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 technical note presents an intriguing new metric fusing together aspects of traditional efficiency and hydrologic signature metrics. The research is highly relevant to HESS, and the technical methodology is well described. The results used to demonstrate the utility of the new method of evaluating model performance are sufficient to support the conclusions of the manuscript. Overall, the material is well structured but there are some aspects which are unclear or insufficiently explained.

We thank the reviewer for his/her helpful comments.

Specific comments

31: I do not see how traditional efficiency metrics only allow a binary choice between 'good' and 'poor'. They provide a gradation of relative performance. This should be rephrased.

We fully agree and rephrase the text accordingly.

61: The justification is missing or misplaced. Why these three and not the other two?

We would like to point out that the justification is placed at line 55ff.

62: The three types of model error are a key point in the manuscript, but this 'definition' is inadequate. Why these three types? What distinguishes the types? Listing potential sources of each type does not define anything. What is the difference between constant error from model parameters and dynamic error from model parameters?

We used these three error types because constant, dynamic and timing errors are common model errors. We would like to emphasize, that each error type is calculated as an individual term in the DE. In order to assign the error types (constant, dynamic, timing) to error sources (input data error, parameters, model structure, etc.) contextual/expert knowledge (e.g. shortcomings of the input data) or statistical analysis (e.g. linking the error types with model parameters) is required. We will rephrase the definition and add further explanations.

71: Superficially, the DE metric looks like KGE (three component terms, covering bias, variability and correlation). The manuscript could be improved with an explicit contrast between the two, to highlight the novel aspects of the DE metric. Section 2.3 would be a good place, as it currently does not include a comparison, only formula regurgitation.

We will strengthen the difference and add a sentence including a comparison in Section 2.3. Furthermore, we would like to point out that the supplement contains a comparison between the *DE* terms and the *KGE* terms for the artificially generated errors (Figure S2, Figure S3 and Table S1) and for the modelling example (Figure S4 and Table S2). These results are discussed in Section 4 (see lines 248ff).

151: 'Mimicking' may not be the best term to describe the artificial errors generated for this demonstration. To mimic is to imitate, and the synthetic errors introduced to the observed time series are not intended to imitate anything in particular.

We agree and will rephrase the term mimicking into generation of artificial errors.

180: The summary table is very useful, but grid lines would improve the readability.

We add grid lines.

240: This paragraph has glossed over one key limitation of the new error metric. The 'negative dynamic error' lumps together high flow underestimation and low flow overestimation. The results presented in Figure 4 are a perfect example of why this is a limitation: all three time series have only low flow overestimation as a prominent error. How is the diagnostic polar plot (Fig 5) more informative than the FDC presented in Figure 4?

We agree that the lumping represents a limitation and we will add a paragraph to the manuscript. In most cases high flow underestimation and low flow overestimation are not equally prominent. We emphasize that with DE and the corresponding diagnostic polar plot only the main error can be identified. In order to explore more specific errors, we recommend to include specific signatures (see Appendix A).

A visual evaluation and comparison of the FDCs (see Figure 4) does not allow the identification of the best parameter set. For example, it would be difficult to find the "best" parameter set from 100 model runs just from the FDC. *KGE* and *NSE* do not provide any information on which parts of the FDC are underestimated or overestimated, respectively. Moreover, a separated interpretation of the FDC and the efficiency metric do not give any hint towards the error type. The strength of our approach is the combined visualization of the overall model performance and the different metric terms which enables the identification of the dominant error type. Figure 5 clearly shows which is the best parameter set and what are the dominant errors although the parameter sets perform slightly different.

253: You have stated that the metric formulation is based on hydrological rather than purely statistical understanding, but this has not come out clearly earlier in the text. After all, one of your three component terms is identical to one used in the KGE. A more explicit justification for the hydrological basis would better support the novelty of your metric.

Since the first two terms of DE are based on the FDC, we argue that this improves the hydrological understanding. We strengthen the hydrological justification in the manuscript. Moreover, we want to stress that the metric terms could be easily replaced with other hydrologic signatures (see Appendix A).

273: If the use of polar plots is limiting the information content, why not use some other type of plot? For example, could a radar chart be used instead?

The polar plot is just one way to visualize multidimensional information. Of course, radar chart could be used instead. The polar plot technique facilitates multiple evaluations (e.g. multiple simulations from different parameter sets or multiple simulations from different models) since points are used instead of polygon shapes.

Technical corrections:

7: Should be 'part of' not 'part for'.

10: Unsatisfactory rather than unsatisfying.

10: Originate not origin.

15: Should be 'these three' not 'the three' as other error types are possible but not account for here.

21: Extra comma after 'suggests'.

31: Should this be "model performance using only a single numerical value"?

44: You do not need two qualifiers in this sentence, use either 'usually' or 'may only be' but not both.

52: This is not the best way to introduce the topic of model error or the stated topic of diagnostic efficiency.

55: 'Sources' may be more appropriate than 'origins' in this context.

96: The word 'does' is extraneous.

Figure 1: The figure could use a y-axis title, and I'm not sure that 'years' is an appropriate unit for dates.

149: Are the underscores appropriate for a caption?

152: In the following what? List, table or section?

285-287: Sentence contains grammatical errors, please correct.

We will include all technical corrections in the manuscript.

Response to Reviewer #2

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.

The authors present an interesting technical note in which they link the idea of diagnostic model evaluation with that of efficiency metrics. They propose a new metric in which they integrate terms to assess constant, dynamic and timing errors. I like the idea and the paper, but I am unclear about the way this metric and its terms are formulated, and how they relate to previous work. Hopefully my comments below help the authors to strengthen their argument.
We thank the reviewer for his/her useful comments.

MAJOR COMMENTS
[1] I understand that the first term of their metric is the relative bias of the FDC. Why is this a more hydrologically relevant and insightful term than other bias estimates? Can you show evidence for this claim?
The relative bias of the FDC (i.e. constant bias) may have similar hydrological relevance than other bias estimates. Since we remove first the constant error (see Eq. 5) before we compute the dynamic error, we used the relative bias of the FDC for reasons of consistency.

[2] Similarly, I would find it more informative if the authors were to compare their terms to the terms in KGE and the non-parametric version by Pool et al. (2018) to really understand the differences. Why are these more informative and can it be shown?
For the comparison between the *DE* terms and the *KGE* terms for the artificially generated errors, we would like to refer to the supplement (Figure S2, Figure S3 and Table S1). Similarly, the comparison between the *DE* terms and the *KGE* terms for the modelling example, we would like to refer to the supplement (Figure S4 and Table S2). We will add a sentence to Section 2.3 (as suggested by reviewer #1) which strengthens the difference between DE terms and KGE terms. In addition, as already suggested in the paper, a non-parametric version of the DE could also be used (replacing Pearson's correlation coefficient with for example the Spearman's rank coefficient)

[3] Would it not be more informative if the different parameter sets in Figure 4 were to show that different errors dominate? Why do they all show essentially identical FDCs? Maybe use more varied examples?
The overall objective of the modelling example in Section 3.2 is to demonstrate the applicability of our approach. Of course we could have used an example for which different error dominates. In order to illustrate, when different errors dominate we would like to refer to Figure 2. The FDCs seem to be almost identical, because we compared three model runs which are among the ten best parameter sets. Figure 5 clearly shows which is the best parameter set and what are the dominant errors although the parameter sets perform slightly different.

[4] Is the main problem one of aggregation? And hence loss of information. See for example the separate use of KGE terms in Gudmundsson et al. (2012). Even your second term is more informative because it leads to less aggregation and loss of information. Is this the key?

We would like to point out, that we try to overcome the problem of aggregation by separating and visualising the model performance in combination with the metric terms. Using the polar plot technique, the results can be visualised in a disaggregated way. However, a certain level of aggregation cannot be avoided since each metric term already reflects an aggregation itself. We add a sentence to Section 4 which will highlight the value of including the metric terms into the model evaluation.

[5] It would be good if the authors would clarify their assumptions better and discuss how these might relate to reality. For example, they assume that precipitation has a consistent input data error. Some previous studies suggest that such an input error varies significantly between rainfall events (e.g. Yatheendradas et al., 2008, WRR). Similarly, for the other errors. It would strengthen the study significantly if the authors where to review the literature thoroughly for studies that discuss how these different errors manifest themselves (the authors lines 61ff). The three assumptions made here are key to the paper, but they are currently not supported by literature. I am not arguing that the authors' assumptions are wrong (though I might disagree partially), but they need to show evidence why these assumptions are reasonable. How to assign these errors is key here, but it is also something many people have argued about before.

We highly appreciate this critical comment. In order to assign the error sources contextual/expert knowledge (e.g. shortcomings of the input data) or statistical analysis (e.g. linking the error types with model parameters) is required. We will rephrase the definition, add further explanations and provide the missing references.

[6] There have been others who raised the question of benchmarks before. For example Jan Seibert (https://eprints.ncl.ac.uk/file_store/production/246998/A084BCF1-F4EA-4EDF-AE6D-9E85C27A9DC4.pdf or Seibert, 2001). It would be good if the authors would review the literature more thoroughly on this topic.

The point we want to make here is that DE does not require any benchmark for an improved hydrological interpretation (see lines 298ff).

[7] Section 3.7 is difficult to follow. Maybe this can easier be summarized in a figure? I find these error combinations difficult to read and compare. Maybe another figure instead of the table?

Unfortunately, there does not exist a Section 3.7. We assume that the comment addresses Section 3.1. We recommend using Figure 3 in combination with Table 1.

REFERENCES

Gudmundsson, L., T. Wagener, L. M. Tallaksen, and K. Engeland (2012), Evaluation of nine large-scale hydrological models with respect to the seasonal runoff climatology in Europe, Water Resour. Res., 48, W11504, doi:10.1029/2011WR010911.

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. Hydrological Sciences Journal, 63(13-14), 1941-1953.

Seibert J. 2001. On the need for benchmarks in hydrological modelling. Hydrological Processes 15 (6): 1063–1064 DOI: 10.1002/hyp.446

Yatheendradas, S., T. Wagener, H. Gupta, C. Unkrich, D. Goodrich, M. Schaffner, and A. Stewart (2008), Understanding uncertainty in distributed flash flood forecasting for semiarid regions, Water Resour. Res., 44, W05S19, doi:10.1029/2007WR005940.

List of all relevant changes

- We rephrased inappropriate terms
- We added further explanations on the linkage of the error types and the error sources
- We 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. ~~Better~~A better understanding of the reasons why hydrological model performance is ~~“good” or “poor”~~unsatisfying represents a crucial part ~~for~~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 ~~only distinguish between “good” and “poor”~~ provide a relative gradation of model performance. Especially in the case of a ~~“poor”~~weak model performance it is important to identify the different errors which may have caused such ~~unsatisfying~~unsatisfactory predictions. These errors may ~~origin~~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 ~~origin~~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 ~~the~~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 ~~errors systematically by mimicking~~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 ~~mimicked~~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 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 only a single numerical value within the range of minus infinity and one. A value close to one indicates a better model performance, whereas with increasing distance to one the model performance deteriorates. From this point of view, the model performance can only be assessed in terms of ~~"good" or "poor".~~a relative gradation. However, cases of ~~poor~~a weaker model performance immediately lead to the following questions: Why is my model performance not satisfying? 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. ~~Usually, diagnostic~~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. In this study, we propose a specific model evaluation approach 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. Remaining errors in hydrological simulations may then be caused by the following ~~origins~~sources:

- model parameters (e.g. Wagener and Gupta, 2005)
- model structure (e.g. Clark et al., 2008;Clark et al., 2011)
- 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 origin of the errors, we define three error types linking to 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; (ii) dynamic error; (iii) timing error. 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).

- ~~Constant error may have its origin in the input data or the model parameters. For example, errors may be caused by consistent input data error or by inappropriate model parameters causing consistent overestimation/underestimation.~~
- ~~Dynamic error may have its origin in the model structure or the model parameters. For example, structural model errors (e.g. deficient storage routine) or deficient model parameters (e.g. parameters of the storage routine) may cause dynamic errors.~~
- ~~Timing error may have its origin in the input data, the model structure or the model parameters. The error may be caused, for example, by input data errors and/or deficient model routines/parameters.~~

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

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

where $\overline{B_{rel}}$ is a measure for constant error, $|B_{area}|$ for dynamic error, and r for timing error. Similar to *NSE* and *KGE*, *DE* ranges from 1 to -∞. *DE* = 1 indicates perfect agreement between simulations and observations.

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) as a third term. $\overline{B_{rel}}$ reflects the constant error and is represented by the arithmetic mean of the relative bias (Eq. 2):

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

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):

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

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):

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

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:

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

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):

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

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. However, the aggregated value ~~does~~ only ~~allow~~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):

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

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):

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

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):

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

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):

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

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 ($DE \leq DE_l$) and a good simulation ($DE > DE_l$).

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

$$Diagnosis = \begin{cases} yes, & |\overline{B_{rel}}| \leq l \,\&\, B_{slope} > l \,\&\, DE \leq DE_l \\ yes, & |\overline{B_{rel}}| > l \,\&\, B_{slope} \leq l \,\&\, DE \leq DE_l \\ yes, & |\overline{B_{rel}}| > l \,\&\, B_{slope} > l \,\&\, DE \leq DE_l \end{cases}, \tag{12}$$

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

$$Diagnosis = \; timing\ error\ only, \qquad |\overline{B_{rel}}| \leq l \,\&\, B_{slope} \leq l \,\&\, DE \leq DE_l, \tag{13}$$

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):

$$Diagnosis = \; no, \qquad |\overline{B_{rel}}| \leq l \,\&\, B_{slope} \leq l \,\&\, DE > DE_l, \tag{14}$$

In this case, the model performance is sufficiently accurate and can be denoted as a good simulation.

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):

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

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.

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 ~~mimic~~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.

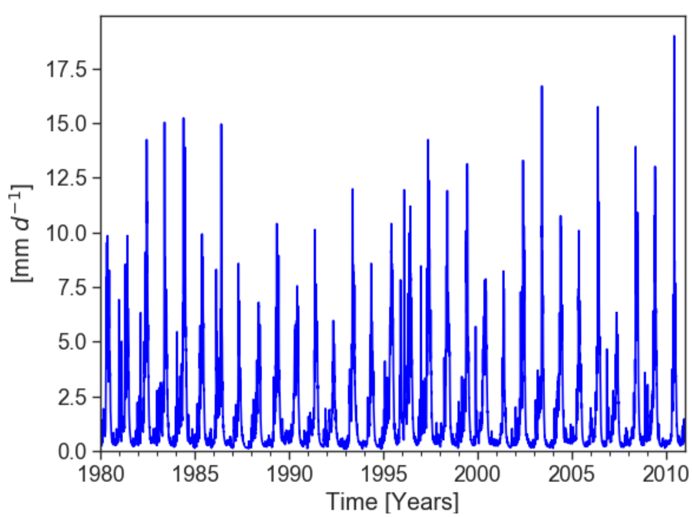


Figure 1: Observed streamflow time series from CAMELS dataset (Addor et al., 2017; gauge~~–~~_id: 13331500; gauge~~–~~_name: Minam River near Minam, OR, U.S.)

3.1 ~~Mimicking~~Generation of artificial errors

In the following section, we portray how we ~~generated the~~ manipulated observed time series ~~mimicking~~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 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 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 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 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.

~~**Table 1: Summary on mimicked 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~~

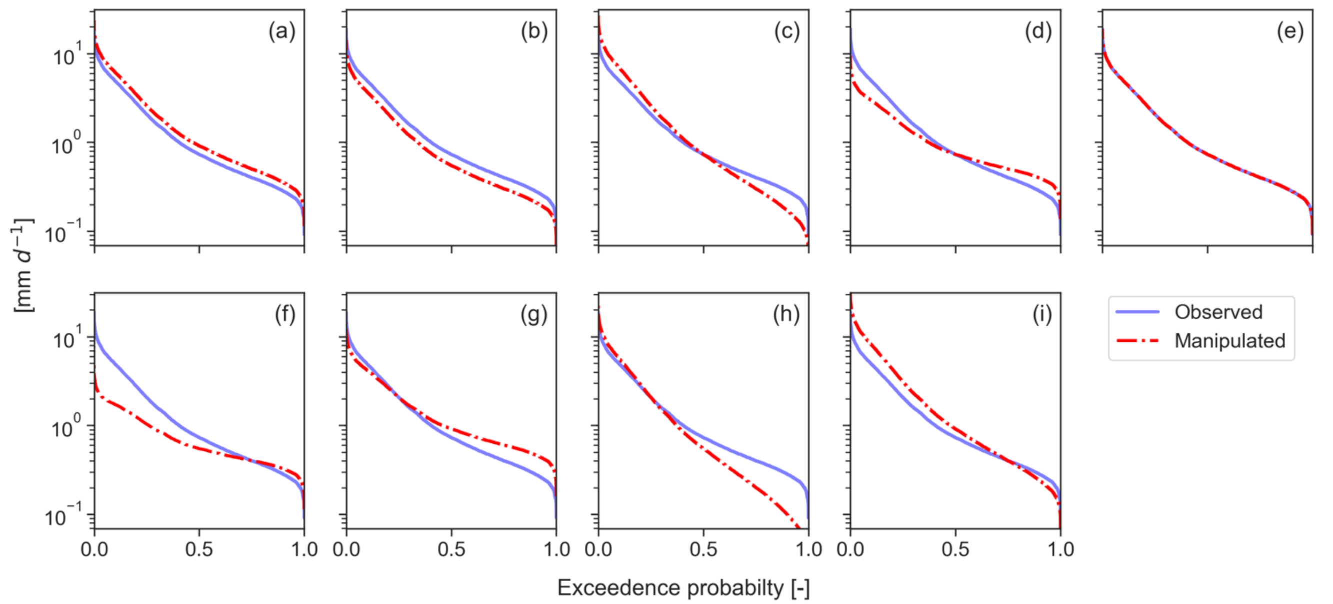


Figure 2: Flow duration curves (FDCs) of observed (blue) and manipulated (dashed red) streamflow time series. Manipulated FDCs are depicted for (a-b) constant errors only, (c-d) dynamic errors only, (e) timing error only, and (f-i) combination of dynamic and constant errors. The combination of constant 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 ~~mimicked~~synthetic error cases is shown in Fig. 3. Interdependently which error has been ~~mimicked~~generated, 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 ~~origin~~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 for individual errors (except for timing error) than for combined errors. Increasing the number of errors added to a time series, leads to lower *DE*. For the numeric values of the individual metric terms, we refer to Table S1.

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. ~~mimicking~~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. ~~mimicking~~generating single error types or multiple error types, respectively, leads to a non-equidistant decrease in performance). This non-equidistant decrease suggests that *KGE* and *NSE* are differently sensitive to the ~~mimicked~~generated errors. For example, lowest *KGE* values for single constant and dynamic errors are obtained by only introducing one error type (Table 2a-d). *NSE* is prone to timing errors (Table 2e), particularly to peak flows (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

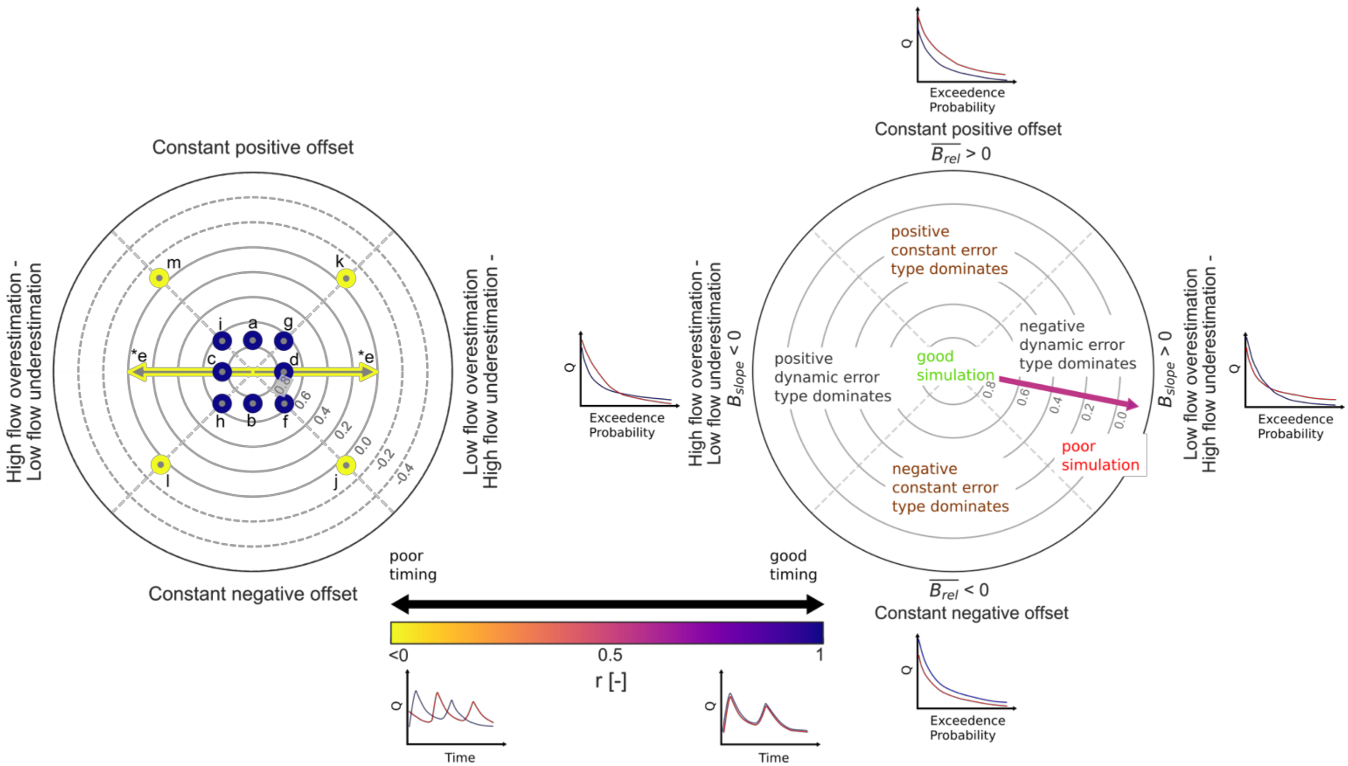


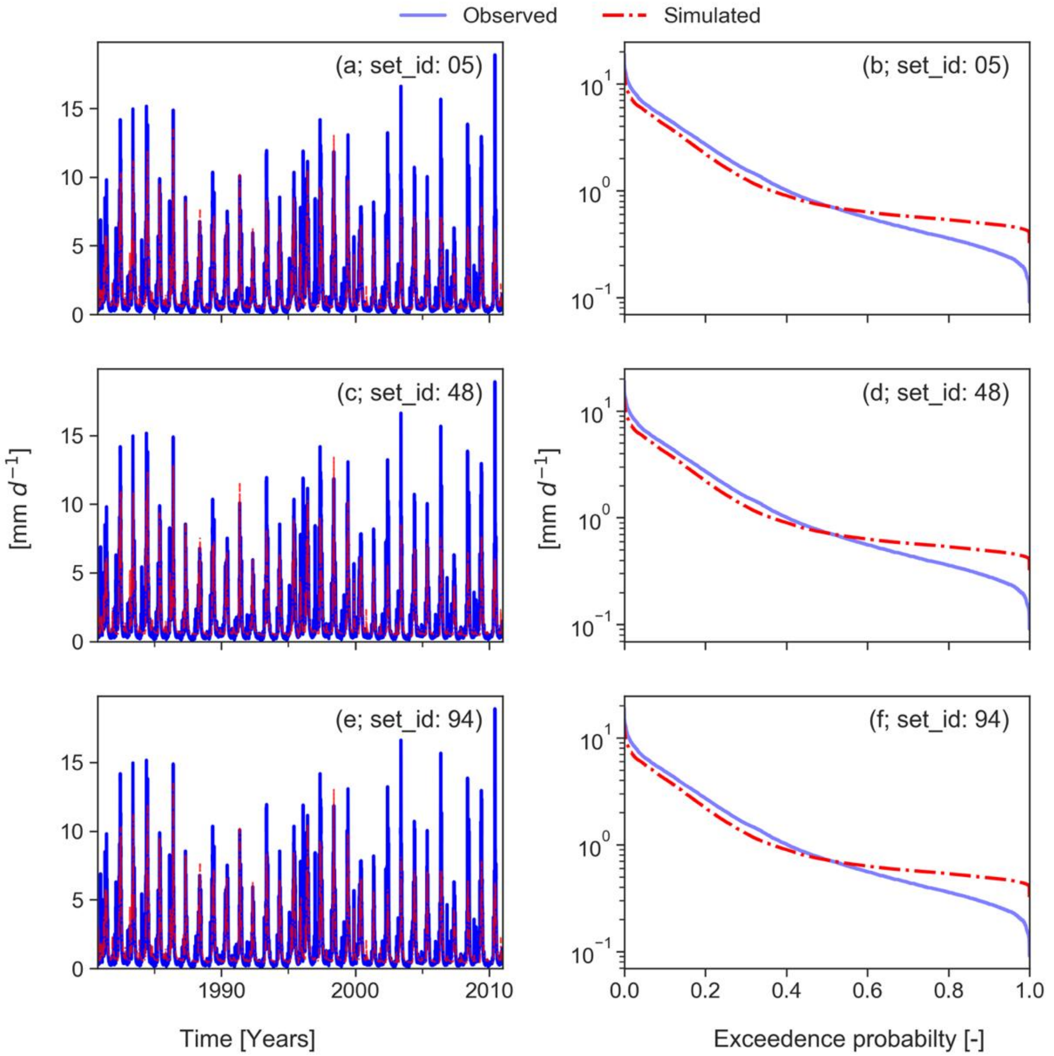
Figure 3: (left) Diagnostic polar plot for manipulated time series generated characterized by ~~mimicking~~ constant errors, dynamic errors and timing errors (a-m) visualizing the overall model performance (*DE*; contour lines) and contribution of constant error, dynamic error and timing error (purple (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 ~~generated~~characterized by ~~mimicking~~ constant ~~error~~errors, dynamic ~~error~~errors and timing ~~error~~errors (a-m). Lowest model performance for each error case is in bold.

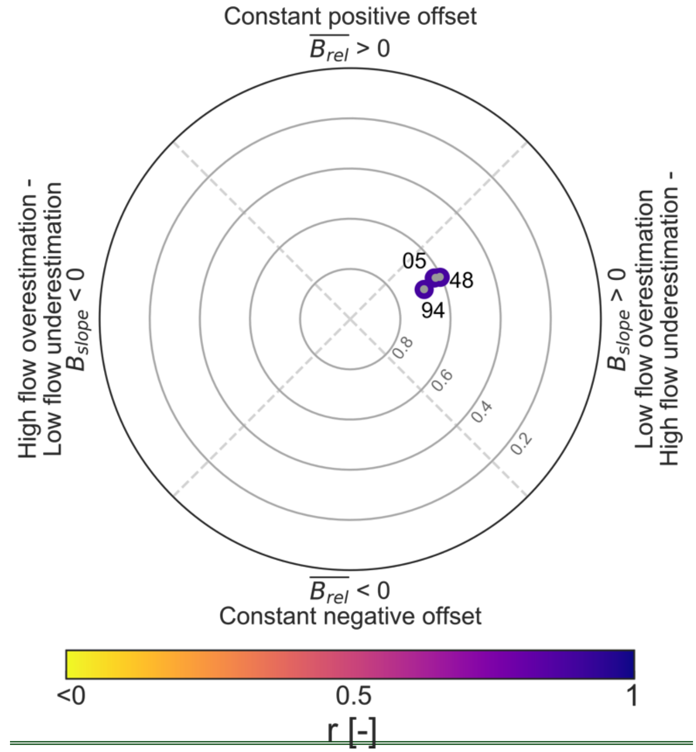
	a	b	c	d	e	f	g	h	i	j	k	l	m
DE	0.75	0.75	0.75	0.75	0	0.65	**0.65**	**0.65**	0.65	-0.06	-0.06	-0.06	-0.06
KGE	**0.65**	**0.65**	**0.43**	**0.43**	0	**0.08**	0.75	0.75	**0.08**	**-0.36**	-0.04	-0.04	-0.36
NSE	0.9	0.9	0.7	0.7	**-1**	0.27	0.94	0.94	0.27	-0.25	**-0.59**	**-1.58**	**-3.26**

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 airtemperature-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



~~**Figure 4: Simulated and observed streamflow time series of modelling example (a, c and e) and the related flow duration curves (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.**~~

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 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 ~~types~~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.

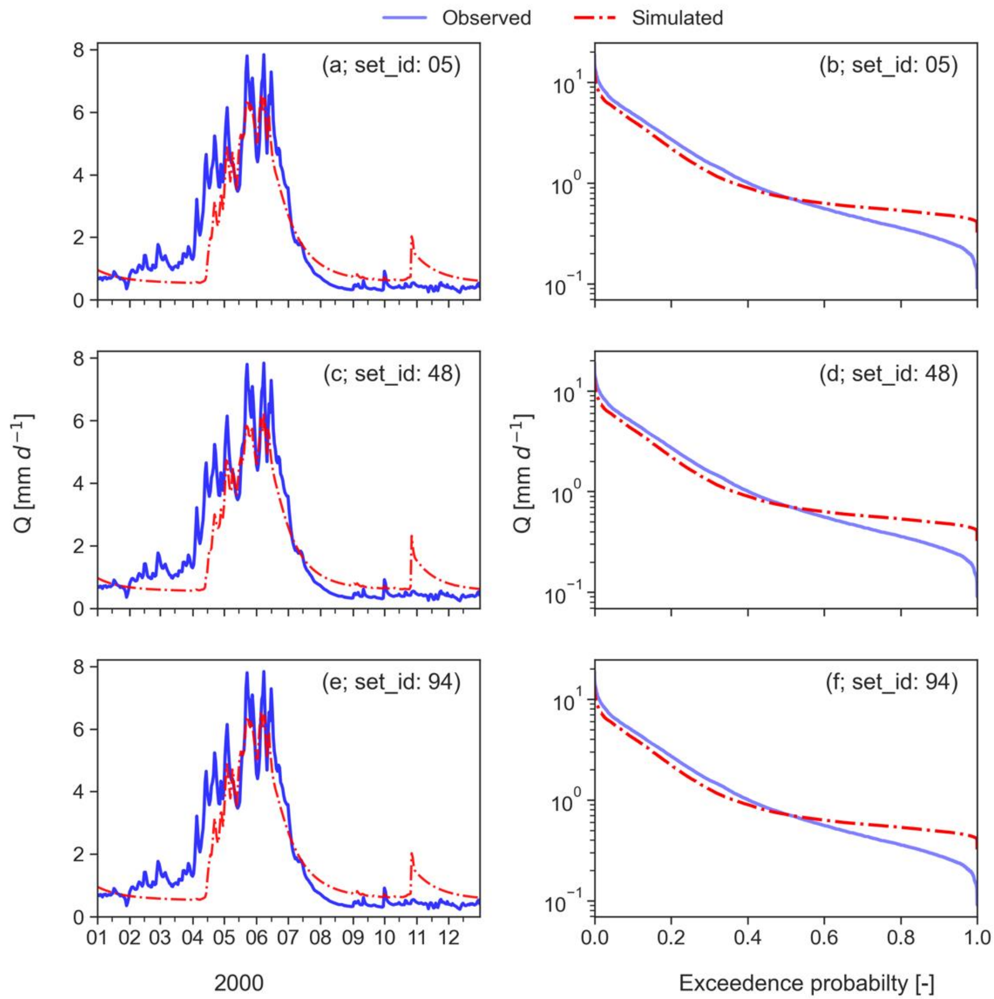


Figure 4: Simulated and observed streamflow time series of modelling example for the year 2000 (a, c and e) and the related flow duration curves for the entire time series (b, d and f). Time series are derived from the CAMELS dataset (Addor et al., 2017). Observations and simulations belong to the same 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).

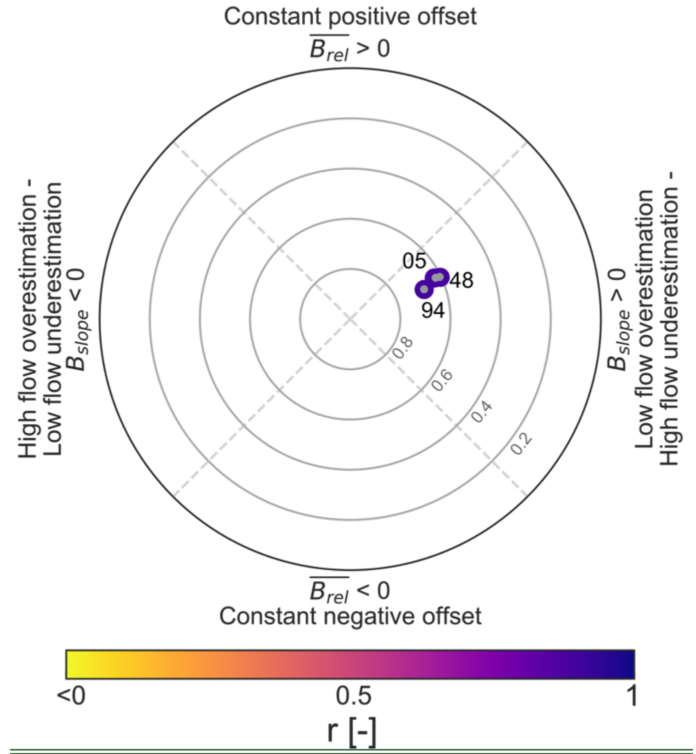


Figure 5: Diagnostic polar plot for modelling example. Simulations were realised with three different parameter sets (05, 48, 94; see Fig. 4). All simulations perform well. However, the remaining error is dominated by a negative dynamic error type while timing is excellent.

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 ~~mimicked~~generated errors, we found an illogical 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*: (i) metric formulation is based rather on a hydrological understanding ~~instead of~~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. 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).

Our approach focuses on model deficiencies. We do not propose a skill score measure for *DE* since skill scores 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), 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 and the metric terms into the evaluation process represents a major advantage. Although errors may have multiple ~~origins~~sources, 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 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. 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 ~~is based~~ 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.

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 ~~mimic~~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 comments on the language style and structure of the 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.

Appendix A

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

$$DE_{ext} = 1 - \sqrt{term_1^2 + term_2^2 + term_3^2}, \tag{A1}$$

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

$$DE_{multi-ext} = \frac{1}{N} \sum_{i=1}^{N} DE_{ext,i}, \tag{A2}$$

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.