Reply to Referee #3

We thank the anonymous Referee #3 for the positive feedback and for providing so many detailed comments. The latter will help us to improve the quality of the manuscript. Here we provide a point-by-point list of author replies (AR) to all issues raised by the reviewer.

General comments
****

RC#3: Overall, I would advise shortening Sections 2.2 and 2.3 and adding some text to better explain and discuss the new aspects of SD that are introduced in this paper. In Section 3, I found very interesting the use of the proposed error to build uncertainty ranges. However, in Section 5.3, I don't understand why the uncertainty ranges are so different even though they were both built to achieve an 80 % coverage. My questions on this section are listed in the detailed comments.

AR: The sections 2.2 and 2.3 actually describe all new aspects of SD. For this reason we are not sure about where to shorten these sections. Could you please clarify? Our answers concerning the questions about section 5.3 are provided in the detailed comments.

Detailed comments
****

RC#3: Lines 151-162: transform this list into a table so that the reader better visualizes the new aspects proposed.
AR: Thank you for this suggestion. We will clarify the novel aspects and compare them to the previous version.

RC#3: Line 176: I think "low-flow" should be replaced with "low flows" when it is not an adjective.
AR: We will do that

RC#3: Line 200 : "this approach has been shown to work well" : could you specify ?
AR: This is difficult to show in a simple graph or statistic. Rather, this is the authors' perception after applying the SD approach to many different discharge time series in both the 'many events' and 'single, long event' mode. 'Shown to work well' in this context means even in the 'single, long event' mode, SD linked parts of obs and sim time series that visually appeared to be matching segments within matching events. As we provide the SD code and test data together with the article, this can easily be tested by interested users. We will include these two points in a revised version of the manuscript.

RC#3: Lines 262-263 : I recommend removing "(or vice versa, if a falling segment is dissolved)". The sentence line 261 follows the example on line 249, and I don't think the information in parenthesis necessary to understand as it is already clear. More generally, I would recommend removing most parenthesis in the article and either include the content in the text or remove it. This could further improve the reading of the article.
AR: Thank you for pointing this out. We will revise the article again with respect to the wording and remove words in parenthesis wherever possible.

RC#3: Lines 268-271: "Since the segments can differ in length. . .between the time series edge nodes." Here, you say that you interpolate between edge nodes to obtain the same number of points to compare between the obs and sim, and, in the case of the obs time series, each edge node

corresponds to an actual observation. Is that right? If so, my question is: if you linearly interpolate a segment between edge nodes in the obs time series, won't you create "fictive" observations by filling a segment? And wouldn't this "falsify" the evaluation of your simulation time series? Could you clarify this?
AR: You did understand right. Due to the linear interpolation we obtain "fictive" observations. We do however not understand why this should falsify the evaluation as the same procedure is applied to the simulated time series. The idea of the SD method is to identify and to compare points which are "hydrologically similar". This essentially requires to compare points which do not share the same abscissa and which do not coincide with the hourly observations.

RC#3: Line 290: "In this case," a space is missing after the comma.
AR: we will add the space.

RC#3: Line 317: "obs an sim", "an" misses the final "d".
AR: we will add the missing "d"

RC#3: Line 367: Shouldn't the uncertainty ranges capture a significant portion of "observed" values rather than "simulated" values?
AR: This is of course correct. We will change it in the revised version.

RC#3: Sections 3.1 and 3.2: I am not sure I understand how you use the relative contribution to the total error to select your sample. For instance, in the one-dimensional case, do you assume that your uncertainty range is symmetrical around your simulated value? Since the relative contribution is positive, do you pick the value v that corresponds to 80% of the error and use it to build both the lower (-v) and upper range (+v) ? If not, the upper left graph of Figure 4 may have misled me. Or do you directly use the error values behind the relative errors within 0 and 80% and select the largest negative one to build the lower range and the largest positive one to build the upper range?
AR: The second case is correct. We directly use the error values behind the relative errors within 0 and 80% and select the largest negative one to build the lower range and the largest positive one to build the upper range. We do hence not assume that the uncertainty range is symmetrical around the simulated value. Please compare also the results from the case study (fig. 5) where this is the case. We do however agree that the sketch in Fig. 4 is a little bit misleading. There reason for this is that the sketch in Fig. 4 is, for the sake of simple presentation, based on "normally distributed" random numbers which do of course yield symmetrically centered envelopes.

RC#3: Line 398: Don't all ds2 sum up to 100?
AR: This is correct, we will change it.

RC#3: Line 449: Remove one of the "period"
AR: Of course.

RC#3: Lines 552 and 773: The reference for Ewen is not formatted properly
AR: You are right, we will correct the formating.

RC#3: Line 661: I am not sure about the use of "relevant" here.
AR: We agree. Maybe "important" fits better. We will change it.

RC#3: Lines 670-672: Just a question, in your opinion, could this be addressed by applying the dressing to adapted time steps? E.g. so that the distance between two dressed time steps is equal to the error in timing? Would there be an optimal time step to apply the dressing (one for which the percentage of sampled errors and the overall coverage would match)?

AR: We are not sure if we correctly understand your question. We are however grateful for asking it as this points needs to be clarified. We expect that there are different reasons for this problem: The most important ones are probably in the selection of the error model and in the definition of the error distributions. In the manuscript we applied SD using a "relative" magnitude error model but an "absolute" horizontal error model. The result is that the same (static) timing error is applied to all time steps. One possible way out would be to formulate the timing error in a relative way, e.g. using average event duration. This would very likely narrow the timing errors. Another alternative would be to further differentiate the error distributions according to streamflow magnitude since we do not expect that the same timing errors do occur at each flow rate. Last, the current version of SD does not account for the auto-correlation of the errors which is typically high in streamflow data. Considering it in both, the calculation and in the application of the errors would very likely narrow the uncertainty envelopes to a significant degree. We will note on these aspects in the revised version of the manuscript.

RC#3: Section 5.3.2: (1) From Figure 6, it is not intuitive that the SD and BM approaches have the same coverage. Is it specific to this example? Do you have cases when there are more observations outside the SD envelope than outside the BM envelope that would compensate for the case you show? (2) Based on how the SD envelope is constructed, you would expect a third of the 20 % outside the envelope occur in low flows, a third in rising limbs and a third in falling limbs. On the opposite, the 20 % falling outside the BM envelope can occur whenever in the time series, and can, for instance, always occur when it is harder to model streamflow, i.e. during events. Does this have an impact here? How does that affect the results?

AR: Good point. The coverages of SD and BM shown in the plot are indeed different. The reason is that the plot contains only a subset of the entire time series: While for the entire time series the overall coverages are indeed equal for SD and BM, the values can deviate for subsets of the time series. For other subsets such as very small events classified as low flow the effect is opposite: Here the 1-D error distributions are applied to the simulation for both SD and BM. These differ however in their extent (compare the lower two panels of Figure 5), which causes that BM has a "higher" upper envelope than SD during low flow. We agree that the case depicted in Fig. 6 is not intuitive with respect to coverage and we will add an explanation to clarify this in the revised version of the manuscript.

RC#3: Lines 700-702: From this sentence, it may seem that low flows also have a 2-d error distribution, but errors in low flows are 1-d. Could you clarify this?

AR: Of course. Thank you for pointing this out. We will change the manuscript accordingly.

RC#3: Lines 719-720: What would be the difference between “uniting” and “intersecting” in this case?

AR: These terms refer to set theory. Intersecting means to use only "elements common to both error components" whereas uniting would mean to use all elements, which is what we have done at the moment. We will clarify this in a revised version of the manuscript.

RC#3: Line 723: “as proposed by” I believe a reference is missing here.

AR: This is a wording error. We will correct it.

RC#3: Figure 5: The dot of the sampled subset in the upper left graph is black whereas the sampled subset is orange.

AR: You are right, we will correct this