Anonymous referee #1:
I think the paper improved. However, the explanation about the Morris Scrrening parametrization was not enough. It is necessary to support the parametrization (and decisions) with other studies (citation is lacking).
We thank Referee #1 for the constructive comments, which have been carefully addressed and incorporated into the revised manuscript.

For example how many iterations did you run and why?
How you decide that threshold to consider the parameter importance?
You need to include citations of each decision that you made with Morris Screening parametrization.
We have now revised the section on the Morris method to clarify the threshold criteria applied for excluding less sensitive parameters, while adding new references.
We have also specified the model outputs considered in the sensitivity analysis and identified the Python library used.
Moreover, we have improved the linkage between the sensitivity results and the corresponding outputs, as illustrated in Figures S5 and S6 of the Supplement.

L. 375-410: For the sampling method of parameters, PiBEACH required the calibration of 43 parameters (Table S3 in the Supplement). The range, i.e. min-max values, of these parameters were defined based either on literature or field data collected in 2012 and 2016 (Lefrancq et al., 2017 and 2018; Alvarez-Zaldivar et al., 2018). These parameters were assumed to be a priori uniformly distributed within these min and max values (Table S3 in the Supplement). To reduce the number of runs required by the GLUE method, three steps were successively applied. First, a pre-sensitivity global analysis based on the Morris method (Morris, 1991; Herman and Usher, 2017; Campolongo et al., 2007) was conducted with the SALib Python library for sensitivity analysis (section 10 of the Supplement) to select the most sensitive parameters. Although the Morris method yields a qualitative indication of relative parameter importance, it is efficient compared to other sensitivity approaches (Gan et al., 2014) that screen for sensitive parameters (Herman et al., 2013). The mean and standard deviation of the elementary effects (EE) for each parameter were calculated as required by the Morris method. The mean represents the overall effect of a parameter on the model output, while the standard deviation captures the potential for interactions or non-linear effects. Parameters with a mean EE of zero or near zero, indicating negligible impact, were excluded, resulting in the removal of 21 parameters. The sensitivity analysis was conducted for the following outputs: S-metolachlor concentrations at the outlet; concentrations in composite topsoil transects (North, Valley, and South); discharge at the outlet; and isotope signatures both in the river at the outlet and within the topsoil composite transects. The EE statistics for the 25 retained parameters are shown in Figures S5 and S6 of the Supplement for each output variable.

The Morris method allowed to reduce the PiBEACH parameter number from 43 to 25 (Table S3 in the Supplement). Second, a Latin-Hypercube sampling (Herman and Usher, 2017) was used to reduce the numbers of runs (n = 2500) to cover the parameter space for the 25 parameters. To further reduce the computation time, the GLUE assessment focused on the growing period (March 19th to July 12th, 2016), where pesticide degradation and exports are of most

significance. Initial hydrological state was estimated from a spin-up period of one full hydrological year (Oct. 1st, 2015 - Sept. 30th, 2016) and hydrological parameters calibrated against observed discharge at the catchment outlet (March 19th and July 12th, 2016) using particle swarm optimization (Bratton et al., 2007).

Page 1 – line 9 (in the initial pdf):
Abstract:
You have a lot of data, yet the abstract only reports qualitative results:
We have now enhanced the abstract by incorporating key quantitative results.

Abstract. *Predicting pesticide dissipation at the catchment scale using hydrological models is challenging due to limited field data distinguishing degradative from non-degradative processes. This limitation hampers the calibration of key parameters such as biodegradation and volatilization half-lives (DT_{50}), and the carbon-water partition coefficient (K_{oc}), often leading to equifinality and reducing confidence in predictions of pesticide persistence in topsoil and transport from agricultural field to catchment outlets. This study examines the use of pesticide Compound-Specific Isotope Analysis (CSIA) data to improve model predictions of pesticide persistence in topsoil and off-site transport at the catchment scale. The study was conducted in a 47-ha crop catchment using the pre-emergence herbicide S-metolachlor. A new conceptual distributed hydrological model, PIBEACH, was developed to simulate daily pesticide dissipation in soils and its transport to surface waters. The model integrates changes in the carbon isotopic signatures ($\delta^{13}C$) of S-metolachlor during degradation to constrain key parameters and reduce equifinality. Model and parameter uncertainties were estimated using the Generalized Likelihood Uncertainty Estimation (GLUE) method. Incorporating $\delta^{13}C$ data and S-metolachlor concentrations from topsoil samples reduced the uncertainty in the estimated degradation half-life DT_{50} by more than half, yielding a value of 18 ± 4 days. This approach also significantly decreased uncertainty in six key metrics of pesticide persistence and transport. Between the day of application (day 0) and day 115, the modelled mass balance components, ranked by relative contribution, were as follows: degradation accounted for the majority at 82 ± 21%, followed by the remaining bioavailable mass in the topsoil at 12 ± 8%. Leaching contributed 4 ± 17%, while export to the river outlet accounted for 2 ± 6%. The irreversibly sorbed mass represented 1.1 ± 2.0%, and volatilization was minimal (<1%). The results highlighted that moderate, targeted sampling effort can identify degradation hot-spots and hot-moments in agricultural soil when stable isotope fractionation is integrated into the model. Overall, integrating CSIA data into PiBEACH model significantly enhances the reliability of pesticide degradation predictions at the catchment scale. In addition, PiBEACH, which accounts for spatial and seasonal variations in topsoil pesticide concentrations, enables coupling with distributed, event-based hydrological models such as OpenLISEM OLP to capture intra-event pesticide transport dynamics more accurately.*

Page 1 – line 31:
Why plural? / You mean pesticide degradation?

We have now modified the sentence as suggested:
While pressure on aquatic ecosystems continues to increase, accurately quantifying and predicting the contribution of individual pesticide dissipation processes in soil, through degradation and off-site transport to the catchment outlet, remains a major challenge.

Page 3 – line 88:
There is a noun missing here.

Indeed, we have now revised the sentence:
L. 92. More recently, CSIA data have been integrated into a lumped transport model using travel-time distributions, improving the interpretation of pesticide transport at the catchment scale (Lutz et al., 2017)

Page 3 – lines 90-92:
Most of this is captured in the travel time distribution. Your reasoning applies only to another type of lumped model. There are models that work with a convolution of application time distributions and travel time distributions, but these, to my knowledge, have not been applied to pesticides. I think your point is still valid, but the reasoning behind it can be improved.
We agree and we have now emphasised that travel time distributions can capture aspects of hydrological behaviour, as demonstrated in the study by Lutz et al. (2017), which to our knowledge represents one of the few applications in this context.

Additionally, we now have incorporated a recent reference addressing nitrate trends (Broers et al., 2024) (Lines 88–90). We have also highlighted that lumped models are limited in their ability to represent spatial variability of key parameters—such as soil moisture and temperature—that are critical for linking land use, pesticide application, and degradation processes.
Consequently, we have revised the sentence to clarify this limitation of lumped modelling approaches.

With a new reference:
Broers, H.P., van Vliet, M., Kivits, T., Vernes, R., Brussée, T., Sültenfuß, J., and Fraters, D.: Nitrate trend reversal in Dutch dual-permeability chalk springs, evaluated by tritium-based groundwater travel time distributions, Sci Total Environ., 15;951:175250. https://doi.org/10.1016/j.scitotenv.2024.175250, 2024.

L 92: More recently, CSIA data have been integrated into a lumped transport model using travel time distributions, enhancing the interpretation of pesticide transport and transformation processes at the catchment scale (Lutz et al., 2017). However, lumped models primarily capture aggregate hydrological behaviour, with some applications in water quality such as nitrate trend analysis (Broers et al., 2024), but they do not account for spatial variability in land use or topsoil parameters, including soil moisture and soil temperature (Fatichi et al., 2016). This limitation restricts their capacity to represent landscape heterogeneity—such as variations in crop distribution and pesticide application—thereby impeding the accurate identification of contaminant sources and degradation hotspots (Grundmann et al., 2007).

Page 4 - lines 108-109:
Isn't this an integral part of GLUE?
The sentence has been simplified to acknowledge that the Monte Carlo method is an integral component of the GLUE approach.

Page 5 – line 119 – figure 1:
I do not always understand the catchment boundary in relation to the ditches.

The landscape is predominantly flat, exhibiting slopes of less than 5.7% ± 2.9% throughout the catchment. Catchment delineation, derived from LiDAR data and validated by direct field observations during significant rainfall events, is strongly influenced by the topography of roads and tracks.

Page 5 – line 130
Most of the attributes are more chemical properties than composition

Indeed, and we have now modified the sentence as suggested:
The soil texture is predominantly composed of silt (61.0 ± 4.5%), followed by clay (30.8 ± 3.9%) and sand (8.5 ± 4.2%). The soil also contains calcium carbonate ($CaCO_3$: 1.1 ± 1.6%), organic matter (2.2 ± 0.3%), and total soluble phosphorus (0.11 ± 0.04 g kg^{-1}), and exhibits a cation exchange capacity (CEC) of 15.5 ± 1.3 cmol kg^{-1}.

Page 5 – line137:
That's from top to bottom in Fig. 1, I presume.

Indeed, we have modified the caption of Figure 1 to explicitly include the names of the three transects: North, Valley, and South.
Figure 1: *The Alteckendorf headwater catchment (Bas-Rhin, France), showing the experimental setup, including three transects, i.e. North, Valley and South (weighted samples collected at green dots along red lines) and plot sampling (black dots). Land use for 2016 is also displayed. The "Other" category includes roads, grass strips and orchards.*

Page 5 – line 140:
I presume you mean 'Samples from the topsoil

Indeed, we have now revised the sentence.
Samples from the topsoils (0-1 cm) were collected from individual plots and upstream-downstream transects across the catchment (Fig. 1 and 140 S1; Alvarez-Zaldivar et al., 2018).

Page 5 – line 142:
Do you mean you retrieved the same mass of soil at x locations along a transect, then mixed all these samples, and then obtained a single subsmample of the mixture, which was then analyzed for S-metolachlor and isotopes?

Indeed, we have not modified this sentence.
A single mixed sample was collected weekly, combining 30, 25, and 27 subsamples (green dots in the figure) for the North, Valley, and South transects, respectively.

Page 5 – line 142:
Were the masses of the samples taken in the field determined at field soil water content?)

The subsamples were collected in the field using a fixed volume rather than a fixed mass, as the latter is not feasible at the field scale.

Page 5 – line 144:
This was not done for the composite samples, or was it?

We have modified the sentence to clarify that these calculations were performed on both plot-scale and composite samples, as detailed below:
The volumetric topsoil water content (m^3 water m^{-3} soil) was calculated from gravimetric measurements obtained after drying samples at 110°C following NF ISO 1146 (Lefrancq et al., 2018). This procedure was applied to all samples, including field samples collected on days 1, 50 and 100 after application, as well as the weekly mixed samples from the three transects. It incorporated seasonal variations in topsoil bulk density, as modelled by PiBEACH and detailed in the Supplement, Section 7.2.

Page 6 – line 151.
In the previous sentence you stated the sampling was flow-proportional, but here the sampling was done weekly. I don't follow.

Indeed, flow-proportional sampling was conducted using an ISCO Avalanche autosampler. However, the collected samples were retrieved only on a weekly basis for subsequent analyses. This has now been clarified

I also do not understand what a fixed weekly discharge volume is. The \(cumulative\) discharge volume between samplings can either be fixed, or you can sample at fixed time intervals, but not both.

To optimize our sampling strategy—balancing the need to collect sufficient water for quantifying S-metolachlor concentrations, particularly for $\delta^{13}C$ analysis, while limiting the number of collection bottles to twelve 330 mL units per week—we implemented a variable threshold based on cumulative discharge volume to trigger sampling. This threshold was progressively increased from March to June to reflect the seasonal rise in baseflow discharge. Specifically, one bottle was filled after every 50 m^3 of cumulative discharge in March, whereas by June, the threshold had increased to 150 m^3 per bottle.

L. 176-184. This has now been clarified:
Runoff discharge at the catchment outlet was measured using a Doppler flowmeter (2150 Isco) with 3% accuracy and a 2 min resolution. Continuous, refrigerated, flow-proportional sampling was carried out using an Isco Avalanche autosampler equipped with twelve 330 mL bottles. Samples were collected based on fixed weekly discharge volumes from 50 to 150 m^3, in order to capture progressive increase in baseflow discharges from April to June 2016 (Alvarez-Zaldivar et al., 2018). To obtain sufficient S-metolachlor for quantification and CSIA, weekly

composite samples were prepared by pooling bottles according to hydrograph phase (base-flow, rising limb, and falling limb), yielding one to four samples per week with volumes ≥ 990 mL (Alvarez-Zaldivar et al., 2018). Piezometric monitoring of the shallow aquifer was not possible due to the absence of observation wells at the study site.

Page 6 – line 158:
river water:

We have modified the sentence accordingly:
To separate dissolved and particulate phases of S-metolachlor, river water samples were filtered through 0.7 µm glass fiber filters.

Page 7 – line 169:
This is still incomplete. The first paragraph is too general to give the reader enough background to understand the later paragraphs. For instance, the way in which the model represents the landscape and topography is not explained at all, and neither is the way the soil is represented. In later paragraphs you refer to cells, the plow layer and a deeper soil layer, without the reader knowing how these fit into the model.

I do not see this part of the acronym in the full name. Is something missing?
Indeed, BEACH is the acronym for *Bridge Event And Continuous Hydrological* modelling; accordingly, the sentence has been corrected accordingly:
The Pesticide-isotopes BEACH model (PiBEACH) was developed in Python based on the conceptual Bridge Event And Continuous Hydrological (BEACH) model (Sheikh et al., 2009).

You have not described the model cells yet. Because you have not explained how the cells are configured, and how the interactions between neighbouring cells are implemented, this text, and much of the rest of this modified text, is unclear.

We have clarified in the "PiBEACH development" section how the model cell is defined in this article and revised Figure 2 to schematically illustrate these cells

L 84: Similar to BEACH, the PiBEACH model employs square cells (x, y) with variable depths (z) corresponding to the soil layers considered (Fig. 2), in order to represent water and pesticide movement within the catchment, as detailed below.

Page 7 – line 197:

This is strange. The dispersion has dimensions, the export coefficient is dimensionless. How can you can compare their numerical values, given that you can change the value of one of them by simply changing the units?

Indeed, we acknowledge that our original sentence could be interpreted in this way. In Gatel et al. (2020), the impact of numerical dispersion was assessed as a dimensionless fraction of

the applied pesticide mass within the overall mass balance. We have now revised the sentence to clarify this comparison.

L 205: This challenge arises because numerical dispersion can affect the mass balance at the catchment scale (Gatel et al., 2020) to an extent comparable to the pesticide export coefficient—defined as the ratio of the mass transported at the outlet to the total mass applied within the catchment—which typically ranges from 0.1‰ to 1% of the applied pesticide load (Lefrancq et al., 2018).

Page 8 – line 222:

Should Z not be a depth interval? If so, what are the boundaries of this interval? The text above suggests that the interval comprises layers z0 - z2, and that the interval therefore is 0-80 cm. But you need to explain that better.

Equation 1 has also been modified to clearly indicate its application across the depth intervals from z_0 to z_3, using the generalised notation z_j (with j=0 to 3).

Figure 2 has been revised to use italic font for variables, in accordance with HESS formatting guidelines. In the updated version, the raster-based structure of PiBEACH is highlighted by depicting square cells with a 2×2 m plan view and variable depths corresponding to the different soil layers described earlier in the document and in the manuscript.

Page 8 – line 226:

This, and other variables like this, does not conform to the HESS guidelines. The notation is also inconsistent with that of other variables. Fonts that are italic in the text are regular here. Please make this consistent, and keep in mind HESS guidelines. Notation not in line with the rest of the paper, and with HESS guidelines. This happens too often to keep flagging it. Please go over the entire paper carefully to fix all occurrences.

We have thoroughly reviewed and revised all variables to appear in italic font, and have updated Figure 2 accordingly to ensure consistency with the HESS guidelines.

Page 9 – line 251:
We have corrected the sentence as suggested

Page 10 – line 266-271:
This limits the applicability of the new approach, does it not? Do you imply that this makes it acceptable to ignore isotope fractionation? If so, state this explicitly.

Indeed, this implies that isotope fractionation induced by sorption can be ignored.
This has now been specified in the manuscript.
L. 323-326. *In our case, since isotope fractionation associated with sorption and ageing processes is expected to be negligible, the observed isotope fractionation can be attributed*

primarily to biodegradation. This justifies the exclusion of non-destructive processes from the isotope mass balance and supports the use of the model to distinguish between destructive, i.e., biodegradation, and non-destructive processes, thereby enabling a quantitative evaluation of the contribution of biodegradation to pesticide dissipation.

Page 11 – line 296:
If these variables denote functions of temperature and water content, then say so.

We have modified the sentence as suggested:
A dynamic degradation rate ($k_{Dynamic}$, d^{-1}) was calculated daily as a function of soil temperature (F_T) and of water content (F_θ):

Page 11 – line 300:
What is this?

As introduced in line 296, a dynamic degradation rate ($k_{Dynamic}$, d^{-1}) was calculated daily from a K_{ref} and a function of soil temperature and of water content. The $k_{dynamic}$ provides a half time $DT_{dynamic}$ (day) calculated as $DT_{50,\ Dynamic} = \ln(2) / k_{Dynamic}$.

We have revised the sentence to clarify this step.
A dynamic half-time $DT_{50,Dynamic} = ln(2) / k_{Dynamic}$ was derived to be compared to $DT_{50,Ref}$.

Page 11 – line 296:
The units of F sub T are unclear.

We have now modified the sentence to clarify that the dependence equation of soil temperature (F_T) and soil moisture (F_θ) are unitless.
A dynamic degradation rate ($k_{Dynamic}$, d^{-1}) was calculated daily as a function of soil temperature (F_T) and of water content (F_θ):

Page 11 – line 315 to Page 12 – line 320:
This text seems out of place here. It contains some observation that should be in the Results section, and does not have a clear connection to the preceding text, even though it is part of the same paragraph. Because the line of thought is broken, I do not understand what point is being made here.

We acknowledge that this part of the text may be out of place. Therefore, we have moved this section, providing intermediate results, in the section results (3.1 Topsoil hydro-climatic dynamics and effect on S-metolachlor degradation rates).

Page 12 – line 329:

We have corrected the font size of the word “pesticide”

Page 12 – line 335:
Is this adjective necessary in this context? and plural for “macropore”

We have now modified the sentence by removing “explicit” and correction on macropores.
However, the integration of macropores at the catchment scale necessitates advanced in situ measurements (Weiler, 2017), and a combination of geostatistical methods, pedotransfer functions or meta-models, i.e., simplified statistical models built with 1D soil reactive transport models such as MACRO (Lindahl et al. 2008).

Page 12 – line 339:
Why is this sentence here?

We have removed this sentence and instead referenced previous work highlighting the negligible contribution of S-metolachlor plant uptake to the S-metolachlor mass balance at the plot scale (Lefrancq et al., 2018).
However, plant uptake of S-metolachlor was not included, as it is likely negligible (Lefrancq et al., 2018).

Page 13 – section 2.6:
Do I understand correctly that you calibrated the model on the full data set, and that, therefore, a validation was not carried out?
In that case, the graphs in the R&D section show how well the model reproduced the measurements, but tells us nothing about its predictive capabilities. this is correct, the discussion needs to reflect this.

We acknowledge the Referee's comments regarding the distinction between calibration and validation. The mention of model validation has been removed, and the following sentence has been added at the end of Section 3.2.
L 526: However, these findings necessitate further confirmation with a validation dataset, which was not available for the targeted catchment.

And the following in the conclusion:
L 620: The next step should involve confirming these findings with a validation dataset, which was not available for the targeted catchment.

Page 13 – line 359:
“Parameter values” / “with”
The sentence has been modified as suggested:
The GLUE method involved a sampling method of PiBEACH parameters values, an objective function incorporating observed dataset (i.e., topsoil S-metolachlor concentration only, then combined with S-metolachlor $\delta^{13}C$), a threshold of this objective function to select behavioural parameter sets, and the calculation of posterior probability distributions for parameters and uncertainties associated to the outputs of PiBEACH.

Page 14 – line 386:

Why past tense? / Between simulated and observed values? / Should S not be sigma?
The sentence has now been revised to correct the verb tense and to clarify the variables used in the calculation of the correlation coefficient, replacing the letter *S* with the Greek symbol σ to ensure consistency in notation.

where r is the linear correlation coefficient between simulated and observed values, $\alpha_{KGE} = \sigma_i / \sigma_0$, *and* $\beta_{KGE} = \mu_i / \mu_0$, *where* σ *and* μ *denoting the standard deviation and mean of simulated and observed values, respectively.*

Page 14 – line 391:
I don't understand what this means.

We have now revised the sentence to enhance clarity and improve reader comprehension:

L. 447-451. *The Kling–Gupta Efficiency (KGE) provides a more balanced assessment of model performance than traditional metrics such as the Mean Squared Error (MSE) or Nash–Sutcliffe Efficiency (NSE), which often favor parameter sets that underestimate output variability.*

Page 14 – line 395:

I think this should be a separate sentence, it is grammatically disconnected from the first part of the sentence, making the meaning of the full sentence (and its grammar) dubious.
Thank you. The sentence has been repositioned to enhance clarity and improve reader comprehension.

These three approaches to aggregating topsoil data were developed to determine the minimum sampling effort for S-metolachlor concentration and $\delta^{13}C$ *required to minimise uncertainties in PiBEACH model outputs.*

Page 14 – line 401:

How did you arrive at these thresholds?
According to the classification proposed by Kling et al. (2012) and subsequently adopted by Towner et al. (2019), the goodness-of-fit between simulated and observed variables is categorised as “very poor” for KGE ≤ 0, “poor” for KGE ≤ 0.5, “intermediate” for KGE ≤ 0.75, and “good” for KGE > 0.75. In this study, the intermediate threshold of KGE > 0.5 was retained for both S-metolachlor concentration in topsoil (KGE_{SM} >0.5) and discharge (KGE_Q >0.5) at the outlet. A more stringent threshold was applied to the weekly topsoil $\delta^{13}C$ (KGE_δ > 0.8) in order to better leverage this information as an indicator of degradation. These thresholds were ultimately retained as a compromise to ensure the selection of simulations of at least intermediate quality, while maintaining a sufficient number of parameter sets to derive outputs with robust confidence intervals.

The sentence has been modified to clarify the rationale underlying the selection of specific KGE thresholds, thereby providing a more transparent justification for their application in this study, supported by the inclusion of two additional references.

Kling, H., Fuchs, M., and Paulin, M.: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, J. Hydrol., 424, 264–277, https://doi.org/10.1016/j.jhydrol.2012.01.011, 2012.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., Coughlan de Perez, E., and Stephens, E. M.: Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin, Hydrol. Earth Syst. Sci., 23, 3057–3080, https://doi.org/10.5194/hess-23-3057-2019, 2019.

*According to the classification proposed by Kling et al. (2012) and subsequently adopted by Towner et al. (2019), the goodness-of-fit between simulated and observed variables is categorised as "very poor" for KGE ≤ 0, "poor" for KGE ≤ 0.5, "intermediate" for KGE ≤ 0.75, and "good" for KGE > 0.75, t*he threshold to retain acceptable model results runs (out of 2500 simulation runs) for topsoil S-metolachlor degradation and transport was set to KGE_{SM} >0.5 and KGE_Q >0.5. *An additional and more stringent criterion (KGE_δ > 0.8) was applied to weekly topsoil $\delta^{13}C$ data to maximise its value as an indicator of degradation processes. These thresholds were ultimately selected as a compromise, enabling the retention of simulations with at least intermediate accuracy while preserving a sufficient number of parameter sets to support the derivation of outputs with robust confidence intervals.*

Page 14 – line 402:

Hypotheses generally belong in the Introduction.
This sentence has been removed, as the underlying hypothesis is already stated in Introduction.

Page 14 – line 406:
This comes out of the blue. The last step mentioned is step 2.
The sentence has been revised to accurately indicate that this step corresponds to the final stage of the GLUE framework.
L. 433: In the final step of the GLUE procedure, the distributions of the 25 most sensitive parameters were extracted from the subset of acceptable parameter sets, i.e. KGE_{SM} >0.5 and KGE_Q >0.5 and KGE_δ >0.8. PiBEACH outputs were then expressed as the mean considering the 95 % confidence intervals based on these parameter sets, excluding the lower 2.5% and the upper 2.5% of acceptable simulations.

Page 14 – line 407:
The population of acceptable parameter sets gives a range of acceptable values for each of these parameters, so how come they have a confidence interval? Did you determine the joint distribution of all parameters? Then I can see how you can arrive at confidence intervals, but

that step is missing. In any case, how reliable are the parameter statistics, bases as they are on a sparse sampling (Latin Hypercube) of a 25-dimension parameter space? I just saw that you explain some of this at the end of the paragraph, which makes the line of thought in the paragraph hard to follow, so please rewrite it. I see from there that you did not pursue the joint distribution (which is not trivial for 25 parameters). It would be nice to know what kind of distributions you found, or did you assume normal distributions for the lot?)

The methodology for deriving the 95% ensemble confidence interval for the six PiBEACH outputs, as detailed from lines 405 to 413, has been refined to improve clarity and remove redundancy. Specifically, this involved the systematic exclusion of the lower and upper 2.5% of acceptable simulations to determine the 2.5th and 97.5th percentiles, which are illustrated in Figures 3 and 5. Additionally, the posterior distributions of the 25 parameters retained were provided in Table S3 of the Supplement, delineated by the 2.5th and 97.5th percentiles. As the distributions were not normal for all parameters, as illustrated with the K_{OC} distribution, (Figure S8 in the Supplement), the range (2.5th and 97.5th percentiles) were derived from exclusion of the lower and upper 2.5% of each parameter.

We have now revised the sentence to enhance clarity and improve reader comprehension:
L. 433: In the final step of the GLUE procedure, the distributions of the 25 most sensitive parameters were extracted from the subset of acceptable parameter sets, i.e. KGE_{SM} >0.5 and KGE_{Q} >0.5 and KGE_{δ} >0.8. PiBEACH outputs were then expressed as the mean considering the 95 % confidence intervals based on these parameter sets, excluding the lower 2.5% and the upper 2.5% of acceptable simulations.

Page 14 – line 407:
This is a valid point, but it needs to be made elsewhere. It does not have anything to do with the purpose of section 2.6)

We have now moved this statement in the section 2.4 dedicated to PiBEACH model description:
L 332: Calibrating the PiBEACH parameters, as detailed in Section 2.6, was challenging and warranted the collection of an extensive dataset across the catchment throughout the growing season. This unique dataset incorporated isotopic signatures and comprised 103 topsoil samples analysed for S-metolachlor concentration and $\delta^{13}C$, 115 daily discharge measurements, and 51 river outlet samples with corresponding S-metolachlor concentrations.

Page 15 – line 414:
This is useful.

Page 15 – line 430:
Simulated topsoil...

The sentence has been modified:

Simulated topsoil (z_0) water contents showed substantial variability, consistent with weekly field measurements (Fig. 3A) and previous application of the BEACH model in catchments with similar soils, crops and conditions (Sheikh et al., 2009).

Page 15 – line 432:
Use this term when you introduce KGE for the first time, not here.

The sentence has been modified:
Simulated discharges at the catchment outlet closely matched observations (Fig. 3B), with a maximum KGE_Q of 0.75, demonstrating the model's ability to capture prevailing hydrological dynamics.

The full term was introduced at its first occurrence.
L. 403: For the second step of the GLUE method, the Kling-Gupta Efficiency (KGE) (Gupta et al., 2009) metric was adopted as the objective function to maximize during calibration.

Page 15 – line 432:
With a KGE of 0.75, that's a bit optimistic, as Fig 3B shows.

The sentence has been revised to moderate the assessment of the PiBEACH model's performance in simulating daily discharge, as it was not primarily designed for this purpose, while still highlighting its ability to capture the prevailing hydrological dynamics.

Simulated discharges at the catchment outlet showed reasonable agreement with observations (Fig. 3B), with a maximum KGE_Q of 0.75, demonstrating the model's ability to capture prevailing hydrological dynamics.

Page 17 – line 452:
"repetitive"

The sentence has been revised to eliminate redundancy and avoid repetition of information.

L. 479: Out of 2500 simulation runs ~~using Latin Hypercube sampling~~, 672 were deemed acceptable based on hydrological and concentration performance criteria (KGE_Q >0.5 and KGE_{SM} >0.5).

Page 17 – line 455:
... ensemble of acceptable simulations to 244.

We have modified the sentence to clarify the term "ensemble":

L. 481: *Applying an additional constraint based on isotope data (KGE_δ >0.8) further reduced the ensemble of acceptable simulations to 244 simulations.*

Page 17 – line 461:

This belongs in methodology, not here.

This introductory sentence has been revised to minimise redundancy regarding the methodological objectives, while retaining a concise reminder of the two calibration strategies and their respective designations.

L. 487: This section underlines the benefit of incorporating topsoil CSIA data during model calibration (WIC: with isotope constraint) compared to no isotope constraint (NIC).

Page X – line Y:
This paragraph suggest that you can make more accurate predictions, but I suspect you are only reporting improved data fitting, because the methodology suggest that you calibrated but not validated your model. See my comment above about the need to have the discussion correctly reflect what exactly you did with your model. If you have indeed calibrated your model on the full data set, the text here would be misleading.

This paragraph illustrates how equifinality in parameter estimation—particularly for DT_{50} related to pesticide dissipation—can be reduced by incorporating both the isotopic signature and concentration of S-metolachlor (WIC), rather than relying solely on topsoil concentration data (NIC). In both cases, the datasets were used during model calibration. Calibration with WIC proved more effective, resulting in a narrower acceptable range for DT_{50} compared to NIC. Notably, the full 2016 dataset was used for both model development and calibration, without a separate validation phase. We have revised the sentence to emphasize the reduction of equifinality during the calibration process.
L. 499: Reducing uncertainty in estimates of pesticide degradation in soil during the calibration of reactive transport models during calibration is crucial, as degradation half-lives can vary by one order of magnitude depending on the compound (Wang et al., 2018), largely affected by hydro-climatic and soil conditions. In this study, the WIC calibration yielded mean $DT_{50,Ref}$ below 20 days, with low standard deviations (SD <7 days; Fig. 4), indicating that aerobic degradation of S-metolachlor, typically reported between 14 and 21 days (Lewis et al., 2016), was the dominant process in Alteckendorf topsoil. In contrast, anaerobic degradation, characterized by longer half-lives (DT_{50} = 23 - 62 days; Seybold et al., 2001; Long et al., 2014), appeared to play a limited role.

Page 19 – line 598:
You are probably correct, but you cannot claim this so strongly based on an unvalidated model, which is what I believe you have here.

We have revised the sentence to highlight that the integration of compound-specific isotope analysis (CSIA) enhances the calibration step of the PiBEACH model.
L. 521: These findings underscore the importance of site-specific calibration in CSIA applications and highlight the value of model ensemble approaches in capturing the range of degradation processes in heterogeneous agro-ecosystems. While previously reported εC values may slightly overestimate degradation in some field settings, the calibration of ensemble modelling with integrated CSIA data provides a more robust and field-relevant assessment of

pesticide transformation. However, these findings necessitate further confirmation with a validation dataset, which was not available for the targeted catchment.

Page 19 – line 514:
This suggests prediction, but it really is just fitting, isn't it?

Observed S-metolachlor exports at the catchment outlet were not used during model calibration and therefore serve as a form of partial validation. Only discharge at the outlet was calibrated, with a Kling–Gupta Efficiency (KGE_Q) exceeding 0.5. Consequently, our classification of the results as a 'slight overestimate' is supported by the comparison between simulated and observed S-metolachlor exports, which was performed independently of direct calibration to those data.
We have revised this sentence to clarify this point:
L. 541: The model slightly overestimated the export of S-metolachlor to the outlet (2 ± 6%) in comparison to the observed values (0.5 ± 0.1%), which were not mobilised during calibration. However, this difference remains within the model's uncertainty range.

Page 19 – lines 515 and 516:
river water ? fits ?

We have revised the sentence as follows:
L. 543: It is important to note that observed exports were based solely on the dissolved phase, as particulate-bound S-metolachlor (> 0.7 µm) remained below quantification limits in all river water samples. The S-metolachlor export metric depends on PiBEACH ability to simulate daily discharge

Page 19 – line 519:
I think you are overstating the model performance here.

We have revised the sentence to moderate this statement:
L. 546: Although PiBEACH was originally designed to initialize sub-hourly, event-based distributed models like Openlisem-OLP (Commelin et al., 2024), it also demonstrated a reasonable ability to reproduce daily discharge dynamics (Fig. 3B), consistent with prior applications in similar catchments (Sheikh et al., 2009).

Page 21 – line 528:
Do the massive confidence intervals not suggest that the criteria for acceptance of a set of parameter values should perhaps have been set more strictly? I am not suggesting you should redo the entire GLUE procedure with new KGE thresholds, but it would be worthwhile taking this up in the discussion. The problem with GLUE is that it will be very hard to estimate a priori the best thresholds for the objective function, while the massive computational demand of a single GLUE run makes it prohibitive to narrow down the range of the thresholds by trial and error. This is not something you can or need to resolve, but you can write a not too long paragraph about it.

We acknowledge the Reviewer's comment and have added a new sentence addressing the question of the optimal thresholds for the objective function.

L. 569: The wide 95% confidence intervals observed for the six metrics suggest that the thresholds used for the objective function, particularly $KGE_{SM} > 0.5$ and $KGE_{Q} > 0.5$, may need to be increased to reduce uncertainty. As previously noted in the application of the GLUE method (Jin et al., 2010), selecting an optimal threshold is inherently challenging, as it involves a trade-off between the computational effort required to retain a sufficient number of acceptable simulations and the resulting width of the confidence intervals.

Page 23 – 568:
Fitted?

We have revised the targeted sentence:

L. 600: The limited spatial variation in simulated DT_{50} across the catchment does not diminish the value of PiBEACH's distributed nature, particularly when considering its future integration with distributed event-based models such as OpenLISEM-OLP.

Page 23 – line 578:
But you did not address that in this paper, did you?

We have removed the validation term to prevent any potential misinterpretation, as our work primarily focuses on improving the model calibration process and addressing the issue of parameter equifinality.

L. 609: This study addresses the gap between the increasing complexity of reactive transport models and the limited availability of field data for their calibration.

Page 23 – line 587:
I am confused. I saw no evidence of model validation in the paper. Did I overlook something?
As for the previous sentence, we have removed the validation term to prevent any potential misinterpretation.

L. 618: *In many cases, carbon isotope data ($\delta^{13}C$) alone may be adequate to provide evidence of in situ degradation, thereby supporting pesticide mass balance closure and improving model calibration at the catchment scale.*

Page 24 – line 595:
There is no need to introduce an abbreviation at the end of the paper.

The abbreviation has been removed from the targeted sentence