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 will be addressed in the revised manuscript.
Best wishes,
Hanwu Zheng (on behalf of all co-authors)

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 will add more relevant references as suggested and thanks also for the recommended reference. We will extend our discussion and review a broader range of literature on how isotopes can help identify structural deficits and improve model predictions in physically based models as suggested by the reviewer.*

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 apologize for the confusion regarding the locations of discharge and isotope measurements. We will indicate the locations in the revised figure 1, although we had already mentioned that they are taken at the outlets of all catchments (Line 250 in original submission).*

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 will add information in the methods on how and when the samples were collected and the data processed.*

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 acknowledge that the use of global isotope products may have introduced some bias into the model performance, though this is likely to be less influential than the reviewer implies. As noted in the discussion, such uncertainty in the rainfall isotope input is largely inevitable due to data limitations. We have local daily data from a rain gauge relatively close (~30km) to the catchments, though as we are modelling over a more extensive area of four catchments, and we know from other work that the catchments in Brandenburg are generally groundwater-dominated, so show relatively limited/slow variation in streamflow isotopes, we preferred the modelled input signal. However, we will check whether our daily data makes any difference.*
However, we are still convinced that there is scientific value in our approach – particular for such largescale investigations where data are coarse and rare. The global precipitation isotope product that we used has been widely applied and its value has been shown and we would argue that it can adequately describe the spatial pattern of rainfall. GNIP station data are only available in Berlin far downstream of this region. Further, we do not think that this product leads to a major bias in our conclusions, as our focus is on average catchment functioning at the seasonal scale. The main differences observed in the Pareto front stem from the simulated isotope seasonal patterns in stream flow signals in groundwater-dominated systems that have high summer ET.
The isotope inputs used in this study are spatially interpolated products based on GNIP stations, their robustness has already been demonstrated by other studies, showing they are suitable for representing the local seasonal patterns of rainfall isotopes. Second, although the seasonality of a specific year from the sampled four datapoints could still be biased, we used three years dataset to mitigate this potential annual bias. Third, our discussion focuses on how management influences these seasonal patterns, for example, water withdrawn reduced base flow discharge and forces the model to simulate a faster runoff process, which in turn exaggerates isotope seasonality and contradicts the observed isotope variations. All our conclusions are based on the seasonal patterns of isotope variations, which could be well represented. In addition, the conclusions are not drawn from a single case in isolation, but from a comparative analysis. In Vetschauer, characterized by similar landscapes and proximity to the poorly performing Berste subcatchment but with limited human influences, we obtained good performances in both simulated isotopes and streamflow, and this supports the validity of the input rainfall isotope product. We will make this clearer in the revision.

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: *Of course, we acknowledge that the added value of using isotopes has also been reported in other studies, but our applications focus on heavily human-influenced catchments and highlight the understanding of ET processes by distinct signatures of isotopes among catchments and the potential bias introduced by discharge only based calibrations, as well as the added value of isotopes in this regard, revealing the potential epistemic errors in the discharge observation caused by human activities. We also acknowledge that not all parameters can be constrained under the calibrated variables and this equifinality inevitably exists, not only in the present study. This is why we conducted sensitivity analysis and identified hydrological processes (or parameters) we can control (under the calibrated variables). Nevertheless, we do not consider our tracer dataset as too minimalistic to support our conclusions. As we mentioned in our discussion and in the above reply, the sampled isotopes, although being relatively coarse in temporal resolution, adequately capture the seasonal patterns, especially as we used three years of data. Our major conclusions are based on these better-constrained processes, e.g., potential human influences are concluded based on the conflicts of simulated isotope seasonal patterns in the pareto front and consistency with our knowledge of the catchment characteristics. Further, the seasonality of isotopes in rainfall is a pronounced characteristic, due to seasonal shifts in temperatures and atmospheric vapour, and streamflow usually follow the similar pattern but are damped or phase-shifted due to storage and mixing effects. Kirchner (2016) has shown the young water fraction can be quantified even in heterogeneous and nonstationary catchments. Seasonal isotope datasets have been used widely around the world to capture catchment functioning at larger spatial scales (Jasechko et al., 2016). We will clarify these points in the discussion section.*

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: *We apologize that some of our arguments may not have been clearly made. We acknowledge that data scarcity can lead to insufficiently constrained processes, and if human alterations mainly affect these uncertain processes, then we will gain less insights from the modelling. However, even a limited number of data points can sometimes provide sufficient information about key characteristics. In addition, we also used insights from "soft data" based knowledge on management measures and their effects on different hydrological processes (e.g. the high ET losses in this region, locations and*

amount of groundwater withdrawal or addition). In a way, these soft data based insights were confirmed by the isotope signatures and model results.
The seasonal pattern of the isotopes in this region (also presented in previous studies, e.g. Chen et al., 2023, were well captured by our seasonally sampled isotopes, and our analysis focussed on these gradually changing seasonal patterns of the simulated isotopes and flow paths (soil or groundwater) in the pareto front. In the Berste sub catchment, the model actually captured isotope and streamflow separately in the two edges of the pareto front, but this comes at the expense of a strongly degraded performance in the other calibrated variable. In other words, the seasonal isotope patterns derived from the calibrated discharge do not aligned with the observed seasonal isotope patterns. In contrast, at the Vetschauer sub catchment, such conflict does not exist, reflecting good consistency among model framework, input dataset and calibrated variables. The major difference between the two sub catchments is the degree of human influence. We also agree with the reviewer that the model may overcompensate structural errors and the actual quantification of ET partitioning, soil and baseflow processes could be biased by data scarcity. We also mentioned this limitation in the discussion. However, we highlight the added values of isotope through the comparisons among calibration schemes, and results presented high degree of alterations after incorporation of isotopes in calibrations, and this is more of a quantitative analysis. Lastly, the poor NSE performance resulted by conflicts in the calibration in our conceptualized model could potentially be improved by other physical-based models (showing better NSE) due to their larger parameter space, but the reasons (faster runoff processes supported by discharge and slower process by isotopes) resulting in the conflicts can still be illustrated in the calibrated parameters in the physical-based models. We will extend our discussion more in this regard.

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: *We hope our detailed explanations to the comments above show that we think we can address the concerns of the reviewer. Despite these comments, we note that they are at odds with the comments of Reviewers 1 and 2 who were more positive about the value and contribution of our paper. We apologize that the mentioned possible biases were not clearly explained in the original manuscript. We will certainly ensure these points made above are all much clearly explained in the revised manuscript. We will alter the discussion section to explain why we think seasonal tracer data are valuable for constraining the model and, along with other catchment knowledge, allow us to hypothesize the main influencing human factors. We will also use explore other datasets (i.e. measured daily rainfall isotopes from a nearby station to confirm the validity of the rainfall product; and water quality data in stream flow and groundwater to test the connection between surface and sub surface storage) to further assess the validity of our model performance*

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: *Thank you for the suggestions and sorry for the confusion. Through comparing modelling performances across different sub catchments, we observed different degrees of conflict between observed isotopes and streamflow, and attributed the major differences to the potential epistemic errors*

caused by human factors in the observed discharge. Specifically, in our study, human managements reduced base flow and the observed discharge likely misled the model into simulating a faster runoff process, which contradicted with slower runoff pattern indicated by isotope patterns. We will clarify these points in the revision.

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 climate, 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.*

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 will refer to and cite 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. Sump water is a term in mine dewatering and it means temporary water stores for groundwater or rainfall which may influence mining activities. Before mining, groundwater is pumped into the sump and later transferred to streams, and "the decline of pumped sump water" means reduction groundwater abstraction (as you noted). We will make it clearer accordingly.*

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: *Please see also our responses to the previous comments. These conclusions are based on the comparisons of different calibration schemes. Of course, as is usually the case, we don't know the exact partitioning of subsurface flow, but still have insights based on other observations and soft data for these systems. We highlighted the contrasting information provided by discharge and isotopes, that is, lower base flow and higher winter peaks in discharge reflected faster runoff, whereas the flattened seasonal isotope variations indicate slower water turnover. We do know: This flattened seasonal isotope variation means a larger water storage mixing and this requires greater hydrological connectivity and higher exchange rates between surface and subsurface flow. For the potential issues brought by data scarcity, we have explained that in the responses above.*
Additionally, we will consider presenting the comparison of hydro-chemical parameters between stream and groundwater to identify that there is apparent connection between surface and sub surface

storage, which is not captured by discharge-only based calibrations. We will clarify this in the discussion.

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: *With respect, this is incorrect: we actually mentioned that such rainfall isotope inputs possibly result in failure of capturing some short-term catchment variations in line 700-701. Besides, our conclusions are mainly based on averaged seasonal patterns, which can be represented well by this product in addition to our observations. We also mentioned the potential issues brought by the coarsely sampled isotopes, e.g., low controls on ET partitioning (line 697-699), and underscore the advantages of higher-resolution data. However, we will further highlight and more clearly describe advantages and limitations of using the coarsely sampled isotopes.*

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: *It was not our intention to claim that the information content of the stable isotopes is "very high", so apologies if this is the impression we gave. Rather, we seek to show that the data are insightful and helpful in using the modelling as a learning tool to hypothesize catchment function. As explained above, the conclusions are based on comparisons between different calibration schemes. More specifically, we used different variables to constrain the model, the model performances objected to different observations presented more controlled processes. In the better performed sub-catchment, incorporation of isotopes clearly narrows down the uncertainty and resulted in limited degraded simulated streamflow, and this is an improvement. We have shown that our simulations presented similar transpiration ratio with a RS product. We will further test the comparison of hydro-chemical parameters between stream and groundwater to identify that there is apparent connection between surface and sub surface storage, which is not captured by discharge-only based calibrations. These additional datasets will further support the present study.*