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.