Response to Reviewer #1

We would like to thank the anonymous referee for his/her interest and the comments on our manuscript. Below, reviewer comments are in italic font and our replies are in plain blue font.

General comments

The manuscript presents a potential useful new metric and method for evaluating hydrologic model inadequacy, which is definitely in the journal scope. The technical note has been somewhat improved by the revisions, and most comments have been addressed to some degree. The methods are valid and fully outlined, the results are adequately discussed and sufficient to support the conclusions. The overall structure of the technical note is fine, but there are some issues with the text and figures which detract from the comprehensibility of the paper. Some minor revisions should be completed prior to publication.

We thank the reviewer for his/her helpful comments.

Specific comments

53-62: This section is not clear. The text states that observations should be checked for accuracy, but then uncertainties in observations remain, therefore we will not focus on uncertainties in observations. This is not helpful in clarifying why some sources of uncertainty were the focus of the new metric.

We rephrased the section and replaced observations with evaluation data.

65-70: This may be a minor point, but what good is it three defined error types to be linked to potential error sources when, in your example, any of the three types can be linked to the same source (precipitation input)? If expert knowledge is need, how is the novel metric helping?

The novel metric shows which of three error types is dominating. The expert can then define which error may dominate and provide measure to either reduce the error or at least better consider the effect of the uncertainty.

98: Polar plots are not very much like clocks, given both angle and distance from the origin matter in polar plots. This comparison is more confusing than helpful.

We agree and removed the comparison.

Figure 1: The y-axis has no title.

We added a title to the y-axis.

Figure 2: The y-axis has no title.

We added a title to the y-axis.

Figure 4: Why are the FDC y-axes in log space? It initially appears as though the simulation only over estimates low flows, when the magnitude of high flow under estimation is in fact similar.

Using a linear y-axis the differences for the low flows would have been barely visible, therefore we decided to use log discharge axis. We are aware that this emphasis the low flow part more, but there is unfortunately no really good approach in comparing high flow and low flow with a similar scale.

Figure 5: One benefit of the novel metric and method which may not be highlighted in the manuscript is its ability to facilitate simulation comparisons. While the FDC curves contain more information than the polar plot, a single

diagnostic polar plot can clearly summarize the errors in dozens or hundreds of simulations (which would be unreadable in a single FDC plot).

We included a sentence which highlights this benefit. Thanks for pointing this out, this was clearly in our mind, but sometimes things get lost at the end.

Technical corrections:

30: 'Measure' twice

191: 'Interdependently' is either being used incorrectly, or was not the word intended

256: 'Illogical' is not a good descriptor for the quirks and irregularities of the NSE and KGE

We included all technical corrections in the manuscript.

The authors may also wish to review the erratic comma usage, although this is not impeding the manuscript's comprehensibility.

We reviewed the erratic comma usage.

Response to Reviewer #3

We would like to thank Wouter Knoben for his interest and the comments on our manuscript. Below, reviewer comments are in italic font and our replies are in plain blue font. We would also like to point out, that some remarks of this new reviewer are more difficult to approach, since changes we did in the first round of this review process were not known or considered or may have even be in a different direction as the reviews in the first round.

General comments

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.

We thank the reviewer for his helpful comments. The provided support on the causes of the specific error types may serve as a general advice how the errors might be linked to their source. This is more thought as an example, certainly not as a strong linkage to specific model deficiencies. We want to emphasize that our goal is not to replace current metrics. Moreover, *DE* is not designed to answer the question whether simulations are "good enough". This is in our opinion a philosophical question and many others have already tried to answer this question. But this is a problem for all metrics except there is a perfect agreement. Instead, we aim to provide a new alternative metric which focuses on the error identification. Moreover, we aim to facilitate the comparison between multiple simulations.

The constant error term represents the average of the relative bias distribution. Positive values for the constant error term can be always interpreted as overestimation on average whereas negative values can be interpreted as underestimation on average. B_{dir} is the integral of residual error (i.e. relative bias of the two FDCs after subtracting the constant error (see Eq. 5)) that reaches from 0^{th} percentile to the 50^{th} percentile. The deviation between the observed FDC and the simulated FDC contains some hydrologic information. 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. B_{dir} is required to calculate B_{slope}.

Specific 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.

We assume that the evaluation data represents the truth. We do not want to question whether the evaluation data is truly representative or not. Instead, we presume that the data quality of the evaluation data is checked beforehand. Answering the question about the "truth" is partly a philosophical one and cannot be answered by this technical note.

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.

We agree and include the term accuracy to describe the model performance.

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

With observations we mean the data used for the evaluation. We rephrase this line.

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?

In general, metrics are based on a comparison between observations and simulations. Consequently, metrics do not have the purpose to quantify the uncertainty of the evaluation data. Accounting for errors caused by initial/boundary conditions require a more complex approach. The approach we presented here is not suitable to quantify errors caused by initial and/or boundary conditions.

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.

We added some formal definitions. And we would also like to point out, that the core of the paper is the new metric and its potential advantages to other metrics, not the error terms.

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?

These three error types are common errors generated by many hydrological models. The three error types might not be sufficient to describe all possible deviations. In order to account for a wider range of deviations, we recommend to use different error terms or choosing a multi-criteria approach (see line 277f).

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".

We agree on using an error score and a perfect fit should indicate a value of 0. The idea of using $DE = 1-Y$ was to make it comparable with KGE. So it may also be discussed, why KGE was accepted the way it is. But we see the point and change DE from $DE = 1-Y$ to $DE = X$. 0 means zero errors.

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

B_{rel} represents the average of the relative bias distribution. Positive values for the constant error term can be always interpreted as an average overestimation whereas negative values can be always interpreted as an average underestimation. Overestimation and underestimation of the simulated FDC are captured by B_{area}. B_{rel} describes only the average behaviour.

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).

L94. The conceptual novelty of DE seems to be that it uses observed and simulated FDC's for two of its three metrics. For $Bre\bar{l}$ 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.

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.

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].

We use the term slope which is referred to the inclination of the dynamic error. For example, Yilmaz et al. (2008) used a similar terminology. In case B_{slope} is positive low flows are underestimated and high flows are overestimated (i.e. left integral of B_{res} is negative and right integral B_{res} is positive). Vice versa, if B_{slope} is negative low flows are overestimated and high flows are underestimated (i.e. left integral of B_{res} is positive and right integral B_{res} is negative).

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

The threshold refers to the letter *l*.

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.

There is only one new variable which is DE_l. The equations 12-14 represent the conditions for which a diagnosis can be made.

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.

The sentence was corrected. It should be time series.

L191. "Interdependently... regions." I don't understand what this sentence means.
We rephrased the sentence.

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".
We removed the sentence.

L207. "For example, lowest KGE values ... (Table 2a-d)." I'm not sure how to interpret this sentence. Can this be clarified?
We rephrased the sentence. The point we want to make here is that *KGE* and *NSE* are differently sensitive to the generated errors.

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.
We removed this pargraph.

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.
We agree and rephrased point (iii).

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.

We have chosen the 5% deviation mainly for illustrative purposes. The threshold can be set to any value. This is more about whether it is worth to diagnose the error and not about whether a model is "good enough". In comparison to *KGE*, the deviation is measured by a percentage of error.

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.

We rephrased the sentence and removed the speculation.

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.

We removed the last part of the sentence. We want to emphasize that our goal is not to answer "Is my model good enough?". The goal is to facilitate the comparison of multiple simulations and to easily identify dominant error types. In addition, we provide a graphical tool (i.e. diagnostic polar plots).

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.

We appreciate this comment, but we decided to keep the Appendix in the manuscript.

Editorial

l29. Replace "can be measured by only" with "with".

l34. "satisfying" > "satisfactory"

l160, l162. "requires" > "required"?

We included all technical corrections in the manuscript.

List of all relevant changes

- We added formal definitions for the error terms
- We changed the formula of *DE* from *DE=1-Y* to *DE=X*
- We further strengthened the difference between *DE* and *KGE*

Technical note: Diagnostic efficiency – specific evaluation of model performance

Robin Schwemmle[1], Dominic Demand[1], Markus Weiler[1]

[1]University of Freiburg, Faculty of Environment and Natural Resources, Chair of Hydrology, Freiburg, Germany

Correspondence to: Robin Schwemmle (robin.schwemmle@hydrology.uni-freiburg.de)

Abstract. A better understanding of the reasons why hydrological model performance is unsatisfying represents a crucial part of meaningful model evaluation. However, current evaluation efforts are mostly based on aggregated efficiency measures such as Kling-Gupta Efficiency (*KGE*) or Nash-Sutcliffe Efficiency (*NSE*). These aggregated measures provide a relative gradation of model performance. Especially in the case of a weak model performance it is important to identify the different errors which may have caused such unsatisfactory predictions. These errors may originate from the model parameters, the model structure, and/or the input data. In order to provide more insight, we define three types of errors which may be related to their source: constant error (e.g. caused by consistent input data error such as precipitation), dynamic error (e.g. structural model errors such as a deficient storage routine) and timing error (e.g. caused by input data errors or deficient model routines/parameters). Based on these types of errors, we propose the novel Diagnostic Efficiency (*DE*) measure, which accounts for these three error types. The disaggregation of *DE* into its three metric terms can be visualized in a plain radial space using diagnostic polar plots. A major advantage of this visualization technique is that error contributions can be clearly differentiated. In order to provide a proof of concept, we first generated time series artificially with the three different error types (i.e. simulations are surrogated by manipulating observations). By computing *DE* and the related diagnostic polar plots for the reproduced errors, we could then supply evidence for the concept. Finally, we tested the applicability of our approach for a modelling example. For a particular catchment, we compared streamflow simulations realized with different parameter sets to the observed streamflow. For this modelling example, the diagnostic polar plot suggests, that dynamic errors explain the ~~model performance~~overall error to a large extent. The proposed evaluation approach provides a diagnostic tool for model developers and model users and the diagnostic polar plot facilitates interpretation of the proposed performance measure as well as a relative gradation of model performance similar to the well-established efficiency measures in hydrology.

1 Introduction

Performance metrics quantify hydrological model performance. They are employed for calibration and evaluation purposes. For these purposes, the Nash-Sutcliffe efficiency (*NSE*; Nash and Sutcliffe, 1970) and the Kling-Gupta efficiency (*KGE*; Gupta et al., 2009) are two commonly used performance metrics in hydrology (e.g. Newman et al., 2017;Towner et al., 2019). *NSE* and *KGE* measure the overall model performance ~~can be measured by~~with only a single numerical value within the range of minus infinity and one. A value close to one indicates a better model ~~performance~~accuracy, whereas with increasing distance to one the model ~~performance~~accuracy deteriorates. From this point of view, the model performance can only be assessed in terms of a relative gradation. However, cases of a weaker model performance immediately lead to the following questions: Why is my model performance not ~~satisfying~~satisfactory? What could improve the model performance?

In order to answer such questions, Gupta et al. (2008) proposed an evaluation approach that includes diagnostic information. Such a diagnostic approach requires appropriate information. Considering only the overall metric values of *NSE* and *KGE* may not provide any further insights. Additionally, an in-depth analysis of *KGE* metric terms may provide more information on the causes of the model error (e.g. Towner et al., 2019). Although including the *KGE* metric terms may enrich model evaluation, due to their statistical nature the link to hydrological process is less clear. Current diagnostic approaches are either based on entropy-based measures (Pechlivanidis et al., 2010) or on process-based signatures (Yilmaz et al., 2008;Shafii et al., 2017). The latter one improves measuring the realism of hydrological processes by capturing them in hydrological signatures. These signatures represent a main element of a powerful diagnostic approach (Gupta et al., 2008).

Although the numerical value of the overall model performance is diagnostically not meaningful, the overall model performance determines whether diagnostic information will be valuable to the modeller or not. Diagnostic information may only be useful if the overall model performance does not fulfil the modeller's requirements. It will then be cumbersome to select the appropriate signatures or measures which may answer the modeller's questions about the causes. Visualising evaluation results in a comprehensive way poses another challenge for diagnostically meaningful interpretation. Therefore, we see a high potential in compressing the complex error terms into one diagram simplifying the interpretation~~.~~ and the comparison of multiple simulations. In this study, we propose a specific model evaluation approach with a strong focus on the error identification which contributes to existing diagnostic evaluation approaches and builds on existing approaches.

2 Methodology

2.1 Diagnostic efficiency

In general, the quality of ~~observations should be verified before simulations and observations are compared against each other. Observations with insufficient accuracy should not be considered for model evaluation. Likewise, accuracy of initial and boundary conditions should be inspected beforehand.~~evaluation data (e.g. streamflow observations) should be verified before

simulations and observations are compared against each other. Model evaluation data with insufficient accuracy should not be considered for model evaluation (e.g. Coxon et al., 2015). Likewise, accuracy of initial and boundary conditions should be inspected beforehand (e.g. Staudinger et al., 2019). Remaining errors in hydrological simulations may then be caused by the following sources:

- model parameters (e.g. Wagener and Gupta, 2005)
- model structure (e.g. Clark et al., 2008;Clark et al., 2011)
- uncertainties in input data (e.g. Yatheendradas et al., 2008)
- ~~uncertainties in observations (e.g. Coxon et al., 2015)~~
- ~~initial and boundary conditions (e.g. Staudinger et al., 2019)~~

Thus, within our approach we focus on errors caused by model parameters, model structure and input data. In order to diagnose the source of the errors, we define three error types which might be linked to potential error sources (e.g. model parameters, model structure and input data): (i) constant error describes the average deviation between simulations and observations; (ii) dynamic error defines the deviation at different simulated and observed magnitudes; (iii) timing error comprises the temporal agreement between simulations and observations. Model errors may have different sources. Assigning the error type to its source requires expert knowledge (e.g. shortcomings of the input data) or statistical analysis (e.g. linking the error types with the model parameters). We provide here some examples how expert knowledge might be used to link the input data with the error type. A constant error might be linked to the precipitation input, for example. Beck et al. (2017) found a negative constant errors in snow-dominated catchments. In case the precipitation input error varies between rainfall events, the input data might be the source for dynamic errors (e.g. Yatheendradas et al., 2008). On the other hand, errors in the spatio-temporal rainfall pattern might be the source for timing errors (e.g. Grundmann et al., 2019).

In order to ~~expand existing diagnostic evaluation approaches~~ quantify the overall error, we introduce the diagnostic efficiency (*DE*; Eq. 1):

$$DE = \text{~~1 -~~} \sqrt{\text{~~}\overline{B_{rel}}^{\text{~~2~~}} + |B_{area}|^{\text{~~2~~}} + (r-1)^{\text{~~2~~}}}\text{~~}\sqrt{\overline{B_{rel}}^2 + |B_{area}|^2 + (r-1)^2}, \quad (1)$$

where $\overline{B_{rel}}$ is a measure for the constant error, $|B_{area}|$ for the dynamic error, and r for the timing error. ~~Similar to *NSE* and *KGE*,~~ *DE* ranges from ~~-1~~0 to ~~-∞.~~∞ and *DE* = ~~1~~0 indicates, that there are no errors (i.e. perfect agreement between simulations and observations~~.~~). In contrast to *KGE* and *NSE*, *DE* represents an error score. This means, that model performance is decreasing for increasing values of *DE*.

First, we introduce the three terms which define the *DE*. The first two terms $\overline{B_{rel}}$ and $|B_{area}|$ are based on the flow duration curve (FDC). Since FDC-based signatures do not include information on temporal performance, we have added correlation (r) between the simulated time series and the observed time series as a third term. $\overline{B_{rel}}$ reflects the constant error and is represented by the arithmetic mean of the relative bias (Eq. 2):

$$\overline{B_{rel}} = \frac{1}{N} \sum_{i=0}^{i=1} B_{rel}(\mathrm{i}), \quad (2)$$

i represents the exceedance probability, N the total number of data points and B_{rel} is the relative bias of the simulated and observed flow duration curve; $\overline{B_{rel}} = 0$ indicates no constant error; $\overline{B_{rel}} < 0$ indicates a negative bias; $\overline{B_{rel}} > 0$ indicates a positive bias. The relative bias between the simulated and observed flow duration curve (B_{rel}) calculates as follows (Eq. 3):

$$B_{rel}(\mathrm{i}) = \frac{Q_{sim}(i) - Q_{obs}(i)}{Q_{obs}(i)}, \quad (3)$$

Q_{sim} is the simulated streamflow at exceedance probability i and Q_{obs} the observed streamflow at exceedance probability i.

The dynamic error is described by the absolute area of the residual bias ($|B_{area}|$; Eq. 4):

$$|B_{area}| = \int_0^1 |B_{res}(i)| \, di, \quad (4)$$

where the residual bias B_{res} is integrated over the entire domain of the flow duration curve. Combining Eq. (2) and Eq. (3) results in:

$$B_{res}(\mathrm{i}) = B_{rel}(\mathrm{i}) - \overline{B_{rel}}, \quad (5)$$

by subtracting $\overline{B_{rel}}$ we remove the constant error and the dynamic error remains. $|B_{area}| = 0$ indicates no dynamic error; $|B_{area}| > 0$ indicates a dynamic error.

To consider timing errors, the Pearson's correlation coefficient (r) is calculated (Eq. 6):

$$~~r = \frac{\sum_{t=1}^{n}(Q_{obs}(i) - \mu_{obs})(Q_{sim}(i) - \mu_{sim})}{\sqrt{(\sum_{t=1}^{n}(Q_{obs}(i) - \mu_{obs})^2)(\sum_{t=1}^{n}(Q_{sim}(i) - \mu_{sim})^2)}}~~ r = \frac{\sum_{t=1}^{n}(Q_{obs}(t) - \mu_{obs})(Q_{sim}(t) - \mu_{sim})}{\sqrt{(\sum_{t=1}^{n}(Q_{obs}(t) - \mu_{obs})^2)(\sum_{t=1}^{n}(Q_{sim}(t) - \mu_{sim})^2)}}, \quad (6)$$

where Q_{sim} is the simulated streamflow at time t, Q_{obs} the observed streamflow at time t, μ_{obs} the simulated mean streamflow, and μ_{obs} the observed mean streamflow. Other non-parametric correlation measures could be used as well.

2.2 Diagnostic polar plot

DE can be used as another aggregated efficiency by simply calculating the overall ~~model performance.~~error. However, the aggregated value only allows for a limited diagnosis since information of the metric terms is not interpreted. Thus, we project *DE* and its metric terms in a radial plane ~~(i.e. similar to a clock)~~ to construct a diagnostic polar plot. An annotated version for a diagnostic polar plot is given in Fig. 3. For the diagnostic polar plot, we calculate the direction of the dynamic error (B_{dir}; Eq. 7):

$$B_{dir} = \int_0^{0.5} B_{res}(i) \, di, \quad (7)$$

where the integral of B_{res} includes values from 0th percentile to 50th percentile. Since we removed the constant error (see Eq. 5), the left half of the integral is positive and the right half (i.e. 50th percentile to 100th percentile) will, thus, be negative and vice versa if the left half of the integral is negative.

In order to differentiate the dynamic error type, we computed the slope of the residual bias (B_{slope}; Eq. 8):

$$B_{slope} = \begin{cases} |B_{area}| \cdot (-1), & B_{dir} > 0 \\ |B_{area}|, & B_{dir} < 0 \\ 0, & B_{dir} = 0 \end{cases}, \quad (8)$$

B_{slope} = 0 expresses no dynamic error; B_{slope} < 0 indicates that there is a tendency of simulations to overestimate high flows and/or underestimate low flows while B_{slope} > 0 indicates a tendency of simulations to underestimate high flows and/or overestimate low flows.

We used the inverse tangent to derive the ratio between constant error and dynamic error in radians (φ; Eq. 9):

$$\varphi = arctan2(\overline{B_{rel}}, B_{slope}), \quad (9)$$

Instead of using a benchmark to decide whether model diagnostics is valuable or not, we introduce certain threshold for deviation-from-perfect. We set a threshold value (l) for which metric terms deviate from perfect and insert it in Eq. (1):

$$DE_l = ~~1 - \sqrt{l^2 + l^2 + ((1-l)-1)^2}~~ \sqrt{l^2 + l^2 + ((1-l)-1)^2}, \quad (11)$$

for this study l is set by default to 0.05. Here, we assume that for a deficient simulation each metric term deviates at least 5% from its best value. l can be either relaxed or expanded depending on the requirements of model accuracy. Correspondingly, DE_l represents a threshold ~~which discerns between a deficient simulation ($DE \leq DE_t$) and a good simulation ($DE > DE_t$).~~to discern whether an error diagnosis ($DE > DE_l$) is valuable.

Finally, the following conditions describe whether a diagnosis can be drawn (Eq. 12):

$$Diagnosis = \begin{cases} ~~yes, \quad |\overline{B_{rel}}| \leq 1 \,\&\, B_{slope} > 1 \,\&\, DE \leq DE_t~~ \\ ~~yes, \quad |\overline{B_{rel}}| > 1 \,\&\, B_{slope} \leq 1 \,\&\, DE \leq DE_t~~ \\ ~~yes, \quad |\overline{B_{rel}}| > 1 \,\&\, B_{slope} > 1 \,\&\, DE \leq DE_t~~ \end{cases} \begin{cases} yes, & |\overline{B_{rel}}| \leq 1 \,\&\, B_{slope} > 1 \,\&\, DE > DE_l \\ yes, & |\overline{B_{rel}}| > 1 \,\&\, B_{slope} \leq 1 \,\&\, DE > DE_l \\ yes, & |\overline{B_{rel}}| > 1 \,\&\, B_{slope} > 1 \,\&\, DE > DE_l \end{cases}, \quad (12)$$

There exists a special case for which timing error only can be diagnosed (Eq. 13):

$$Diagnosis = timing\ error\ only, \quad |\overline{B_{rel}}| \leq ~~1 \,\&\, B_{slope} \leq 1 \,\&\, DE \leq DE_t,~~ 1 \,\&\, B_{slope} \leq 1 \,\&\, DE > DE_l, \quad (13)$$

If DE and its metric terms are within the boundaries of acceptance, no diagnosis is required which is expressed by the following conditions (Eq. 14):

$$Diagnosis = no, \quad |\overline{B_{rel}}| \leq 1 \,\&\, B_{slope} \leq 1 \,\&\, DE ~~> DE_t~~ \leq DE_l, \quad (14)$$

In this case, the model performance is sufficiently accurate and ~~can be denoted as a good simulation~~errors are too small.

2.3 Comparison to *KGE* and *NSE*

In order to allow a comparison to commonly used *KGE* and *NSE*, we calculated the overall metric values and for *KGE* its three individual metric terms. We used the original *KGE* proposed by Gupta et al. (2009):

$$KGE = 1 - \sqrt{(\beta - 1)^2 + (\alpha - 1)^2 + (r - 1)^2}, \quad (15)$$

where β is the bias error, α represents the flow variability error, and r shows the linear correlation between simulations and observations (Eq. 16):

$$KGE = 1 - \sqrt{\left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + (r-1)^2}, \quad (16)$$

where σ_{obs} is the standard deviation in observations, σ_{sim} the standard deviation in simulations. Moreover, we applied the polar plot concept (see Sect. 2.2) to *KGE* and the accompanying three metric terms. In contrast to *DE* (see Sect. ~~2.1) the formulation of *KGE* is entirely based on statistical signatures. By replacing the first two terms of *KGE* with FDC-based signatures, we aim to improve the hydrological focus and provide a stronger link to the error sources.~~2.1), KGE ranges from 1 to -∞ and the metric formulation of *KGE* is entirely based on statistical signatures. By replacing the first two terms of *KGE* with FDC-based signatures, we aim to improve the hydrological focus and provide a stronger link to hydrological processes (e.g. Ghotbi et al., 2020).

NSE (Nash and Sutcliffe, 1970) calculates as follows (Eq. 17):

$$NSE = 1 - \frac{\sum_{t=1}^{t=T}(Q_{obs}(t) - Q_{sim}(t))^2}{\sum_{t=1}^{t=T}(Q_{obs}(t) - \mu_{obs})^2}, \quad (17)$$

where T is the total number of time steps, Q_{sim} the simulated streamflow at time t, Q_{obs} the observed streamflow at time t and μ_{obs}. NSE = 1 displays perfect fit between simulations and observations; NSE = 0 indicates that simulations performs equally well as the mean of the observations; NSE < 0 indicates that simulations perform worse than the mean of the observations.

3 Proof of concept

To provide a proof of concept any perennial streamflow time series coming from a near-natural catchment and having sufficiently long temporal record (i.e. > 30 years) may be used. We selected an observed streamflow time series from the CAMELS dataset (Fig. 1; Addor et al., 2017). In order to generate specific model errors, we systematically manipulated the observed time series. Thus, we produced different time series which serve as a surrogate for simulated time series with a certain error type which we call manipulated time series. These manipulated time series are characterised by a single error type or multiple error types, respectively. We calculated *DE* for each manipulated time series and visualised the results in a diagnostic polar plot.

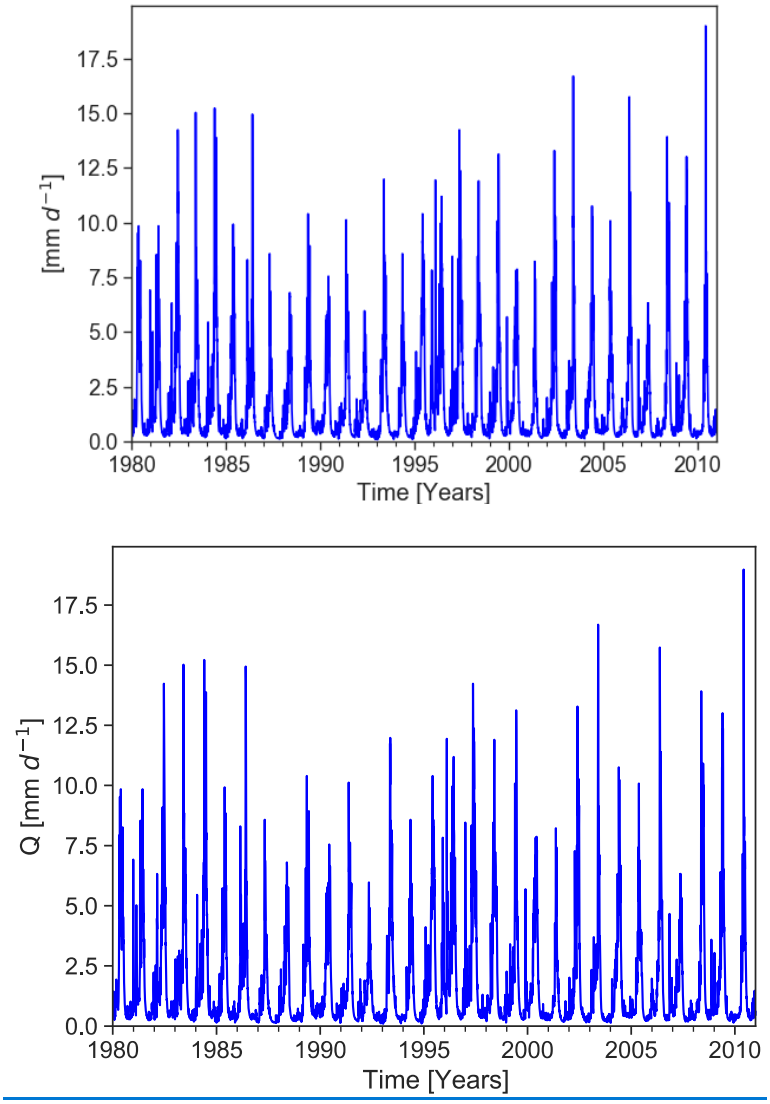


Figure 1: Observed streamflow time series from CAMELS dataset (Addor et al., 2017; gauge id: 13331500; gauge name: Minam River near Minam, OR, U.S.)

3.1 Generation of artificial errors

In the following section, we portray how we manipulated observed time series to generate artificial modelling errors. Table 1 provides a brief summary on the error types and how we combined them. The resultant FDCs are illustrated in Figure 2. For the corresponding time series, we refer to the supplement (Fig. S1). We first describe the genesis of the time series for individual errors:

(a) Positive constant error: We generated a positive offset by multiplying the observed time series with a constant 1.25 (see Fig. 2a and Fig. S1a). Constant ~~requires~~required to be > 1.

(b) Negative constant error: We generated a negative offset by multiplying the observed time series with a constant 0.75 (see Fig. 2b and Fig. S1b). Constant ~~requires~~required to be < 1.

(c) Positive dynamic error: We built a linearly interpolated vector (1+*p*, …, 1, …, *p*) with *p* set to 0.5. We then generated the error by multiplying the observed FDC with the linearly interpolated vector. With that~~,~~ we increased high flows and decreased low flows. As a consequence, hydrological extremes are amplified (see Fig. 2c and Fig. S1c). Note that the original temporal order of the time series is maintained.

(d) Negative dynamic error: We built a linearly interpolated vector (*p*, …, 1, …, 1+*p*) with *p* set to 0.5. We then generated the error by multiplying the observed FDC with the linearly interpolated vector. With that~~,~~ we decreased high flows and increased low flows. As a consequence, hydrological extremes are moderated (see Fig. 2d and Fig. S1d). Note that the original temporal order of the time series is maintained.

(e) We reproduced a timing error by randomizing the order of the observed time series (see Fig. 2e and Fig. S1e).

We then assembled the individual techniques (a-d) for the genesis of time series which are characterised by a combination of constant error and dynamic error. The two errors contribute with an equal share:

(f) Negative constant error and negative dynamic error (see Fig. 2f and Fig. S1f)

(g) Positive constant error and negative dynamic error (see Fig. 2g and Fig. S1g)

(h) Negative constant error and positive dynamic error (see Fig. 2h and Fig. S1h)

(i) Positive constant error and positive dynamic error (see Fig. 2i and Fig. S1i)

and time series which contain constant error, dynamic error (again both errors are contributing with an equal share) and timing error (a-e):

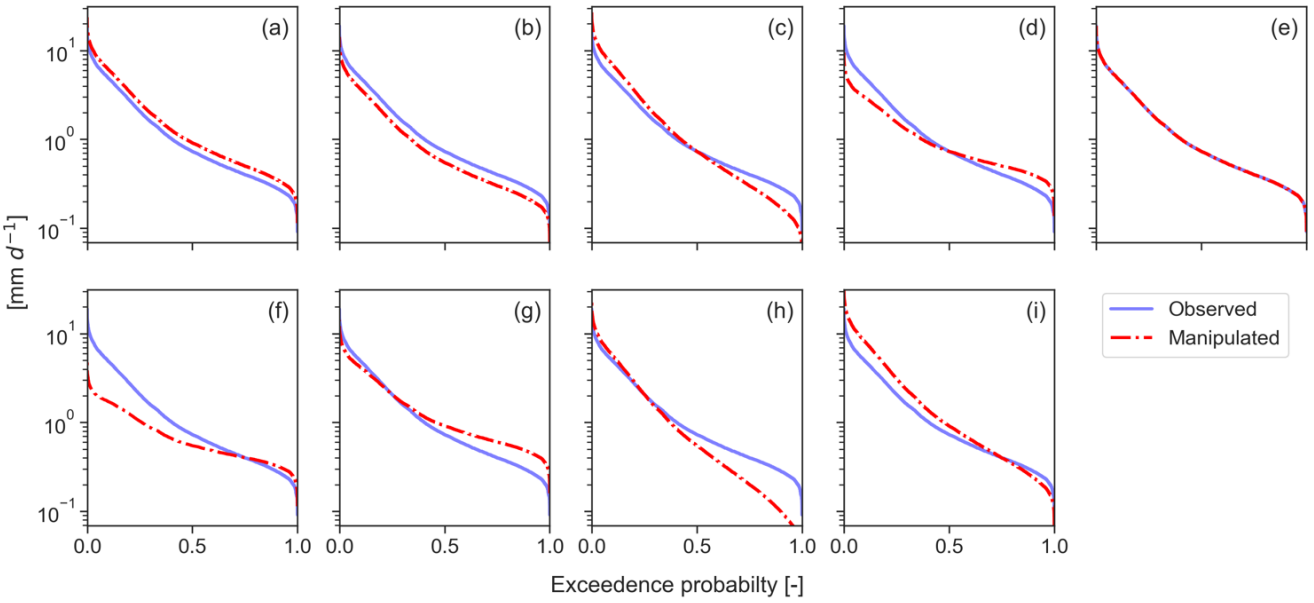
(j) Negative constant error, negative dynamic error and timing error (see Fig. S1j)

(k) Positive constant error, negative dynamic error and timing error (see Fig. S1k)

(l) Negative constant error, positive dynamic error and timing error (see Fig. S1l)

(m) Positive constant error, positive dynamic error and timing error (see Fig. S1m)

Note that for j-m FDCs are identical to f-i and are therefore not shown in Figure 2.



~~**Figure 2: Flow duration curves (FDCs) of observed (blue) and manipulated (dashed red) streamflow time series. Manipulated FDCs are depicted for (a–b) constant errors only, (c–d) dynamic errors only, (e) timing error only, and (f–i) combination of dynamic and constant errors. The combination of constant errors, dynamic errors and timing error is not shown, since their FDCs are identical to f–i. Y-axis is shown in log-space.**~~

The diagnostic polar plot for synthetic error cases is shown in Fig. 3. ~~Interdependently which~~Since each synthetic error ~~has been generated~~case is different, related points are located in different error regions. For individual errors (a-d), related points are placed in the four cardinal directions of each region (Fig. 3). Within these regions the dominant error type can be easily identified. The more central the direction of the point, the more dominant is the error type. In case there is only a timing error present (e) an arrow with two ends instead of a point is used (Fig. 3). This is because dynamic error source becomes arbitrary (i.e. high flows and low flows are being both underestimated and overestimated (see Fig. S1e)). For combinations of constant and dynamic error (f-i), related points are located on boundaries of constant error and dynamic error meaning that both errors are equally dominant (Fig. 3). The same applies for combinations of constant error, dynamic error and timing error except that points shifted towards outer scope of the plot due to added timing error. Numeric values of *DE* are listed in Table 2. *DE* values are ~~greater~~lower for individual errors (except for timing error) than for combined errors. Increasing the number of errors added to a time series, leads to ~~lower~~greater *DE* values. For the numeric values of the individual metric terms, we refer to Table S1.

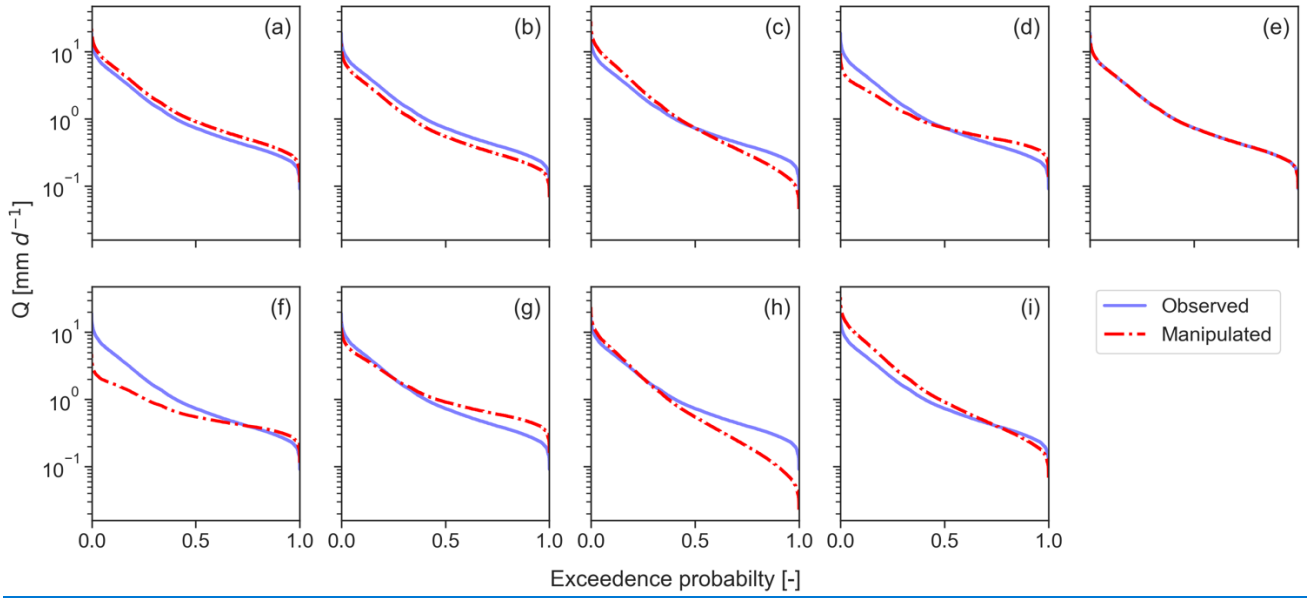


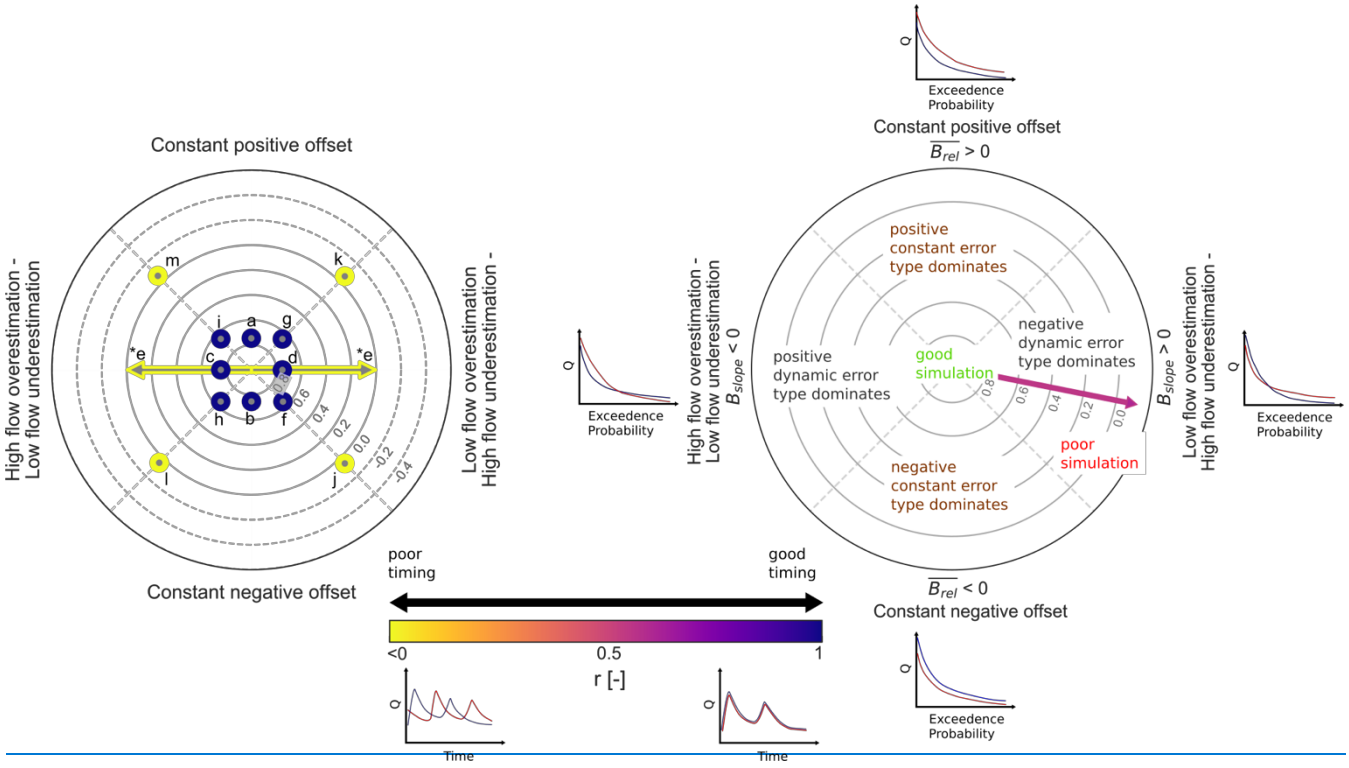
Figure 2: Flow duration curves (FDCs) of observed (blue) and manipulated (dashed red) streamflow time series. Manipulated FDCs are depicted for (a-b) constant errors only, (c-d) dynamic errors only, (e) timing error only, and (f-i) combination of dynamic and constant errors. The combination of constant errors, dynamic errors and timing error is not shown, since their FDCs are identical to f-i. Y-axis is shown in log space.

A comparison of *DE*, *KGE*, and *NSE* calculated for the manipulated time series is shown in Table 2. ~~Numerically, *DE* generally indicates a better performance than *KGE* and *NSE*.~~ Moreover, values for *DE* exhibit a regular pattern (i.e. generating single error types or multiple error types, respectively, leads to an equidistant decrease in performance). By contrast, values for *KGE* and *NSE* are characterised by an irregular pattern (i.e. generating single error types or multiple error types, respectively, leads to a non-equidistant decrease in performance). This non-equidistant decrease of *KGE* and *NSE* scores suggests that *KGE* and *NSE* are differently sensitive to the generated errors. ~~For example, lowest *KGE* values for single~~ *KGE* is more sensitive to constant errors and dynamic errors ~~are obtained by only introducing one error type~~ (Table 2a-d~~).~~), whereas *NSE* is ~~prone~~more sensitive to timing errors (Table 2e~~), particularly to~~). Particularly, the spurious timing of the peak flows leads to an strong decrease of *NSE* (Table 2m). When combining positive constant error and negative dynamic error, and vice versa (see Table 1g,h), *KGE* and *NSE* display better performance (Table 2g,h) than for single constant and dynamic error types (Table 2a-d).

Table 1: Summary on error types and its combinations as described in Sect. 3.1 (a-m). + (-) reflects a positive (negative) error type. For timing error, only one error type exists (x).

	a	b	c	d	e	f	g	h	i	j	k	l	m
Constant error **(+/-)**	**+**	**-**				**-**	**+**	**-**	**+**	**-**	**+**	**-**	**+**

Dynamic error **(+/-)**			+	-		-	-	+	+	-	-	+	+
Timing error **(x)**					x					x	x	x	x



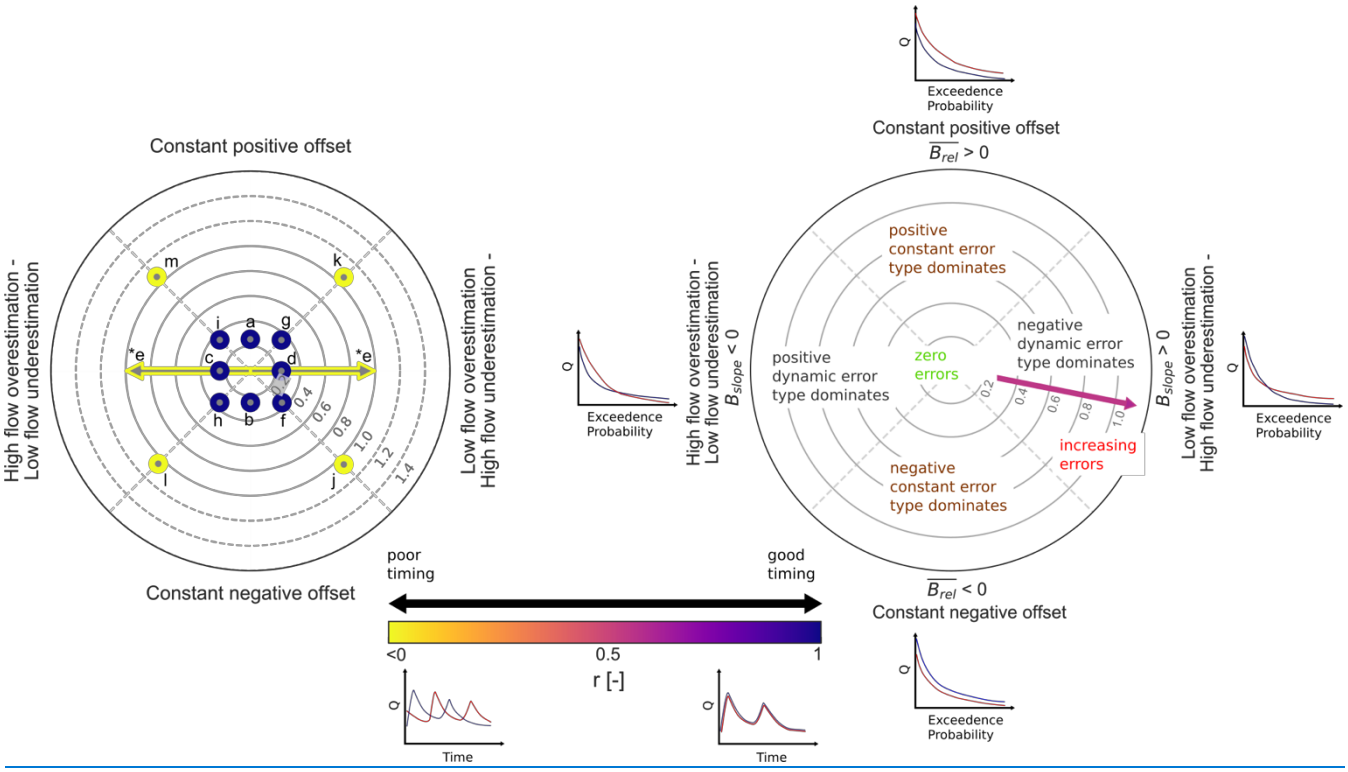


Figure 3: (left) Diagnostic polar plot for manipulated time series generated characterized by constant errors, dynamic errors and timing errors (a-m) visualizing the overall model performance (*DE*; contour lines) and contribution of constant error, dynamic error and timing error (purple (yellow) indicates temporal match (mismatch)). (e*) timing error only: type of dynamic error cannot be distinguished. (right) Annotated diagnostic polar plot illustrating the interpretation (similar to Zipper et al. (2018)). Hypothetic FDC plots and hydrograph plots give examples for the error types.

Table 2: Comparison of *DE*, *KGE* and *NSE* calculated for manipulated time series characterized by constant errors, dynamic errors and timing errors (a-m). Lowest model performance for each error case is in bold.

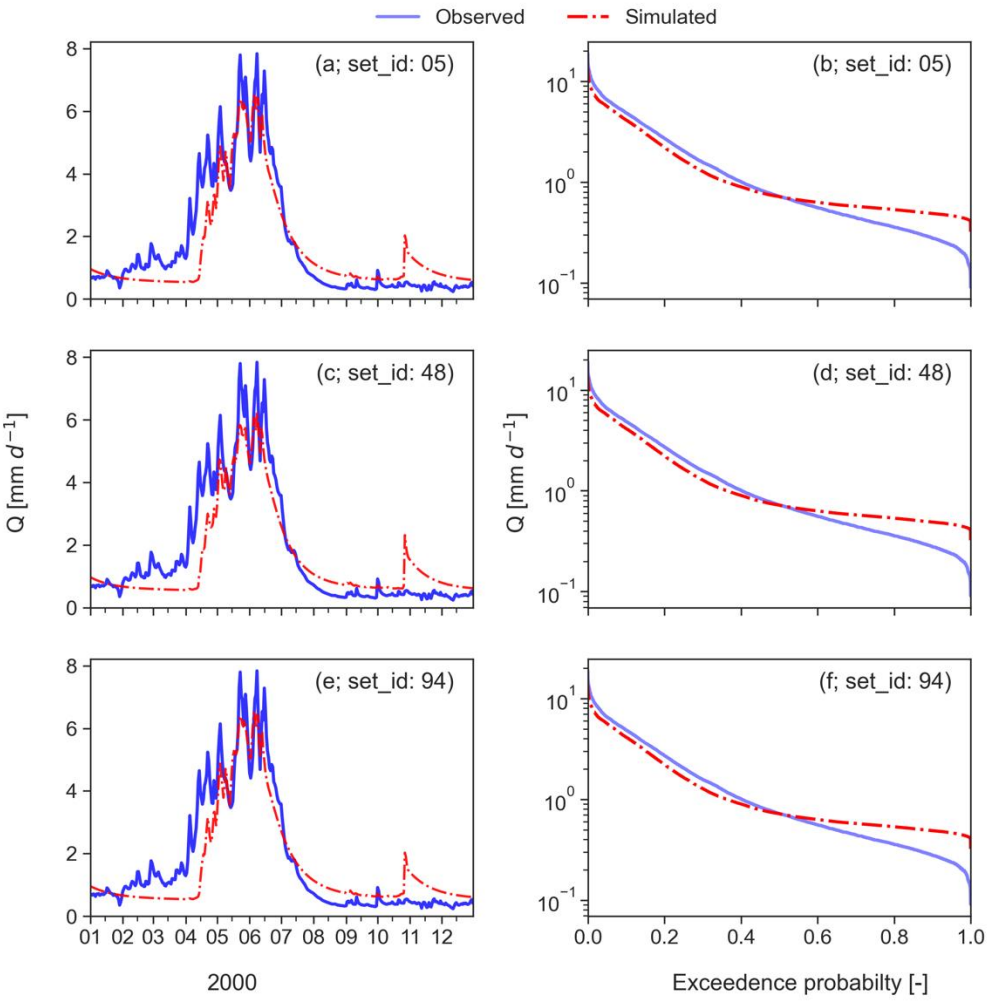
	a	b	c	d	e	f	g	h	i	j	k	l	m
DE	0.75 25	0.752 5	0.75 25	0.75 25	01	0.65 35	**0.65 35**	**0.65 35**	0.65 35	- 01.06	- 01.06	- 01.06	- 01.06
KGE	**0.65**	**0.65**	**0.43**	**0.43**	0	**0.08**	0.75	0.75	**0.08**	**-0.36**	-0.04	-0.04	-0.36
NSE	0.9	0.9	0.7	0.7	**-1**	0.27	0.94	0.94	0.27	-0.25	**-0.59**	**-1.58**	**-3.26**

3.2 Modelling example

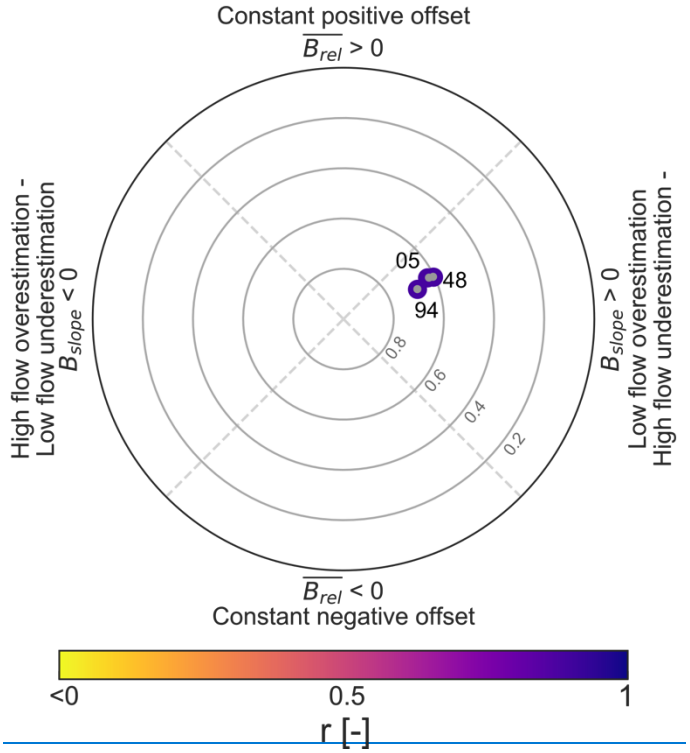
In order to demonstrate the applicability, we also use simulated streamflow time series which have been derived from Addor et al. (2017). Streamflow time series have been simulated by the coupled Snow-17 and SAC-SMA system for the same catchment as in Fig. 1. We briefly summarize here their modelling approach consisting of Snow-17 which "is a conceptual air-

temperature-index snow accumulation and ablation model" (Newman et al., 2015) and SAC-SMA model which is "a conceptual hydrologic model that includes representation of physical processes such as evapotranspiration, percolation, surface flow, and subsurface lateral flow" (Newman et al., 2015). Snow-17 runs first to partition precipitation into rain and snow and delivers the input for SAC-SMA model. For further details about the modelling procedure we refer to Sect. 3.1 in Newman et al. (2015). In particular, we evaluated three model runs with different parameter sets, but the same input data. Simulated time series and simulated FDCs are shown in Fig. 4. The diagnostic polar plot for the three simulated time series is provided in Fig. 5. Simulations realised by parameter set with set_id 94 outperform the other two parameter sets. 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. The modelling example highlights one advantage of the proposed evaluation approach that multiple simulations can be easily compared to each other. For the case of the modelling example, model performance of slightly different parameter sets can be clearly distinguished although the parameter sets are characterized by a similar error type. After identifying the error type and its contributions, these results can be used in combination with expert knowledge (e.g. model developer) or statistical analysis to infer hints on improving the simulations.

~~After identifying the error type and its contributions, we can infer hints on how to improve the simulations. From a process-based (perceptual) perspective, the apparent negative dynamic error described by high flow underestimation and low flow overestimation suggest that process realism (e.g. snow melt, infiltration, storage outflow) appears to be deficient. Measures for improvement could start with adjusting the model parameters (e.g. refining the calibration procedure). If necessary, a follow-up measure could be to alter the model structure (e.g. adjusting the model equations). Additionally, there is a positive constant error available. Because a constant error may be linked to input data errors, this implies that adjusting the input data (e.g. precipitation correction, estimation of evapotranspiration) might improve the simulations.~~



~~**Figure 4: Simulated and observed streamflow time series of modelling example for the year 2000 (a, c and e) and the related flow duration curves for the entire time series (b, d and f). Time series are derived from the CAMELS dataset (Addor et al., 2017). Observations and simulations belong to the same catchment as in Figure 1. Simulations were produced by model runs with different parameter sets (set_id) but same input data (see Newman et al., 2015).**~~



~~**Figure 5: Diagnostic polar plot for modelling example. Simulations were realised with three different parameter sets (05, 48, 94; see Fig. 4). All simulations perform well. However, the remaining error is dominated by a negative dynamic error type while timing is excellent.**~~

4 Discussion

Aggregated performance metrics (e.g. *KGE* and *NSE*) are being criticised for not being hydrologically informative (Gupta et al., 2008). Although we systematically generated errors, we found ~~an illogical~~a disjointed pattern for *KGE* and *NSE* (Table 2) which makes the interpretation of *KGE* and *NSE* more difficult. Particularly, in-depth analysis of the *KGE* metric terms revealed, that the β term and α term are not orthogonal to each other (see Fig. S2 and Fig. S3c). We also lump model performance into a single value, but *DE* ~~differs mainly in two points from the *KGE* and the *NSE*~~has the following advantages: (i) metric formulation is based rather on a hydrological understanding than a purely statistical understanding; (ii) the combined visualization of the efficiency metric and the different metric terms enables the identification of the dominant error type; (iii) diagnostic polar plots facilitate ~~exploration of model deficiencies and diagnostics.~~comparison of multiple simulations. Using *DE* as an error score improves the interpretation of the numerical value. *DE* equals zero can be cleanly interpreted as zero errors. Additionally, numerical values of the first and the second metric term of *DE* equal to zero can also be interpreted as zero errors. Compared to KGE, the included FDC-based measures may be easier linked to different hydrologic processes than

purely statistical measures. For example, slow flow processes (e.g. baseflow) control the low flow segment of the FDC while fast flow processes (e.g. surface runoff) control the high flow segment of the FDC (Ghotbi et al., 2020). When using *KGE* and *NSE* for evaluation purposes, we recommend a comparison to hydrologically meaningful benchmarks which may add diagnostic value to *KGE* (e.g. Knoben et al., 2019) and *NSE* (e.g. Schaefli and Gupta, 2007). Based on such benchmark, skill scores have been recently proposed to evaluate simulations (Knoben et al., 2019;Towner et al., 2019;Hirpa et al., 2018) to communicate model performance and to improve hydrologic interpretation. So far a way to define hydrologically meaningful benchmarks has not been extensively addressed by the hydrologic modelling community (Knoben et al., 2019).

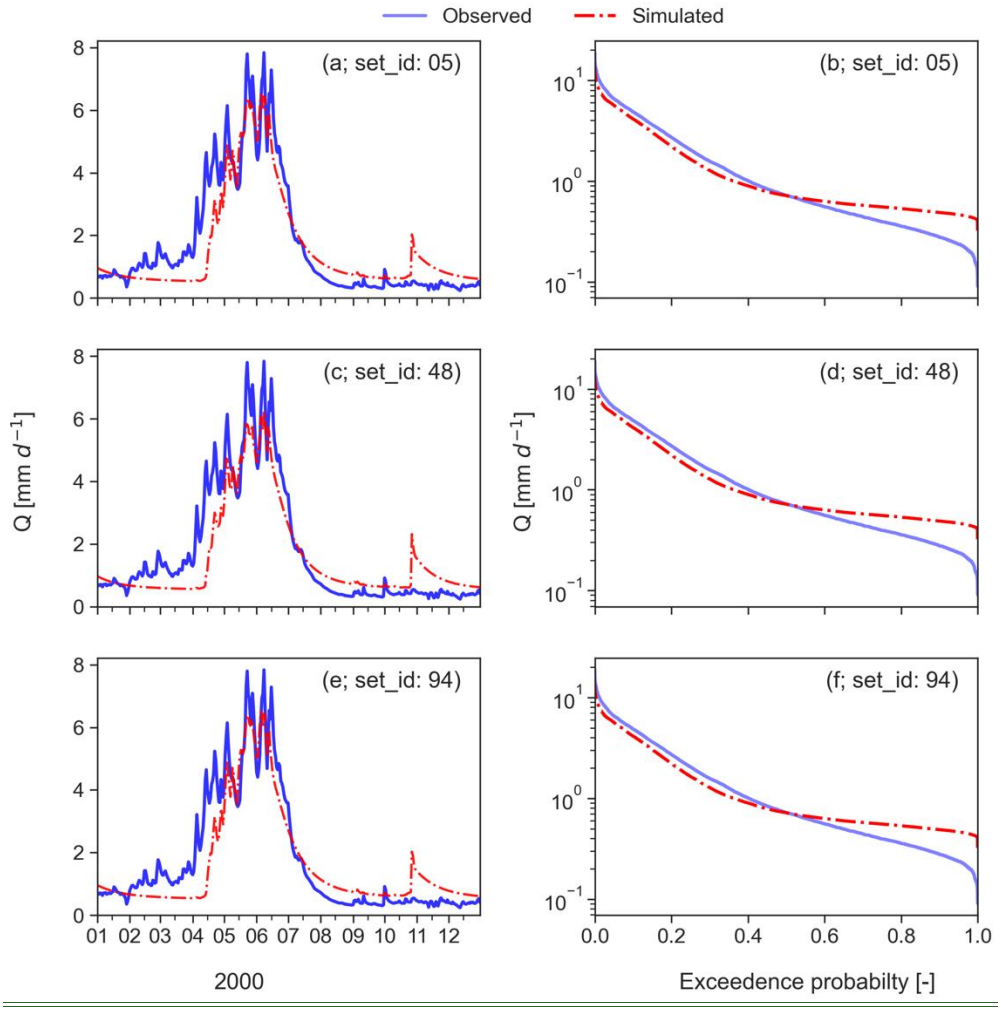


Figure 4: Simulated and observed streamflow time series of modelling example for the year 2000 (a, c and e) and the related flow duration curves for the entire time series (b, d and f). Time series are derived from the CAMELS dataset (Addor et al., 2017). Observations and simulations belong to the same catchment as in Figure 1. Simulations were produced by model runs with different parameter sets (set_id) but same input data (see Newman et al., 2015).

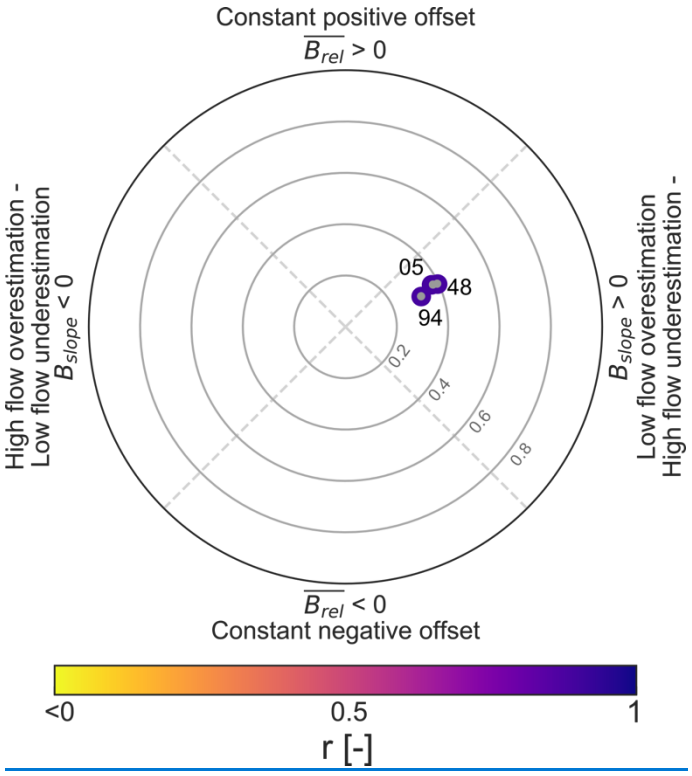


Figure 5: Diagnostic polar plot for modelling example. Simulations were realised with three different parameter sets (05, 48, 94; see Fig. 4). All simulations perform well. However, the remaining error is dominated by a negative dynamic error type while timing is excellent.

Our approach focuses on model ~~deficiencies. We~~errors. Since the *DE* can be interpreted as an error score, we do not propose a skill score measure for *DE* ~~since skill~~. Skill scores are known to introduce a scaling issue on communicating model errors (Knoben et al., 2019). *DE* does not rely on any benchmark to decide whether model diagnostics are required or not. Without considering any benchmark, *DE* may be interpreted as a deviation-from-perfect, measured by its constant error, dynamic and temporal error terms. In Sect. 2.2 (see Eq. 11) we introduced certain threshold for deviation-from-perfect (e.g. *DE*=0.~~91~~09), if all error terms deviate by a certain degree (e.g. 5%; $\overline{B_{rel}}$=0.05, $|B_{area}|$=0.05, r=0.95). Only for simulations in which deviation-from-perfect is sufficiently large, model diagnostics will be valuable.

By including FDC-based information into *DE*, we aimed for capturing rainfall-runoff response behaviour (Vogel and Fennessey, 1994) where different aspects of the FDC are inherently related to different processes (Ghotbi et al., 2020). But the way the dynamic error term is calculated (see Eqs. 4,5 and 7) limits the applicability to catchments with perennial streamflow. Moreover, the second metric term of *DE* (see Eq. 1) is limited to measure only the overall dynamic error. The question whether high flow errors or low flow errors are more prominent cannot be answered. Measuring the timing error by linear correlation may also have limitations. Linear correlation can be criticised for neglecting specific hydrological behaviour (Knoben et al.,

2019), for example, flow recession or peak flow timing. But *DE* could also be calculated for different time periods and hence specific periods (e.g. wet periods versus dry periods) could be diagnosed separately.

Combining *DE* and diagnostic polar plots is, however, limited to three metric terms, because higher dimensional information cannot be effectively visualised by polar plots. We emphasize that the proposed metric terms of *DE* might not be perfectly suitable for every evaluation purpose. For more specific evaluation, we suggest tailoring the proposed formulation of *DE* (see Eq. 1) by exchanging the metric terms with, for example, low-flow-specific terms (e.g. see Fowler et al., 2018) or high-flow-specific terms (e.g. see Mizukami et al., 2019), respectively. Moreover, we suggest that different formulations of *DE* can be combined to a multi-criteria diagnostic evaluation (see Appendix A).

5 Conclusions

The proposed approach is used as a tool for diagnostic model evaluation. Incorporating the information of the ~~model performance~~overall error and the metric terms into the evaluation process represents a major advantage. Although ~~errors~~different error types may have ~~multiple sources~~different contributions, these may be explored visually by diagnostic polar plots. 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~~confirmed the ~~help~~applicability of ~~expert knowledge or a statistical analysis.~~our approach. Particularly, diagnostic polar plots facilitate interpretation of model evaluation results and the comparison of multiple simulations. These plots may advance model development and application. The comparison to Kling-Gupta Efficiency and Nash-Sutcliffe Efficiency revealed, that they rely on a comparison to hydrological meaningful benchmarks to become diagnostically interpretable. 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.~~. More generally, our approach may serve as a blueprint for developing other Diagnostic Efficiency measures in the future.

Supplement. The supplement related to this article is available online at: https://doi.org/10.5194/hess-2020-237-supplement.

Code availability. We provide a Python package *diag-eff* which can be used to calculate DE and the corresponding metric terms, produce diagnostic polar plots or generate artificial errors. The stable version can be installed via the Python Package Index (PyPI), and the current development version is available at https://github.com/schwemro/diag-eff.

Data availability. The observed and simulated streamflow time series are part of the open-source CAMELS dataset (Addor et al., 2017). The data can be downloaded at https://ncar.github.io/hydrology/datasets/CAMELS_timeseries.

Author contributions. RS came up with initial thoughts. RS, DD and MW jointly developed and designed the methodology. RS developed the Python package, produced the figures and tables, and wrote the first draft of the manuscript. The manuscript was revised by DD and MW and edited by RS.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. We are grateful to Kerstin Stahl ~~and Julia Dörrie~~ for ~~their~~her comments on the language style and structure of the manuscript. We thank Wouter Knoben and two anonymous Reviewers for their constructive comments, which helped us to clarify and improve this manuscript.

Financial support. This research has been supported by Helmholtz Association of German Research Centres (grant no. 42-2017). The article processing charge was funded by the Baden-Wuerttemberg Ministry of Science, Research and Art and the University of Freiburg in the funding programme Open Access Publishing.

Review statement. This article was edited by Genevieve Ali and reviewed by Wouter Knoben and two anonymous Reviewers.

Appendix A

We briefly describe how *DE* could be extended to a tailored single-criteria metric (A1):

$$DE_{ext} = \sqrt{term_1^2 + term_2^2 + term_3^2}, \quad \text{(A1)}$$

Multiple single-criteria metric can be combined to a multi-criteria metric (A2):

$$DE_{multi-ext} = \frac{1}{N} \sum_{i=1}^{N} DE_{ext,i}, \quad \text{(A2)}$$

For a multi-criteria approach, diagnostic polar plots can be displayed for each single-criteria metric included into A2.

References

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, in, version 2.0 ed., Boulder, CO: UCAR/NCAR, 2017.

Beck, H. E., van Dijk, A. I. J. M., de Roo, A., Dutra, E., Fink, G., Orth, R., and Schellekens, J.: Global evaluation of runoff from 10 state-of-the-art hydrological models, Hydrology and Earth System Sciences, 21, 2881–2903, 10.5194/hess-21-2881-2017, 2017.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, 10.1029/2007wr006735, 2008.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, 10.1029/2010wr009827, 2011.

Coxon, G., Freer, J., Westerberg, I. K., Wagener, T., Woods, R., and Smith, P. J.: A novel framework for discharge uncertainty quantification applied to 500 UK gauging stations, Water Resources Research, 51, 5531-5546, 10.1002/2014wr016532, 2015.

Fowler, K., Peel, M., Western, A., and Zhang, L.: Improved Rainfall-Runoff Calibration for Drying Climate: Choice of Objective Function, Water Resources Research, 54, 3392-3408, 10.1029/2017wr022466, 2018.

Ghotbi, S., Wang, D., Singh, A., Blöschl, G., and Sivapalan, M.: A New Framework for Exploring Process Controls of Flow Duration Curves, Water Resources Research, 56, 10.1029/2019WR026083, 2020.

Grundmann, J., Hörning, S., and Bárdossy, A.: Stochastic reconstruction of spatio-temporal rainfall patterns by inverse hydrologic modelling, Hydrol. Earth Syst. Sci., 23, 225-237, 10.5194/hess-23-225-2019, 2019.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling theory with observations: elements of a diagnostic approach to model evaluation, Hydrological Processes, 22, 3802-3813, 10.1002/hyp.6989, 2008.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377, 80-91, 10.1016/j.jhydrol.2009.08.003, 2009.

Hirpa, F. A., Salamon, P., Beck, H. E., Lorini, V., Alfieri, L., Zsoter, E., and Dadson, S. J.: Calibration of the Global Flood Awareness System (GloFAS) using daily streamflow data, Journal of Hydrology, 566, 595-606, 10.1016/j.jhydrol.2018.09.052, 2018.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331, 10.5194/hess-23-4323-2019, 2019.

Mizukami, N., Rakovec, O., Newman, A. J., Clark, M. P., Wood, A. W., Gupta, H. V., and Kumar, R.: On the choice of calibration metrics for "high-flow" estimation using hydrologic models, Hydrol. Earth Syst. Sci., 23, 2601-2614, 10.5194/hess-23-2601-2019, 2019.

Nash, J. E., and Sutcliffe, J. V.: River flow forecasting through conceptual models part I - A discussion of principles, Journal of Hydrology, 10, 282-290, 10.1016/0022-1694(70)90255-6, 1970.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance, Hydrol. Earth Syst. Sci., 19, 209-223, 10.5194/hess-19-209-2015, 2015.

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., and Nearing, G.: Benchmarking of a Physically Based Hydrologic Model, Journal of Hydrometeorology, 18, 2215-2225, 10.1175/jhm-d-16-0284.1, 2017.

Pechlivanidis, I., Jackson, B., and McMillan, H.: The use of entropy as a model diagnostic in rainfall-runoff modelling, International Congress on Environmental Modelling and Software, Ottawa, Canada, 2010,

Schaefli, B., and Gupta, H. V.: Do Nash values have value?, Hydrological Processes, 21, 2075-2080, 10.1002/hyp.6825, 2007.

Shafii, M., Basu, N., Craig, J. R., Schiff, S. L., and Van Cappellen, P.: A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework, Water Resources Research, 53, 3279-3301, 10.1002/2016wr019736, 2017.

Staudinger, M., Stoelzle, M., Cochand, F., Seibert, J., Weiler, M., and Hunkeler, D.: Your work is my boundary condition!: Challenges and approaches for a closer collaboration between hydrologists and hydrogeologists, Journal of Hydrology, 571, 235-243, 10.1016/j.jhydrol.2019.01.058, 2019.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., and Stephens, E. M.: Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin, Hydrol. Earth Syst. Sci., 23, 3057-3080, 10.5194/hess-23-3057-2019, 2019.

Vogel, R. M., and Fennessey, N. M.: Flow Duration Curves. I: New Interpretation and Confidence Intervals, Journal of Water Resources Planning and Management, 120, 485-504, 10.1061/(ASCE)0733-9496(1994)120:4(485), 1994.

Wagener, T., and Gupta, H. V.: Model identification for hydrological forecasting under uncertainty, Stochastic Environmental Research and Risk Assessment, 19, 378-387, 10.1007/s00477-005-0006-5, 2005.

Yatheendradas, S., Wagener, T., Gupta, H., Unkrich, C., Goodrich, D., Schaffner, M., and Stewart, A.: Understanding uncertainty in distributed flash flood forecasting for semiarid regions, Water Resources Research, 44, 10.1029/2007wr005940, 2008.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resources Research, 44, 10.1029/2007wr006716, 2008.

Zipper, S. C., Dallemagne, T., Gleeson, T., Boerman, T. C., and Hartmann, A.: Groundwater Pumping Impacts on Real Stream Networks: Testing the Performance of Simple Management Tools, Water Resources Research, 54, 5471-5486, 10.1029/2018wr022707, 2018.