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

Summary

The authors introduce a new efficiency metric called Diagnostic Efficiency (DE), which replaces the bias and variability components of the Kling-Gupta efficiency with two statistical measures derived from the observed and simulated flow duration curve. The goal is to improve the connection between efficiency metric score and specific errors in the modelling setup, so that diagnosing model deficiencies becomes easier. Several synthetic test cases are shown to describe the metric’s functionality and the metric is also used on a real-world test case.

I have read this paper with interest and I think that metrics that help diagnose model failures are a relevant area of research. I do think that this paper needs some improvements before it can published. In my opinion, the benefits of DE are currently a bit overstated and there are some methodological choices that would benefit from a clearer explanation. In particular, (1) I do not think that the interpretation of DE’s error types as being caused by specific types of model deficiencies is currently well-supported, and (2) I do think that DE needs to be used in combination with some form of hydrologically meaningful benchmark/threshold if it is to be used to determine if model simulations are deficient or not. Regarding the methodology, I have some questions about (1) how the constant error term can be interpreted in cases where the simulated FDC is both above and below the observed one; (2) how the calculation of Bdir works, and (3) how deviations between two FDCs can be used to trace model deficiencies. More details are in the comments below.

Kind regards,

Wouter Knoben

Comments

l11. “Input data”; it may be more accurate to rephrase this as “data uncertainty” because errors in the evaluation data could equally lead to unsatisfactory model performance although in that case it may be that what we consider as the “truth” is faulty and not the simulations.

l31. “value close to one indicates a better model performance”. Given the nature of the paper, it would be good to define what is meant by “better model performance” and similar words and phrases. Some readers may interpret this as meaning that the model is an appropriate representation of the catchment in question (high model “fidelity”), but, as the authors indicate, a NSE or KGE score of 1 only indicates a perfect numerical match between observations and simulations (high model accuracy, but not necessarily for the right reasons). The metrics themselves do not provide any interpretations about how well the model simulations represent real-world hydrology and it would be good to be explicit about this.

L59. Do input data errors and observation uncertainty (l60) not fall under observations with insufficient accuracy mentioned in line 54?

L62. It's not immediately obvious to me why this paper addresses three out of the 5 error sources mentioned in lines 57-61. Can it be clarified why these three errors are the focus of this work?

L64. Using these three error terms seems the core assumption of this paper. The provided examples help in understanding what they mean but I think formal definitions of each error term should be included here.

L64. I would expect some form of justification to support the choice of these three types of errors. Are they sufficient to describe all possible deviations from observations that simulations could show?

L71. I expect the authors chose an equation of the form DE = 1-X to match the way NSE and KGE are formulated, but I think this makes the metric vulnerable to wrong interpretation. NSE is a skill score, where any simulations with NSE > 0 can be said to have outperformed the mean flow benchmark model included in the NSE equation. KGE has no such benchmark, but due to its formulation as KGE = 1-Y it is an easy mistake to assume that KGE = 0 has some distinct meaning even though it does not. DE does not seem to be a skill score either and I don't immediately see that DE = 0 has any special meaning. I would strongly recommend to reformulate the metric as DE = X, so that DE = 0 can be cleanly interpreted as "there are zero errors".

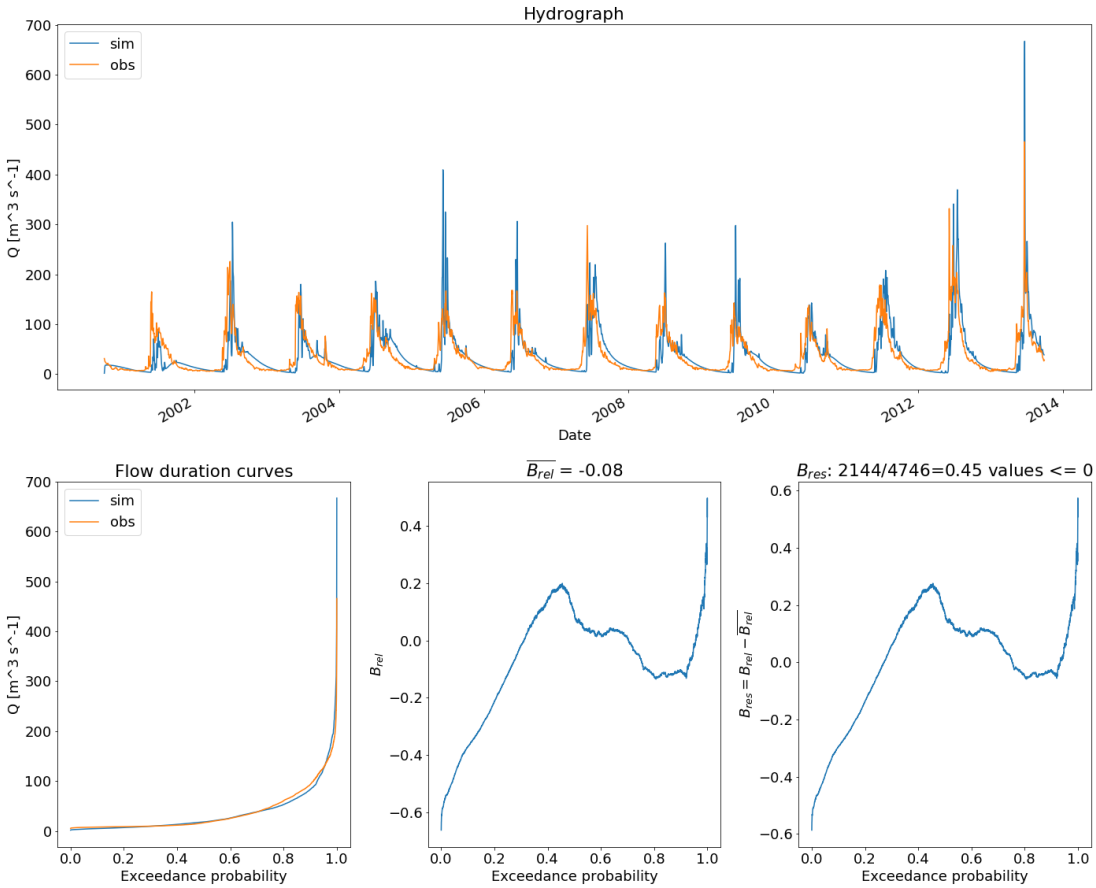
L77. It's not fully clear to me why $\vec{Bret}$ indicates a constant error. What happens in cases where the simulated FDC both overestimates and underestimates the observed one? $\vec{Bret}$ will likely still fall on one side of zero but that does not mean that all simulations showcase a constant error of $\vec{Bret}$.

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.

L94. The conceptual novelty of DE seems to be that it uses observed and simulated FDC's for two of its three metrics. For $\vec{Bret}$ 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-timestep 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 timesteps 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.

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?



```
# sort hydrographs to create FDC (scaling can be done in plot)
qsim = np.asarray(sorted(sim))
qobs = np.asarray(sorted(obs))

# calculate B variables
brel = (qsim-qobs)/qobs # Eq. 3
brel_bar = np.mean(brel) # Eq. 2
bres = brel - brel_bar # Eq. 5
```

L106. Why is this referred to as a slope? This equation seems to only change |Barea| back into Barea through a somewhat roundabout way. "Slope" implies some value with units [distance/distance].

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

L126. I understand that section 2.2 tries to outline different scenarios for Brel_bar, Bslope, DE and Del but I find this quite difficult to follow. I'm struggling to follow the reasoning that leads to equations 12-14 and feel a bit lost with all these variables that I'm seeing for the first time. Maybe a longer

explanation, or a graphical example, or placing the scenarios in a table or even a flowchart could help to clarify this section.

L166. “Note that the original temporal order is maintained.” Is this correct? The FDC contains no temporal information. Should the mention of “FDC” on line 164 be “time series” instead? Also on line 168.

L191. “Interdependently… regions.” I don’t understand what this sentence means.

L202. “Numerically, … NSE.” I suggest to remove this sentence. The fact that DE scores are higher than NSE and KGE scores is irrelevant (there is no reason why these scores can or should be compared in a relative sense) and referring to this as “better performance” may be confusing to readers who associate “better [model] performance” with “more accurate representations of real-world hydrology”.

L207. “For example, lowest KGE values … (Table 2a-d).” I’m not sure how to interpret this sentence. Can this be clarified?

L237-243. I find this attribution of causes to certain error types very speculative. For example, underestimation of high flows and overestimation of low flows could equally indicate that precipitation input is smeared out over time, which tends to happen with gridded forcing products interpolated from station data, or with climate models that have a tendency to drizzle. Equally, a constant positive error (overestimation of flows) may indicate a model structure issue such as an inappropriate evaporation routine (not enough water returns to the atmosphere) or a “impervious runoff” routine that allows part of the incoming precipitation to bypass the soil moisture routine entirely or a “subsurface water exchange process” that imports water from an underlying aquifer. Parameter issues could also play a role here, for example if soil moisture storage capacity is set too low and part of the incoming precipitation directly goes into streamflow as saturation excess runoff, or if evaporation is limited by some form of inappropriately set wilting point. I suggest to either remove this section or better support why certain types of errors must (or are at least most likely to) be generated from the causes described here.

L261. I think point (iii) is a somewhat optimistic view. The link between error type and associated model deficiencies is a bit tenuous in the current manuscript (see previous comment) and needs to be better supported before this can be presented as a feature provided by DE.

L269. DE may not use a benchmark simulation, but it does have the same issue that it is difficult to say which DE scores indicate that a model is “good enough”. The authors have not justified their use of a 5% deviation threshold on each of the DE components, which I assume was chosen for illustrative purposes only. Therefore, DE has the same interpretation challenges as KGE (which is also a deviation-from-perfect kind of metric) and the recommendations for KGE should apply to DE too. I suggest to clarify this in the text.

L291. “A proof of concept and the application to a modelling example showed that errors coming from input data, model parameters and model structure can be unravelled with the help of expert knowledge or a statistical analysis. Particularly, diagnostic polar plots facilitate interpretation of model evaluation results. These plots may advance model development and application.” This seems to be mostly speculation in the current manuscript (see comment about lines 237-243). I suggest to either improve the support for this statement or remove it from the conclusions.

L294. “We tried to base the formulation of the newly introduced diagnostic efficiency on a general hydrological understanding and can thus be interpreted as deviation-from-perfect, we do not need to define benchmarks.” This seems a bit optimistic. DE cannot answer the question “is my model good enough” without a statement about the level of deviation-from-perfect that is considered acceptable for a given purpose. Justifying where this level is set is functionally equivalent to specifying a benchmark. I suggest to remove the last part of this sentence.

L319. I appreciate this stepping stone to more work on efficiency metrics but the provided equations seem a bit trivial and in the case of A1 perhaps even overly specific. There is no real need for future metrics to be of the shape De = 1 – X (arguably, DE = X would lead to less ambiguity) nor do such metrics need to have three components and not two or four or some other number. I expect that this appendix can be removed without harming the main manuscript.

Editorial

l29. Replace “can be measured by only” with “with”.

l34. “satisfying” > “satisfactory”

l160, l162. “requires” > “required”?