as part of the Pearson correlation. What I hoped but failed to accurately convey with this comment is to urge caution with statements that do assign a temporal component to conclusions based on just comparing both FDCs. See also the general comment above. This is important in e.g. lines 255-257: "All simulations have in common, that positive dynamic error type (i.e. high flows are underestimated and low flows are overestimated) dominates accompanied by a slight positive constant error. Timing contributes least to the overall error." It is very clear in Figure 4 (FDC column) that all three models simulate a narrower range of flows than observed, by simulating lower high flows and higher low flows. Equally in Figure 4 (hydrograph column) it can be seen that, if we arbitrarily classify the peak between months 4 and 8 as "high flows" and the remainder as "low flows", the model both underestimates (months 2-4) and overestimates (months 8-1) the low flows, instead of only overestimating the low flows as the text indicates. This particular aspect of model failure cannot be deduced by just looking at the FDC and $\overline{Bret}$ values.

I would again strongly encourage the authors to add a discussion paragraph where this issue is discussed. I think that the need to be very careful about interpreting FDC and $\overline{Bret}$ results is something that can easily be missed if the manuscript does not devote sufficient attention to it.

L102. I tried implementing equations 2, 3 and 5 with real data but could not reproduce 50% of the integral being positive values and 50% being negative. Instead, my Bres plot alternates between being negative and positive and only ~45% of its values are negative (see figure). I have included my code below. Assuming that I didn't make any mistakes, can the authors clarify whether the equations and assumptions in the manuscript are correct?

The equations are correct. We do not assume that 50% of the be either entirely positive or entirely negative. In your case (i.e. ~45% of its values are negative) it means the left part of the FDC is mostly underestimated and the right part of the FDC is mostly overestimated.

I'm glad I implemented the equations correctly. I based this comment on line 111: "Since we removed the constant error (see Eq. 5), the left half of the integral is positive and the right half (i.e. 50$_{th}$ percentile to 100$_{th}$ percentile) will, thus, be negative and vice versa if the left half of the integral is negative." I suggest to update/remove this sentence.

L120. It's not entirely clear to me why |Brel_bar| has this specific threshold at 1.

The threshold refers to the letter *I*.

My mistake, thanks for clarifying.

Editorial

L122. "we introduce certain threshold" > "we introduce a certain threshold"