Reply to Referee #2

We sincerely thank Referee # 2 for the thorough review and for providing excellent suggestions. Our manuscript will considerably benefit from it.

RC#2 elaborates on two general aspects:

A) Coarse-graining, i.e. our pattern matching procedure which breaks the simulated and the observed hydrographs into segments (step 1) and conducts the matching and comparison of the two time series using "fine scale linear interpolation" as RC#2 put it (step 2).

RC#2 correctly describes the calculation of the error in SD using the Equation R1.

The major point of criticism raised by RC#2 with respect to coarse-graining is that "there is simply too little discussion, exploration and testing of the coarse graining algorithm in the manuscript (...) to be scientifically sound, the coarse graining algorithm needs to be explored fully for several types of hydrograph, and tested properly, in detail, against an appropriate benchmark".
Later in the review RC#2 provides an important additional note which is "that the use of linear interpolation in the fine-scale pattern matching adds to inflation because it neglects local information about timing. Rather than depending on local (i.e. within segment) timing, Eq. R1 shows that the local timing errors are assumed to depend only on the t and T values generated in coarse graining."

Our reply to this major point is:
1) We do agree that the coarse graining algorithm has not been fully explored yet. The analysis and evaluation of coarse graining in quantitative terms is however very difficult as the breaking of the hydrographs into segments is subjective and thus, to a certain degree arbitrary. Since there is no reference for the breaking of the hydrograph available it is difficult to meaningfully compare coarse-graining to a benchmark approach. Essentially, only a visually comparison seems meaningful to us. This would however require many plots of different hydrographs and flow regimes which is difficult to realize within in a paper or supplement. For this reason we decided to provide software such that any interested reader can find out for him/herself whether the proposed method suits his or her needs.
2) RC#2 is correct that the accuracy of the SD procedure crucially depends on t and T, and thus, on the accuracy of the coarse-graining. Poor coarse-graining yields incorrect matched segment pairs (compare for instance those in the upper panels of Fig. 3). These inevitably introduce large timing and possibly even large magnitude errors and may thus also add to false inflation. For this reason we clearly state that coarse-graining is of fundamental importance for SD (line 542): "...the quality of the segment matching largely determines the quality of the subsequent matching of obs and sim points and hence the quality of the SD error calculation".

Due to the large importance of coarse-graining we propose to illustrate the sensitivity of the weighting factors in the objective function (Eq. 2) in more detail in the revised version of the manuscript.

B) The second major point of criticism refers to the "error dressing" concept which is regarded unconvincing by RC#2. As reason RC#2 states that only "few results for error dressing are shown and that the method introduces considerable false inflation (overly large error clouds), which makes the clouds difficult to interpret physically and limits their usefulness operationally." RC#2 also raises the

questions whether the entire section on error dressing should be removed from the manuscript to make more room for demonstrations and discussions on coarse graining.

Our answers to this major point is:

1) We consider the joint visualization of timing and magnitude errors a fundamental and important part of the manuscript as this is rarely done in hydrological papers.

2) The reasons for false inflation are manifold and cannot be attributed to error dressing alone. Possible causes for false inflation include: i) poor coarse-graining results as these will yield large timing errors, ii) poor definition of the error pools where the samples are taken from in error dressing, iii) the use of an absolute timing error model (Eq. 6) instead of a relative one like for the magnitude errors and iv) the neglecting of the local, within segment, auto-correlation of timing errors. Each of these sources can cause large timing errors and thus, contribute to false inflation.

For these reasons we would prefer to keep the section on error dressing as it is. We provided it as one possible application of SD although there are of course more elaborate methods available. We propose however to discuss the reasons for false inflation in more detail and to specify possible ways forward.

Other points

****

RC#2: (Line 25) The word “elaborated” does not work her

AR: We don't see the problem of the use of “elaborated” in this context. Could you please specify the reason why it does not work. Change to "detailed", "sophisticated", "ambitious",???

RC#2: (Line 241) Is some normalizing factor or term needed here to make ISEG sum to unity

AR: ISEG sums to unity as it is calculated based upon the relative duration ($dt*(i)$), and the relative magnitude change ($dQ*(i)$). We obtain $dt*(i)$ and $dQ*(i)$ by normalizing $dt(i)$ by the total duration and $dQ(i)$ by the sum of the absolute magnitude changes of the entire event. Equation 1 does hence essentially sum up to unity.

RC#2: (Line 245) This process seems to reduce the number of segments by two. What happens if an odd number needs to be eliminated?

AR: RC#2 is correct, each step of aggregation reduces the number of segments by two. It is however not possible that an odd number of segments needs to be eliminated. During the pre-processing we trim the time series, if required, to ensure that both hydrographs do start and end with an identical segment type, i.e. either a rising or falling segment.

RC#2: (Line 254) It is not entirely clear why the name “coarse graining” was chosen, especially given that this required citing two otherwise irrelevant papers about other things that are commonly called coarse graining.

AR: Good point. We spent a lot of time thinking about a name that has least overlap with existing procedures and found this to be the best. However, we like the suggestion by the referee and will change the wording to 'pattern matching procedure (PMP)' in a revised version of the manuscript.

RC#2: (Line 330) What is done if there is no local minimum in the objective function

AR: The minimum is taken in every case independent if it is a local minimum or not. Local minima occur in the coarse graining of "complex" multi-peak events with large number of coarse graining steps. In "simple" events where no or little coarse graining is required the objective function values often increase fairly linear. We will add a brief comment on this issue at line 330.

RC#2: (Line 454) It is the gold standard in work like this to use split-sample testing because this is the best way to test how the method would work when used operationally. Split-sample testing is trivial to apply, so there seems no reason not to use it here. The defence that the example is used simply to aid “discussion of the SD concept” (line 453) is very weak.

AR: We agree that split-sampling is widely used in this context. However, we do not see the benefit in providing additional information on the method by means of using split-sampling here. Again, the main purpose of our case study is to illustrate the method using real-world data and not whether the achieved coverage differs by some percentage between "calibration" and "validation".

RC#2: (Line 561) An advantage is claimed for SD that “unique relationships of points in obs and sims are established”. This advantage, however, comes from using linear interpolation, which, as discussed earlier, comes at the cost of neglecting local (within segment) timing information.

AR: In our perception it is important to perceive the temporal order of both the simulated and the observed values. For this reason we uniquely compare the first point in obs to the first point in sim, the second to the second, etc. (compare lines 206-211). It is this assumption which justifies the use of linear interpolation, not vice versa.