I have reviewed all the contents of the document and found it to be clearly written. This work applies a multi-objective optimization strategy to a two zone temperature and solute transport model to determine which data sets improve calibration and provide the most information to effectively narrow the parameter ranges.

I have a single, major concern about this manuscript and several minor concerns.

MAJOR CONCERN:

As stated in lines 10-23 on p. 8313 the goals of the manuscript are to use the proposed approach to gain insights regarding 1) parameter sensitivity and parameter bounds, 2) importance/worth of data used for calibration, and 3) prediction uncertainty.

It seems that one of the key findings of the proposed approach is that improved parameter estimates (as measured by the extent of the plausible range of parameter values) can be obtained by including multiple data sets. These results should not be surprising and is well established in the literature on inverse modeling. In particular, I am wondering how many of these insights listed above could be obtained by conventional (and computationally more efficient) local sensitivity analysis, which has been used for decades to make similar assessments.

For example, the results in Figure 10 would correspond to parameters that showed higher sensitivity (that is, parameters with greater sensitivity typically would have narrower bounds). To include the effects of parameter correlation, one could look at the parameter variance, as obtained by the variance-covariance matrix of the parameter sensitivities. Parameters with lower variance have corresponding narrower parameter ranges.

To assess the worth of the various data types, plots of the sensitivity to each parameter could be examined. For a composite measure, integrated metrics like Cook's D or similar influence measures indicate which data have the most influence on calibrated parameter values. Furthermore, examination of the residuals after calibration would indicate which data sets were better fit than others.

To assess the importance of data on prediction uncertainty, measures such as OPR and PREDUNC available in inverse modeling software UCODE and PEST, respectively, can provide these insights. Specifically, they measure the effect of data (current or potential new data) on the variance of predictions of interest.

Finally, many of these local sensitivity techniques are described by Hill, which is listed in the references, but not actually cited in the body of the manuscript. This makes me wonder if this reference, or other local sensitivity analysis literature, was consulted.

MINOR CONCERNS:

Below I list minor concerns that are intended to improve clarity of the manuscript.

1) P. 8311, lines 19-20 – Tradeoff between what? Implies that at least 2 factors need consideration, but it is unclear which factors.

2) P. 8311, line 23 – Unclear. Global optimum does not exist or finding it is unrealistic?

3) P. 8311, line 24 – There is no guarantee that the local optimum "bound" the global. Note that on p. 8318, line 65-8 local vs global is defined. Perhaps this should be moved here (p. 8311) to make clear.

4) P. 8314, lines 5-10 – Q1 = 1.06; Q3=1.96; Qgain = 0.17; This does not add up. Is the rest coming from the pond?

5) P.8315, lines 1-5 – How about calculating a percent mass recovered. That would be a direct measure of loss.

6) P. 8317, line 4 – 11 parameters (not 10)

7) P. 8317, eqn. 1 – What is 'N'? Number of data of each type? Also, how do you calculate the mean when the data have different units? Some additional details on the use of this equation should be provided.

8) P. 8318, line 5 – to do this you would actually have to compare different calibration algorithms, the results of which were not reported in this study. Were different algorithms compared?

In summary, this is an interesting study and will be of interest to users of hydrologic models, but I am concerned about the insights gained via the proposed approach. Specifically, I am wondering if the major insights can be obtained using standard measures based on local sensitivity analyses, which require much less computational overhead as the proposed method. Lack of discussion of local sensitivity methods and how this work relates to those methods is a shortcoming of this manuscript.