Author response to Referee #1 comments:

We thank reviewer 1 for the detailed and careful review of our work. We hereby provide our point by point responses how the comments by referee #1 will be addressed in the revised manuscript.

Best,
Hanwu Zheng

Anonymous Referee #1
General comments:
The study falls within the scope of HESS and is well written, with clear structure and fluent language. The quality of the figures is mixed and the methods used were insufficiently robust to provide any confidence in the generalizability of the results or conclusions. The study is broadly similar to several previous publications on multi-objective optimization using isotope tracers, and the new contribution, beyond replication of previous findings in a new location, is not yet clear. With revisions, this could be an excellent publication for HESS.

Reply: *We apologise if it was perceived that the quality of the figures was mixed: we will revise all accordingly following the reviewer's suggestions (see details below). The methods will also be more clearly described according to the specific comments, but we do feel these are robust. We acknowledge that such multi-objective optimization using isotopes and streamflow was successfully used in previous publications. We do not claim that using this method is the main novel aspect of this study, but rather we build on these previous applications to confirm the robustness of this approach. However, the performance of this methodology applied in large catchments under intensive management has not yet been studied. We will make this contribution clearer in the revision. In retrospect, we can see that that the validity of such sparse seasonal isotope data in constraining hydrological processes in such a large ET-dominated catchments, with heterogeneous land use has not been clearly explained. and we will improve this in our revision. The new contribution of this study rests mainly on this application and will be clarified in revision. We anticipate that this will realise the potential of the paper that the reviewer kindly acknowledges.*

Specific comments:
I see three areas in need of substantial revision: study differentiation, calibration methodology and presentation of results.
The study looks quite similar to previous studies in other areas, some of which have not yet been referenced in the introduction or discussion; multi-objective optimizations using flow and isotopes have been coming out for many years, e.g.: (He et al., 2019; Holmes et al., 2023; Nan & Tian, 2024; Tafvizi et al., 2024; Tunaley et al., 2017). The novelty is currently unclear, and the authors should revise to highlight the specific aspects that are new (this will likely involve only minor changes to the text). Is it the study site (agricultural with substantial groundwater pumping) or the spatial discretization of the model? Or something else, perhaps relating to the analysis of the results?

Reply: *Thanks for these suggestions and we agree. The papers related to multi-objective optimizations using isotopes and streamflow will be referred in the revision. We will also highlight the actual novelty of the paper much more clearly in the revision: 1. We show the value of streamwater stable isotopes in improving understanding of hydrological processes in in large, intensively managed catchments; 2. We also show that even a sparse dataset of isotopes is valuable in constraining streamflow and ET partitioning in such heavily managed systems, although limitations exist (which we discuss); 3. We demonstrate that epistemic uncertainties from unrecorded human activities can be identified by trade-offs between streamflow and isotopes in hydrological calibrations. These novel contributions will be clarified in the revision.*

A more fundamental issue with the present version is the methodology applied. Given the central importance of calibration to the study, the methods applied are not as robust and defensible as they ought to be for a publication. In particular:
The model was calibrated to optimize NSE. This metric has lost support as a calibration objective because as a squared error metric, it overemphasises peak flow timing, and leads to erroneously damped simulation variability (Gupta et al., 2009). Unsurprisingly, the presented model results had erroneously damped variability (low flows too high, high flows too low). Further, for sparse datasets (like the isotope series here) it is highly sensitive to individual points, as noted in the text. Why was this metric used in spite of its well-known deficiencies?

Reply: *Whilst we recognise that the limitations of NSE have been increasingly acknowledged, it is still widely used in the hydrological modelling community. However, of course the reviewer is correct in pointing out all the limitation of NSE (which are aware of). We actually also used KGE as the calibrated metrics, but only minor differences to the NSE results were found. However, given these comments we will replace NSE by KGE in the calibration to exclude any potential misleading impacts from NSE.*

There was no validation or clear evaluation of the model. Shen et al. (2022) was referenced to justify this omission, but this does not excuse the absence of some other method than split-sample validation to test the calibrated models. There is currently no clear evidence that the final models are at all reliable and not just overfit to the calibration data. This might be corrected by using satellite or other data to justify the 'trustworthiness' of the models but it should be an explicit evaluation.

Reply: *As the catchments were influenced by unrecorded managements, and the measures could be different in contrasting period, the split-sample validation may not be appropriate (as the reviewer hints at). We employed products from MODIS, PML remote sensing ET, and a flux tower nearby the studied four sub-catchments as a comparison to our model performances, we will compare these ET datasets with simulations in different spatio-temporal scales, and make the evaluation more explicit, and justifications of the trustworthiness of the model will be added in the result section*.

It seems only a single calibration trial was performed for each objective type. The final calibrated models will vary depending on the initial population for the genetic algorithm, and on the random seed used in mutating new solutions. It is therefore important to run several independent calibration trials for each objective, as a single trial may be an outlier or fail to generate solutions near the 'true' Pareto front (i.e., solutions that are actually as good as the model can do). Without multiple independent calibration trials, it remains possible and plausible that the poor quality solutions for Berste were simply a fluke.

Reply: *Thank you for this recommendation, we will replicate the calibration with different initial population, and collect the final Pareto front, and treat it as the final result in the revision.*

The presentation of the results would benefit greatly from revision in a few areas. In no particular order:
The presented time-series results have only the extreme end points of the pareto front, not the 'compromise' solutions, basically throwing out the 'multi-objectiveness' in favor of one simulation or the other. Why show only outliers?

Reply: *The idea of presenting of the end points of the pareto front was to show how streamflow and isotopes pull the model into different directions, and to better explain how management measures lead to trade-offs between streamflow and isotopes in the calibration. However, the metrics of the compromised solutions were actually already shown but we will add the compromised solutions in the time-series results in the revision. Thanks for this suggestion.*

Figure 4 is mislabeled as showing the Pareto fronts, but it actual has both dominated and non-dominated solutions from the calibration. Either the figure or label needs to change.

Reply: *The Pareto fronts were calculated based on NSE of both streamflow and isotopes, but we showed KGE values of isotopes in the plot. In the revision, we will make this consistent and only show the KGE results*

Labeling can be challenging to decipher. For example, subfigure 7 c2 is apparently 'BSI in schemes 2-5 for wet year of 2023' while figure 8 c2 is 'Vetschauer compromised solution in scheme 2-5' (I don't know which compromise solution, just that it is one). Some figures are quite reader-friendly (Figure 5 and 6 for example can be followed without taxing decoding). However, I was quite unable to read the alphabet soup of Table 4 even after writing out a 'key' on scrap paper to track the 4 item deep 'respectively' label linking processes to letters (I think at least one comma is missing from the list).

Reply: *Sorry for this confusion, we will clarify and double check all figures and add the description of the "compromised solution" clearly in the method section. We will make figure 7 and figure 8 consistent and show the difference between isotope-aided and streamflow-only calibrations. We will adapt the table 4 to be readable.*

Returning to the mysterious compromised solution, the actual solution is not defined, only that it comes from the 'middle part of the Pareto front'. Is it the optimal solution when equal weight is given to the flow and isotope KGE or was it just sort of eyeballed?

Reply: *We used the equal weight to select the optimal compromised solution and we will articulate this more clearly mention this in revision. Sorry about the confusion.*

A final, minor, point: it was frustrating to be told about finicky model details like roughness coefficient values without knowing any of the model basics, which were relegated to the supplement. Certainly, detailed model descriptions are out of scope but it would be lovely to at least have a couple sentences so the reader knows how many soil layers there are or if there is lateral groundwater flow between cells without hunting down a separate document.

Reply: *Thanks for this suggestion (though we respectfully not agree that we presented "finicky model details"). We will add a more detailed description of the major aspects of the model structure.*

Technical corrections:
The precipitation isotope input is referenced as coming from Bowen et al. (2003) which covers annual averages, but the inputs seem to be the monthly average estimates. The monthly estimation method comes from the subsequent 2005 paper (Bowen G. J., Wassenaar L. I. and Hobson K. A. (2005) Global application of stable hydrogen and oxygen isotopes to wildlife forensics. Oecologia 143, 337-348, doi:10.1007/s00442-004-1813-y.).

Reply: *Thank you for spotting this. We will correct this accordingly.*