Dear Prof. Yongping Wei,

Thank you for sending the detailed and careful reviews of our work. We hereby provide our summary of how the manuscript has been revised according to reviewer's comments. We are grateful to the reviewers for their constructive criticism which motivated us to re-do some of the modelling and re-frame the paper in terms of using tracer-aided modelling as a learning tool in complex, heavily managed catchments. The result is a thoroughly revised manuscript.

First of all, we recalibrated the model using KGE of both discharge and isotope, and repeated the calibrations five times. We clearly defined three solution types (i.e., discharge-dominated, balanced (equally weighting discharge and isotopes) and isotope-dominated) in the first Pareto front of each calibration. Each solution type included 25 parameter sets, rather than single parameter set as in the previous version, which increased the robustness of the results and now better shows model uncertainty. Accordingly, we revised all figures, and showed balanced solutions in the main manuscript, while the discharge- and isotope- dominated solutions were shown in supplementary materials. Further, we used spatially averaged MODIS, PML ET values for additional informal validation, and the high KGE further support the new modelling approach. The increasing pattern of simulated subsurface storage also aligned with the inter-annual variations in locally measured groundwater levels which we now include in the supplementary material.

Second, we further discussed the potential biases introduced by seasonally sampled isotopes and monthly rainfall isotope inputs. Since the evaluation of trade-offs primarily relies on the seasonal isotope patterns, we believe that the sampled streamwater isotopes adequately capture this feature. The monthly rainfall isotope inputs were also validated against daily rainfall isotope (aggregated to monthly value for comparison) from a station in the downstream region, showing strong agreement and supporting their use in the modelling. We also acknowledged that such coarsely sampled isotopes could underestimated seasonally aggregated fractionation, and resulted in overestimation of transpiration volumes.

Third, in response to the reviewers' concern regarding the trade-offs associated with human factors, we extended our discussion accordingly. Based on the isotope-aided calibrations and comparisons among the four catchments, we found that the water celerity inferred from the dampened isotope variations was inconsistent with the observed runoff behaviour at Berste, indicating potential influences of water abstraction on the hydrological response. In addition, we also discussed the potential for future improvements by explicit processes representation on trade-offs between isotope and discharge, and their limitations as well.

Fourth, we highlighted the novelty of the present work at the start of the discussion. It lies in evaluating the use of seasonally sampled isotopes for calibrating a tracer-aided model applied in an ET-dominated and heavily managed region. Contrasting discharge-isotope trade-offs among the studied catchments provide preliminary insights of the catchment functioning, and potential influences by human factors.

We feel that these changes have substantially strengthened the manuscript and look forward to its reconsideration.

Best wishes,

Hanwu Zheng (on behalf of all co-authors)

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 were 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: *Thanks for the comments. We replotted all figures and improved their clarity. The methods were refined, and the model was recalibrated using the KGE of both discharge and isotopes, with five independent calibration runs conducted. Comparisons among calibrations were based on ensembles of parameter sets rather than single parameter sets, thereby enhancing the robustness of the conclusions. A key novelty of this study lies in evaluating the use of seasonally sampled isotopes to calibrate a tracer-aided model applied in an evapotranspiration-dominated and heavily managed region. We have emphasized this perspective at the beginning of the discussion section. (line 502-517).*

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 optimization using isotopes and streamflow have been cited in the revised manuscript. We also emphasized the novel contributions of this study in the Discussion section (lines 502-517), along with its limitations.* 1. *The inclusion of isotopes improves the modelling of water partitioning, albeit in heavily managed regions. 2. Coarsely sampled isotope data remain valuable, although underestimation of isotope fractionation can introduce uncertainties.3. The contrasting discharge-isotope trade-offs among the studied catchments provide preliminary insights of catchment functioning and reflected the influences of human factors.*

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: *We replaced NSE with KGE and repeated the calibrations five times. The solutions located on the first Pareto front of each calibration were used for the analysis. Although a slight improvement was observed at Berste, this catchment still exhibited the strongest trade-offs between isotopes and discharge among the four catchments. Overall, the main conclusions remain consistent with the previous version but are now supported in a more robust modelling approach and result.*

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: *We feel that the relatively short run of isotope data precludes split samples calibration/validation that many prefer. However, we additionally calculated the KGE of spatially averaged ET for each calibration scheme using both MODIS and PML products (Table 4). The results were generally satisfactory, except at Berste, where the incorporation of isotopes led to a clear degradation in simulated ET performance, likely due to trade-offs between isotope and discharge objectives. We have also discussed the limitations of using discharge and isotopes to constrain the spatial patterns of ET. In addition, the annually increased subsurface storage aligned with the interannual variation of the averaged groundwater level over 3 wells in the studied region, further supporting the modelling.*

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 repeated the calibration five times using different initial populations, and the first Pareto front from each calibration was used for the analysis. The relatively poor solutions persisted at Berste, and the main conclusions remained consistent with those of the previous version.*

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: *We made the revision accordingly*. *The balanced (compromised) solutions from each calibration scheme are presented in all figures of the main manuscript, while the edge solutions of the Pareto fronts are shown in the supplementary materials.*

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: *In Figure 4, we plotted the first Pareto front from each of the five individual calibrations (five Pareto fronts in total for each scheme).*

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. The balanced, discharge-dominated, and isotope-dominated solutions from the isotope-aided calibrations were clearly defined in the methods section (lines 320–330). All figures in the manuscript consistently show the balanced (compromised) solutions for scheme 1 and schemes 2–5. In addition, the previous Table 4 was replaced with a new figure (Figure 5) illustrating the sensitive parameters identified under different objective functions.*

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 clearly defined the balanced ("compromised" in the previous version), discharge-dominated, and isotope-dominated solutions from isotope-aided calibrations in the method section (lines 320-330).*

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: *We briefly mentioned the structure of the STARR model in the method section (lines 191-193), e.g., the basic storages and fluxes.*

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: *We corrected this accordingly.*

Author response to Referee #2 comments:

We thank reviewer 2 for the detailed and careful review of our work. We hereby provide our point by point responses how the comments by referee #2 were addressed in the revised manuscript.

Best,
Hanwu Zheng (on behalf of all co-authors)

Anonymous Referee #2
This manuscript applies a large-scale tracer-aided modeling (TAM) approach to disentangle ecohydrological processes in the heavily managed Middle Spree catchment (MSC), Germany, an evapotranspiration-dominated region facing strong anthropogenic pressures. By integrating stable water isotopes ($\delta^{18}O$ and $\delta^{2}H$) with streamflow into the distributed STARR model and calibrating with a multi-objective NSGA-II algorithm, the study evaluates runoff generation, groundwater contributions, and evapotranspiration (ET) partitioning across four sub-catchments (Berste, Wudritz, Vetschauer, Dobra). The key contribution lies in showing how streamflow–isotope trade-offs emerge as diagnostic signals of epistemic errors from unrecorded human impacts, such as irrigation or mining legacies. While isotope inclusion sometimes reduced discharge simulation performance, it significantly improved process representation such as subsurface mixing. Overall, the study demonstrates that even sparse seasonal isotope datasets can provide critical constraints in TAM for complex, human-altered hydrological systems, offering new insights into ecohydrological partitioning and informing future water management under anthropogenic and climatic pressures. From a reader's perspective not deeply familiar with isotope tracer methods, I have several comments and suggestions for clarification.

Reply: *We thank reviewer 2 for the detailed and careful review of our work and acknowledgment of the novel contribution showing the value of even coarse isotope data for insights into ecohydrological functioning.*

Points for the Authors to Consider
1.Clarifying the added value of isotopes
The added value of incorporating isotopes over other hydrological variables remains somewhat unclear. For instance, while the introduction emphasizes human influences, isotope integration did not appear to improve the model's ability to capture these anthropogenic effects, which raises questions about the practical contribution of isotopes in this context.

Reply: *Thanks for the comments. We have highlighted the added value of incorporating isotopes into the calibration process and the novelty of this study in the Discussion section (lines 502-517). The emphasis on human influences was intended to illustrate the complexity of process representation in the studied catchments. Our results showed that incorporating isotopes in model calibration can still improve performance in such managed regions, demonstrating the potential for gaining informative insights from broader use of isotope data. In addition, the trade-offs between isotope and discharge objectives improves the understanding of local hydrological regime and potential influences by human factors, pointing out the potential future adaptations on the model structure. To avoid misleading readers into expecting that isotopes were used to explicitly identify anthropogenic influences, we made some adaptations in the introduction and more clearly stated the novelty in the discussion.*

How would the results compare if ET data were used in a multi-objective calibration of the STARR model?
Reply: *We extended our discussion in this aspect, and emphasized that isotopes can also constrain ET, which is similar to ET data. However, the ability of water partitioning by isotope is unique compared with other datasets (lines 557-559; lines 576-579). In addition, the spatially averaged ET were used for informal validation (line 429), and the current calibration targets already showed satisfied ET performance.*

Could the process descriptions be refined to more clearly illustrate the unique role isotopes play relative to other potential data sources?
Reply: *We revised the discussion section and more clearly demonstrated how modelling of water partitioning (unique role) were improved by isotopes (section 4.2). The limitations due to the coarse isotope data were also shown (lines 581-584, 608-610, 640-648).*

2.Improving figure clarity and linkage to discussion
Figures 5–8 combine multiple dimensions (temporal, spatial, and calibration metrics), making them information-rich but sometimes challenging to interpret. The figure captions and related explanations in the text could more directly highlight the core message of each figure. Including a short statement of motivation or the specific hypothesis addressed by each figure would help guide readers and improve accessibility. Moreover, because the figures are complex and the key messages are not always clearly highlighted, the subsequent discussion section becomes less convincing. Readers may find it difficult to fully trust the discussion, as the results and the interpretations are not always tightly aligned. Strengthening the clarity of figures and explicitly linking their core findings to the corresponding discussion points would improve the manuscript's overall persuasiveness.
Reply: *We revised all figures and made them more reader-friendly, figure captions were clarified accordingly. An overall description of findings for each figure were added in the result section.*

Specific Comments
Lines 127 and 140: Please clarify the meaning of SE and m.a.s.l.
Reply: *Sorry for the confusion, we clarified accordingly (line 127, line 140).*

Lines 240–243: Rainfall inputs are provided at daily resolution, whereas precipitation isotope inputs are monthly. How does this temporal inconsistency affect the results, and is this assumption reasonable?
Reply: *Since our main conclusions were based on seasonal dynamics, this assumption should be reasonable. We also pointed out the limitations of the monthly isotope inputs in the discussion section (lines 581-584; 640-648).*

Lines 249–251: Although a citation is provided, the manuscript would benefit from more detail on the isotope observations. Were these instantaneous grab samples, or integrated/accumulated values?
Reply: *They were instantaneously collected "grab" samples. We added this information in the method section (line 257).*

Table 3 (Scheme 1): Please clarify whether the calibration was performed jointly across all basins, or if each basin was calibrated independently.
Reply: *The scheme1 was calibrated jointly across all basins, and this was explained in lines 310-311.*

Figure 3: Why are only δ^2H time series presented, while $\delta^{18}O$ observations and simulations are not shown? It would also help readers unfamiliar with isotope applications if key concepts such as LMWL and VSMOW were briefly explained.
Reply: *Since the ^{18}O and ^{2}H have the similar fractionation and mixing processes, normally we use only one variable to constrain the model and just show the variation of the used parameter. We explained LMWL in line 369, while the VSMOW were explained in line 261.*

Figure 4: KGE is used for isotopes and NSE for streamflow. Why not use the same performance metric for both, to improve comparability?
Reply: *Sorry for the confusion, we recalibrated the model using KGE of both discharge and isotope, and corrected figures and manuscript accordingly.*

Table 4: The description of Table 4 appears in the first paragraph of the Results, though the table is first referenced in Section 3.2.2. Consider relocating the description for consistency.
Reply: *We used figure 5 to replace this table to better show the sensitive parameters, and put the figure in the corresponding section.*

Author response to Referee #3 comments:

We thank reviewer 3 for the detailed and careful review of our work. We hereby provide our point by point responses how the comments by referee #3 were addressed in the revised manuscript.
Best wishes,
Hanwu Zheng (on behalf of all co-authors)

Anonymous Referee #3
This study addresses an important field in ecohydrological analyses, namely the explicit modelling of tracers to better understand (eco)hydrological systems. In the case of this study, the focus lies on the modelling of water stable isotopic signatures in rural catchments where the (eco)hydrological dynamics are heavily affected by human activities. Not only are the dynamics in the studied area affected by human activities today, but the areas were subject to heavy mining in before the 1990s, and the subsurface hydrology is thus highly altered. The study thus presents the very important advance in (and analysis of) the explicit modelling of water stable isotopes in complex watersheds impacted by human activities, moving away from the focus on process representation in mostly natural and remote systems. The main outcome or the central draw of the study lies in fact not in the perfect model representation of all human-induced changes to the system, but instead in the identification of the unknown and from a model-perspective structurally unrepresented processes and dynamics. These abesence of these prior-to-modelling unknown processes and dynamics in the model structure are described to become evident in the mismatch between the water stable isotopic simulations and the actual isotopic data.
Reply: *We thank reviewer 3 for the detailed and careful review of our work and this positive assessment of the importance of our work.*

On the upside, the study reads very well, with some exceptions the sites and data are nicely presented, the figures are clean and the results, discussion and conclusions are written in clear manner. I have some minor recommendations for the improvement of the text, particularly the abstract which could improve in clarity about the achieved results.
Reply: *Thank you for this positive assessment.*

I also find that some sentences in the introduction of the study and the study sites are a bit unclear. I do miss some broader discussion and literature outside of the grey box-type rainfall-runoff modelling domain, especially when it comes to understanding the worth of tracers for physically-based models and to using fully integrated or fully explicit physically based models to identify structural deficits in models by comparing them against tracers. Outside of the grey-box type rainfall-runoff modelling domain, many studies have looked at the worth of tracers for tracer aided modelling. Be this by postprocessing tracer data to become comparable to standard model outputs, or by semi-explicitly or fully-explicitly simulating tracer processes in physically based models. The insights gained from these exercises have helped in understanding the information content of different types of tracers, improving model predictions and in identifying model structural problems. I suggest adding some more references to the many other studies that our there, and I provide a reference to a review that has summarized the findings from many studies up to the year 2019.
Reply: *Thank you for these constructive suggestions. We referred the study recommended by the reviewer in both introduction and discussion. We also extended our discussions on how explicit process representation help on incorporation of isotopes (line 691-717). We mainly focused on explicit representation of runoff generation and ET partitioning in this regard, which highly related to isotope variations.*

On the downside, in terms of methodology, I do have critical concerns regarding the model-data interaction and the validity of the conclusions:
For the forcing of the isotope component of the model, a global model was used to define the monthly constant input signal. Subsequently, the model was calibrated against two types of data, namely seasonal stable water isotope measurements per catchment ((hence 4 per year, for 3 years = only 12 datapoints per subcatchment) and daily streamflow observations from discharge stations of the

subcatchments. During calibration, 35 different model parameters were inversely identified. This entire onset and procedure raises several questions that are critical for the interpretation of the results.
First of all, neither the discharge gauging stations nor the locations of the stable water isotope measurements are indicated in figure 1. I assume that the measurements were taken at the outlet of the subcatchments, but this is just a guess. Please indicate the locations of the measurements.
Reply: *We replotted the figure 1 and pointed out the outlets in each catchment, and highlighted the sampling locations in line 253,258.*

Subsequently, it is unclear what the 4 stable water isotope datapoints per year represent. Are these simple grab samples? Were they taken after rainfall events or do they represent pure baseflow? Or are these cumulative samples taken over the course of a season? It is not enough to say that the sampling procedures can be read elsewhere, because there are huge implications for the model calibration (and interpretation) from what the samples represent.
Reply: *They were instantaneously collected "grab" samples and the sampling campaigns were conducted outside rainfall events roughly one per season. We added more information related to the sampling of isotopes (lines 257-262).*

Beyond the fact that it is mostly unclear what these tracer datapoints actually represent, forcing a model with some global model-derived isotope product instead of locally sampled or robustly characterized rainfall input signals introduces a major bias into the model which even by calibration may not be resolved, and which could cause some or evan all of the biases that the authors associate to the absence of some human-/land-use-/infrastructure-related model structural deficits. I personally seriously doubt that it is possible to differentiate between the origin of the biases with such a "minimalistic" dataset relative to the large complexity of the modelled systems, especially if one is using a lumped parameter or grey-box modelling approach.
Reply: *Thank you for the comments. We compared the monthly values of the global product with a station dataset (originally at daily resolution and aggregated to monthly values) collected downstream of the study area. The two datasets showed good consistency in their seasonal patterns, indicating that the global product is representative of seasonal isotope inputs (lines 245-249). Since our conclusions regarding the evaluation of trade-offs are primarily based on the seasonal isotope patterns (line 719-728), the use of this global product therefore seems reasonable. In addition, we also compared contrasting trade-offs and human factors among the studied four catchments, and further explained conflicts between discharge behaviors and catchment celerity inferred by isotopes (line 664-688). Certainly, the coarsely sampled dataset is hard to constrain every aspects of the model, and we also highlighted the uncertainty potentially brought by this monthly rainfall inputs in lines 581-584, 608-610, 640-648.*

Of course, it can be shown that even a little bit of tracer data can improve model calibration, but that is not new and has been looked at in countless studies and synthesized in extensive detail in multiple review papers on the matter. Moreover, this relatively minimalistic tracer dataset with respect to the complexity of the studied system and the model structure was used to calibrate an entirety of 35 model parameters. Yes, daily streamflow data was alos considerd, but as was already introduced in the introduction by the authors themselves, these data are extremely ambiguous with respect to identifying correct parameter values in such catchment scale surface-subsurface hydrological models, even if of the lumped parameter type. There is simply no way that this dataset contains sufficient information to constrain so many model parameters - a fact that was also introduced by the authors in the introduction via references to the "right answers for the wrong reasons". Yes, the calibration aimed at pareto front identification, but even if the objective function and calibration approach is tailored to this situation, the lack of information in both the observation data as well as the forcing functions cannot be overcome.

Reply: *We identified the value of coarsely sampled isotopes in catchments influenced by human activities, which represents one of the key novelties of this study and has been highlighted at the beginning of the discussion section (lines 502-517). Of course, it would be better to have more data, but in reality the intensive multi-year data available from research catchments in rarely available in more applied studies such as this one. We also clarified that our focus lies on the seasonal isotope*

patterns, which can be adequately captured by three years of seasonally sampled data (lines 7178-727). In addition, we emphasized that such coarse isotope sampling inevitably introduces some uncertainty into the modeling, particularly in accurately partitioning ET. (lines 581-584, 608-610, 640-648). Whilst we are, of course, aware that in such a relatively complex model that not all parameters can be constrained, but the sensitivity analysis highlighted the influence of sub-surface drainage characteristics and these were more reasonably constrained by the isotope data.

I may have missed something important in the study, but how I understand it at the moment, unfortunately, I am not convinced that the present approach can overcome this data scarcity problem to a degree that the insights gained from the study with respect to model structural deficits are unbiased enough to enable the detection of missing information on human infrastructure and alterations to the system. Or even allow a rating of the representativeness of ET partitioning, soil and baseflow processes. Many unresolved problems could simply, and do most likely, stem from inappropriate stable isotope forcing functions, too little tracer data for calibration, and too many parameters featuring into the calibration objective function. In other words, if your forcing/input function is sufficiently wrong, you will never be able to match both stable isotope records in streamflow as well as streamflow volumes against the same combined dataset. And if there is so little data used to calibrate so many parameters, then if one would be able to match both types of observations simultaneously (isotopes and discharge), there is zero guarantee that 35 parameters that were calibrated do not overcompensate for structural model problems. Ok, this latter version of the same problem did not manifest, but the first version of this problem did, and I don't see any convincing arguments that would tell me that the problem of the mismatch lies in structural model deficits from unknown human alterations and not from a problem in the isotope forcing function.
Reply: *Sorry for the confusion in the manuscript. In the present study, we used seasonally sampled isotopes to enhance our understandings of the local hydrology (under heavily human influences), and this dataset showed its valuable function, although the temporally coarse dataset also introduced uncertainties (This was further clarified in the discussion section). We made several adaptations to increase the robustness of the modelling. We replicated the calibrations five times and each solution type (i.e., compromised, discharge-dominated and isotope-dominated) included an ensemble of parameter sets, rather than single parameter set as in the previous results, which increased representativeness of the conclusions. We explained more clearly how the trade-offs were potentially related to human factors, through comparisons among observations and across the four catchments (lines 664-688). Since this conclusion regarding trade-offs is based on the seasonal patterns of isotopes, such coarsely sampled dataset is sufficient to calibrate the model (line 719-728). In addition, we also pointed out the seasonally aggregated fractionation might be underestimated and lead to overestimated transpiration ratios (lines 608-610). Of course, in an ideal world we would have preferred more data, but we had to work with the data we had. Moreover, it is clear that we are not pretending to have the perfect model of the study but simply use what data and tools we have to help better understand practical problems in a constructive way.*

Ultimately, unless the authors present some additional hard data that support the claims on model validity, and unless the possible biases from model forcings, limited information content of the scarce tracer data, and the use of a grey-box model, are discussed and can convincingly be dismissed, I unfortunately can't support the manuscript for publication in HESS.
Reply: *In addition to extending the discussion on why seasonally sampled isotopes are sufficient to support our modeling and acknowledging the limitations associated with coarsely sampled isotope data (lines 664-688), we also calculated KGE of spatially averaged remotely sensed ET and showed measured interannually and spatially averaged groundwater levels, both presented reasonable alignment with simulations. The potential bias from model forcing (lines 581-584, 608-610, 640-648) and structure (line 690-716) were further discussed in the discussion section. Sorry that we didn't show any chemistry of the stream water as promised in previous reply, as they didn't provide any additional information.*

Specific comments

abstract: The abstract should provide the reader with information about the type of analysis that was done, but also for what this type of analysis can be used specifically. The first part is ticked off by the existing abstract, but the second part not so well, as the author's don't provide any clear examples of what kind of epistemic errors may found with their approach. This is because the section on the epistemic errors in the abstract reads very general, and it is difficult to infer what exactly the author's mean by "epistemic errors manifested as strong trade-offs between the information content..." The next sentences remain similarly unclear as to which kind of epistemic error, or which specific source for it, could be a likely cause of the "trade offs in information content"". It is alluded to that the model can help to identify the sources of these errors, ("potential for informative insights"), even when one only has sparse isotopic data to complement streamflow. But the exact use of the approach remains unclear. Here I would strongly suggest to provide one or two examples of which kind of sources for epistemic erros can be identified, and have been identified in this study.
Reply: *We adapted the abstract and pointed out the benefits of the discharge-isotope trade-offs in enhancing understandings of hydrological processes in the managed catchments. We also clearly pointed out that the human factors (i.e., water withdrawn) potentially led to such conflicts between discharge and isotopes (lines 31-32).*

l49: "non stationary climate inputs": what is meant by this? the "climate" usually is a longer term phenomenon, i.e. one assessed over a 30-year period conventionally. I think here something else is meant than a varying climate, namely the inter-annual variation, and therefore not a climate signal?
Reply: *Sorry for the confusion. Here we mean the catchment functioning under changing rainfall, e.g., wet or dry periods (and their duration, magnitude, frequency) might change, and the ecohydrological processes may also vary accordingly. This changing climate input could be inter-annual variations, or intra-annual patterns. We changed it to "rainfall" instead (line 47).*

l66-67: A large number of studies has looked at the benefit of tracers for model calibration, some have even quantified the information content. An extensive review on this has been published in 2019, but article is not in the list here.
Schilling, O. S., Cook, P. G., & Brunner, P. (2019). Beyond classical observations in hydrogeology: The advantages of including exchange flux, temperature, tracer concentration, residence time and soil moisture observations in groundwater model calibration. Rev. Geophys., 57(1), 146-182. https://doi.org/10.1029/2018RG000619
Reply: *Thanks for this suggestion. We cited this paper.*

l167f: this sentence is unclear to me. "...the decline of pumped sump water volumes has been faster than the replenishment of the groundwater deficit". What do you mean exactly by "sump water", and do you want to say the reduction groundwater abstraction was faster than the groundwater recharge, i.e. the recovery of the water table didn't happen as quickly as stopping in abstracting groundwater? It seems to be quite a complicated way to say something that isn't so complicated. Could you reformulate to make it clearer?
Reply: *Sorry for the confusion. We corrected it accordingly (line 165).*

l300: "and isotope." seems unfinished
Reply: *It is finished. Here we mean calibrations based on discharge and isotope.*

Discussion: The discussion is written as if the authors know which model performs best for soil water storage and flow as well as groundwater recharge, storage and flow. However, no comparison between actual data and these simulated components are made, and the entire discussion is based on high level observations and assumptions about the catchment's functioning and the assumption that the calibration approach and information contained in tracers would allow these insights to be gained. But as critically mentioned above, unless I see hard data on the validity of the isotope input function and the soil and groundwater components, I am convinced that the available data is not sufficient to derive the conclusions that are discussed in the discussion section.
Reply: *The balanced solutions in isotope-aided calibration showed good performance metrics in both calibrated variables at Vetschauer, Wudritz and Dobra, while discharge-only based calibration showed*

extremely deviated isotope values. Isotope values are influenced by storage mixing and fractionation processes, and therefore discharge-only based calibrations presented incorrect inference in this aspect, and isotope-aided calibrations are more reliable. Further, we clarified and evaluated the strong discharge-isotope trade-offs at Berste, through analyzing the reasons and potential influences by human factors (lines 664-688). Moreover, we calculated KGE of spatially averaged remotely sensed ET and showed measured interannually and spatially averaged groundwater levels, both presented good alignment with simulations (lines 429-494). Since our conclusions are mainly based on seasonal patterns of isotopes, the coarsely sampled dataset is sufficient (lines 719-728). Our conclusions rely not only on the comparisons of information contents by discharge and isotopes, but also comparisons across the four catchments. The inferences were also compared with other studies in the nearby regions, and showed good agreements after incorporation of isotopes (lines 591-603, 605-606). We focused on how different calibrated variables constrain and lead the model, and how they showed conflicted information in the modelling. We made the explanations more clearly in these aspects in the discussion section. In addition, we acknowledged the limitations by the conceptual model (lines 690-716) and coarsely sampled data (lines 581-584, 608-610, 640-648).

In the entire discussion, the lack of information on the true stable isotope input signals as well as the possible minimal information content of the stable isotope measurements from the 4 seasonal streamflow samples remains unmentioned.
Reply: *We further highlighted the limitations in this aspect (lines 581-584, 608-610, 640-648).*

Instead, it is repeatedly claimed that the information content of stable isotopes is very high, and these assumptions are supposedly supported by information on soil water storage overestimation, correct ET partitioning and underestimation of baseflow etc. However, as stated previously, no hard data on all these processes are used to compare to the model outputs, and therefore all these claims remain relatively unsupported.
Reply: *With respect, we were not meaning to claim the information content of the isotope data is "very high", but rather that it adds substantially more to calibration than stream flow data alone. The balanced solutions in isotope-aided calibration showed good performance metrics in both calibrated variables at Vetschauer, Wudritz and Dobra, while discharge-only based calibration showed extremely deviated isotope values. Isotope values are influenced by storage mixing and fractionation processes, and therefore discharge-only based calibrations presented incorrect inference in this aspect, and isotope-aided calibrations are more reliable. Further, we clarified and evaluated the strong discharge-isotope trade-offs at Berste, through analyzing the reasons and potential influences by human factors. Finally, we repeated the calibrations five times to make the methodology more robust than in the previous version and further explained why isotope data were sufficient (lines 719-728), while acknowledging the limitations associated with the coarsely sampled isotope data (lines 581-584, 608-610, 640-648).*