Authors Response to Referee #3 (Jim Stagge)

Here the authors address two unique research questions. First, the authors define a multi-objective approach to calibrating a hydrologic model to consider low flows, high flows, and water balance. Second, they use this approach to reconstruct flows for rivers throughout the UK beginning in the 1891, made possible by recovered meteorologic datasets.

The paper is well-written, of strong interests for HESS readers and a novel piece of research. I have some concerns about a general lack of reference to the hydrologic calibration literature, particularly with relation to prior multi-objective approaches. The authors' application is certainly novel and they made choices to weight their multiple objectives a priori, which is a realistic approach when repeating this for many watersheds. However, there are more advanced multi-objective schemes that should be mentioned for context (and potentially for follow-up research). Because of this weighting approach, there must be some discussion of how the objectives are related to one another and how these weightings affect results.

Overall, I recommend this article for publication pending the major revisions to provide a better literature context and to better explain the objective weighting scheme's effects.

We thank you for your kind words Jim, and are glad that you deem the research novel and of strong interest to HESS readers. We appreciate your concern for the current lack of reference to the literature regarding multi-objective calibration procedures, and will ensure that this is addressed in the revised manuscript.

Major Comments

1. I have a concern that there is a wide body of calibration/optimization literature not being referenced in this paper. Many approaches have been used for hydrologic model parameter calibration, and although the paper mentions some, there are gaps that could put this work in context. I suggest to at least mention PEST, which is a single objective optimization scheme, but almost ubiquitous in the U.S. hydrologic community. Wallner (2012) "Evaluation of different calibration strategies for large scale continuous hydrological modelling" provides a good overview of these calibration strategies.

Thank you for noticing this oversight, we will insert reference to this area of research in to the introduction, and methods sections.

2. Although the words "multi-objective optimization" aren't often written together in the text, this approach appears to be an a priori multi-objective optimization. By using the sum of each objective's rank as your objective, you have defined weightings a priori to merge multiple objectives into a single objective function. Please include at least one or two sentences explaining this and mentioning the difference between this and a posteriori multi-objective optimization (below).

Please see our response to point 3. below

I mention this because you state that "multi-objective optimization methods have been advancing since the turn of the century", but this area has a pretty rich literature that goes back well into the 1990s. Additionally, most optimization researchers think of a posteriori (not a priori) when they think of multi-objective optimization. A posteriori approaches try to find a set of non-dominated Pareto optimal solutions and then select the best compromise afterwards. You might include references to other multi-objective papers that take this approach like:

"Multiobjective Automatic Parameter Calibration of a Hydrological Model" (Jung et al, 2017) "Comparing multi-objective optimization techniques to calibrate a conceptual hydrological model using in situ runoff and daily GRACE data" (Mostafaie et al. 2018) "Automatic calibration of HEC-HMS using single-objective and multi-objective PSO algorithms" (Kamali et al. 2013) "Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm" (Shafi and de Smedt 2009)

Or consider some of their references for older publications.

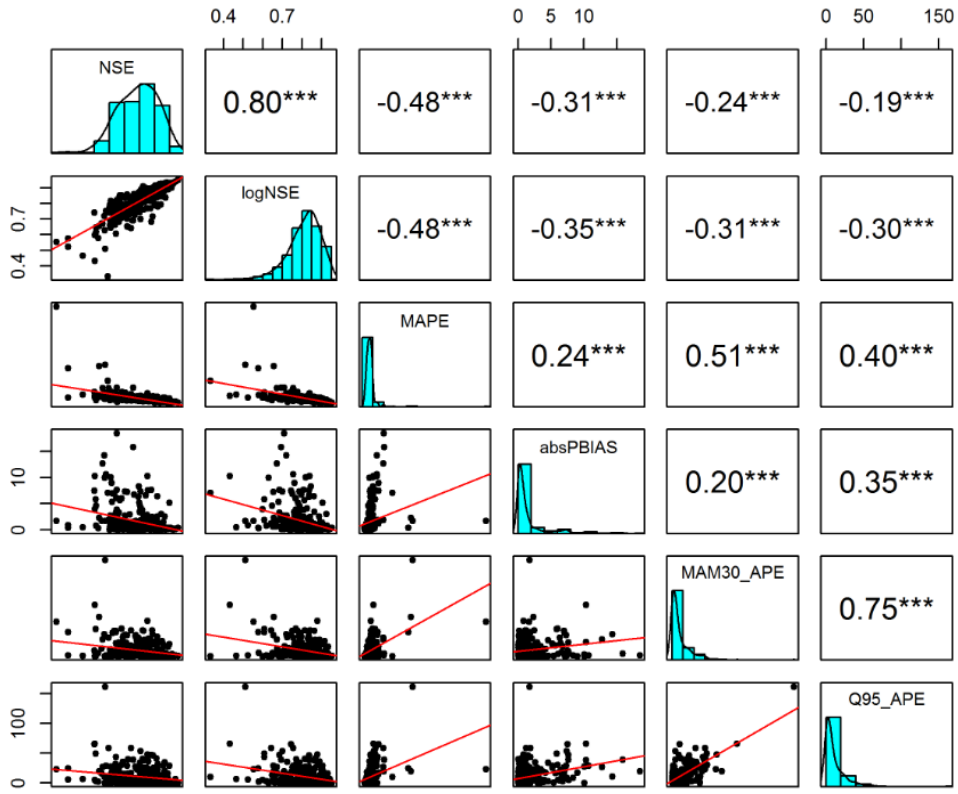
Yes we can see that this area of literature has been overlooked in the manuscript. We will add a few sentences on these approaches to the introduction.

3. Because of the a priori weighting (Comment #2), please provide information about how the multiple objectives are related to one another. Are some highly correlated? Negatively correlated? If, for instance, the rankings from the 4 high/water balance objectives operate as one and the 2 low flow indices operate as one, is there a concern that you are overweighting towards high flows?

We are not sure whether our approach would be considered a priori or a posteriori. Traditionally, a GLUE type approach would assign an a priori distribution to sample parameter values from, we chose to make no a priori assumptions and chose to sample from a uniform distribution across all 4 parameters. A GLUE approach would then weight the "behavioural" parameter sets a posteriori according to their metric scores. We have chosen 6 metrics, and have not "weighted" the runs by their scores, merely extracted the top 500. Yes, the 6 metrics we chose could be implicitly weighted according to their similarity (if two metrics were very similar, then they would hold more weight together than any one of the others). Thus, the graphic below demonstrates a quick look into their correlations. We have taken the LHS1, the "best run", for each catchment here, aside from the fact R is not easily capable of reading in 303 matrices of 500,000 rows, the sets of 500,000 for each catchment contain some truly awful parameterisations, and the metric interactions among these were understandably very odd. Here, we can generally see that there are significant correlations between each of the metrics: generally where catchments score well for one metric, they score well for all metrics. Bear in mind that for NSE and logNSE a high score is a good score, whilst for the other 4 a low score is a good score. The NSE and logNSE have the highest correlation, which would be expected, and the $MAM30_{APE}$ and $Q95_{APE}$ are also highly correlated as they are both errors in low flows. Correlations between MAPE and absPBIAS are positive with each other, and the low flows metrics, and negative with the NSE metrics. MAPE and $MAM30_{APE}$ are quite strongly correlated, as they are both mean percent error metrics. The metrics were carefully chosen to cover different elements of goodness of fit as follows:

- *NSE – good at magnitude and timing of peak flows*
- *LogNSE – NSE on log flows in an attempt to match magnitude and timing of lower flows*
- *MAPE – overall magnitude of variability*
- *absPBIAS – total water balance*
- *$MAM30_{APE}$ – error in the lowest of flows*
- *$Q95_{APE}$ – fitting the tail of the FDC.*

We would say that, if anything, these 6 metrics are together slightly more biased towards matching low flows than high flows, which we were happy with given their intended purpose for use in drought research.



4. Line 245-250: I find it surprising that there is a single very poor fit among nearly perfect fits, for example in Cornwall. As you are mentioning the reasons for poor fits in this paragraph, it is important to mention there does not appear to be a spatial pattern. Presumably, the same abstractions and groundwater issues affect the 0-10% threshold poor fit as its > 90% good fit neighbours. Are there any other feasible explanations?

We would argue that there IS a spatial pattern, there is generally good performance across the country, with the exception of two areas:

- *some upland catchments in Scotland and Northern England that experience snowmelt contributions, and*
- *highly permeable catchments or those with significant human influence in south and south-eastern England. The more local scale variability across the south is likely due to the spatial variability in the geological units.*

You have identified one exception to these two broad categories, which is the Warleggan in Cornwall. This catchment fails the thresholds due to the Nash Sutcliffe Efficiency metric alone: the peak flow magnitudes are significantly underestimated, the other metric scores are acceptable. This could be due to the fact that the catchment sits on a granite outcrop, so is less permeable than surrounding catchments, but the calibration process ought to be able to account for this; it would require further investigation to identify the cause of this specific insufficiency. We will add a comment about this exception to the manuscript.

Minor Comments

Line 45 – Suggest 1 or 2 more references to fill out the discussion of low flow climate projections for the UK.

We have added Wilby and Harris (2006), Christierson et al (2012), and Prudhomme et al (2012)

Wilby, R. L., and Harris, I.: A framework for assessing uncertainties in climate change impacts: Low-flow scenarios for the River Thames, UK, Water Resour. Res., 42, W02419, 10.1029/2005wr004065, 2006.
Christierson, B. v., Vidal, J.-P., and Wade, S. D.: Using UKCP09 probabilistic climate information for UK water resource planning, Journal of Hydrology, 424-425, 48-67, https://doi.org/10.1016/j.jhydrol.2011.12.020, 2012.
Prudhomme, C., Young, A., Watts, G., Haxton, T., Crooks, S., Williamson, J., Davies, H., Dadson, S., and Allen, S.: The drying up of Britain? A national estimate of changes in seasonal river flows from 11 Regional Climate Model simulations, Hydrological Processes, 26, 1115-1118, 10.1002/hyp.8434, 2012.

Line 70 – You may want to mention some proxy-based reconstructions; for example Jones et al (1984) "Riverflow reconstruction from tree rings in southern Britain" or the Old World Drought Atlas (Cook et al 2015) "Old World megadroughts and pluvials during the Common Era" which covers the UK.

We have added these references.

Line 193 – Please define LHS500. This is the first time it is included in the text (only in the abstract).

The first reviewer also noticed this error, we have amended it to "The upper and lower daily limits of the 500 top ranking parameterisations (see Section 3.4 for details on the ranking process) were used to calculate…"

Table 2 – If possible, please try to fit the ranges on a single line of this table.

We've corrected this

Lines 273 – You do a great job of describing a low UncW and low ContR as biased and under-sensitive - this is a helpful translation for readers. As a reader, I would also like a description of the converse. What does high UncW and high ContR mean?

We will add a sentence to this effect.

Line 344 - Can you provide a description of which objective function(s) is driving the best fit parameter set in the Avon to consistently overestimate low flows?

We will look into the parameter values of the best run compared to the other LHS500 members, as well as the metric scores to see if we can notice anything here.

Line 372 – Please add the words "we consider" before "SSI values. . .". The thresholds of -1 and -1.5 are largely arbitrary and more of a convention than a true definition.

Valid point, we have added this to the manuscript

Figure 9 – Please mention that you are plotting the mid-point of each event in the caption. It is currently only in the text (Line 417).

We have added this to the caption

Figure 9 – For the Crimple watershed, there are 3 unique drought events for the Modeled data shown in the period 1975-1979. But Figure 8 shows only 2 crosses of the -1 threshold. Please confirm what is going on here.

The three modelled data circles suggest some discrepancy in the timing of the 1975/76 event among the LHS500, rather than 3 distinct events. The timing of the drought events is characterised by the dates the SSI crosses 0 (though we're only showing the events in Fig 9 where at least one month crosses SSI -1.5). The individual LHS500 runs demonstrate quite some width in the ascending limb as the SSI crosses into positive values in the Crimple in 1976/1977. This discrepancy in the end date of the drought event will affect its midpoint, and from Fig 9 it looks as though the LHS500 are grouped in to 3 main possibilities for timing. However, the thickest circle (demonstrating a higher number of runs) is the central one which best agrees with the timing of the observed event. The Greet and the Bush also show many circles for this event, and also have a wide band of grey LHS500 runs as the SSI crosses 0 in Fig 8. The Bush in particular doesn't cross back above SSI 0 until 1979 for some of the LHS 500 runs. We will re-read this section and make sure that it is clear that overlapping black circles suggest timing discrepancy rather than multiple events.