Responses to reviewers

Reviewer 1

General Comment:

Identifying the sources of problematic sediment in watersheds is a rapidly growing area of research, given its potential to address issues associated with excessive soil erosion and the delivery of sediment (and associated contaminants) to rivers, lakes etc. This research paper aims to make the sediment fingerprinting method more suitable by use for watershed managers and researchers alike. More specifically, it focuses on the differences in outcomes based on tracer selection, emphasizing conservatism and discrimination of tracers. It explains and then uses various range test methods, followed by the Kruskal-Wallis H-test and explores the impacts of using a discriminant function analysis. Additionally, it uses a newer method, the consensus method to compare model outputs with those derived from the conventional three-step method (TSM). Finally, it uses virtual mixtures to compare model output, using data collected from a lake in Japan and the various sources contributing to the lake. The researchers found that the relatively new consensus method (CM) can be too restrictive, as can certain range tests, and that testing the output of models that used and didn't use the DFA is advised. Overall, this is a very useful paper for the sediment fingerprinting community but would benefit from certain clarifications and changes as suggested below.

We thank the reviewer for his/her general positive evaluation of our manuscript. Please find below our replies to the specific comments.

Specific comments:	Authors responses
One area that needs clarification is on the mixed use of the terms theoretical, predicted, and observed, particularly when referring to Figures 5-7. This is also unclear when comparing results from the virtual mixtures and the results from the sediment core, as the term 'predicted results' is used interchangeably. Finally, more clarification is needed when explaining the contributions from the sediment core modeling, you note that the contributions were outside of the range of predicted contributions on the virtual mixtures. I am unsure exactly what this means. This is primarily in section 3.2 (starting at line 377).	Regarding the use of "theoretical", "observed" and "predicted" contribution terms, when generating virtual mixtures, a set of contributions is defined. These contributions can be referred to as "observed" or, as done in our manuscript, "theoretical". We prefer to use the term "theoretical" as the contributions are defined by the user, whereas – in our opinion – the use of the term "observed" would be more appropriate to refer to real observations. The "predicted" contribution refers to model outputs and can be associated with either virtual mixtures or field samples. In our opinion, making this distinction is needed when comparing "theoretical" and "predicted" contributions for virtual mixtures. We paid more attention to the definition of these terms.
There is a lot of repetition, especially Section 1 (Introduction) and Section 2.5 (Tracer selection). The authors need to address this.	In order to introduce concepts as clearly as possible for the reader, we tried to develop terms and definitions, which, as you mentioned, leads to some repetitions. With the help of your comments, we hope that we clarified our statements.

Line 73: if specifics around the sources that contribute to the formation of target sediments are going to be mentioned in the parenthetical, I would suggest adding more common sources, like banks or roads. Is suspended matter a source when often suspended sediments are the target sediment? This seems a little confusing.	Thanks, we added more common sources as suggested (LL. 76-77). Indeed, suspended matter are often a target sediment. However, it is possible to use them as source when assessing contribution from different rivers converging a lake, a pond or a bay for example. We removed it, as it was confusing (LL. 76).
Line 74: Is sediment transport the physical mechanism or is it a product of river/stream discharge and other erosive behaviour? My thinking is the latter.	Thanks, we clarified the definition of sediment transport as a physical mechanism depending and resulting from water/discharge and other erosive processes (LL. 79-80).
Line 79: <63 µm is the most commonly used size fraction, but many studies use a wide range of sizes for different reasons (e.g., targeting rivers with high concentrations of fine sand, etc.). It would be useful to add why <63 µm is not always used.	Thanks. Indeed, 63 µm is a meaningful threshold when studying properties contained in or sorbed onto clays or silts (e.g. radionuclides, pesticides, heavy metals, etc) we added specification about that (LL. 85-87). Nevertheless, other particle sizes (i.e. sand or clays) can provide useful information (e.g. about mineralogy, geology) for catchment heterogeneity description.
Line 86 (and elsewhere): when giving details about each of the three steps in the TSM, it would be useful to continuously refer back to which step equates with each test. For example, in line 86, start with: "The first step in the TSM, which assesses property conservativeness, …" Referring back to paragraph that outlines each step (lines 66-69) should be done throughout section 1.	We have clarified to which step of TSM each test corresponds to. Please see L. 94, 105 and 111.
Line 94: the authors do a very good job of explaining the different statistical analyses for the source material, but do not mention the source samples and whether they are just looking at the mean sediment samples, the min/max, etc. I think most readers will understand all sediment samples should fall in between the range of source values, but it worth explicitly noting.	Thank you for your comment, we added some details to ensure that range tests will be well understood (LL. 96-103).
Line 98: I think more detail is needed as to what the results of the KW test look like and/or specifically do. Does a significant value for a single tracer denote that it is discriminatory across all sources? It's also important to note the use, in some studies, of further post-hoc testing (such as the Dunn's Test) to determine the discrimination potential of each individual source (e.g., forests vs. agriculture vs. roads).	We added some details about the meaning of Kruskal-Wallis H-test results and the use of post-hoc tests for more precise description of the discrimination potential with two samples tests (LL. 107-110).
Line 99: You should mention here, as you do later, that the Mann-Whitney U test can be used for 2 sources.	We added a mention of two samples tests for description of the discrimination potential with two-samples tests such as Dunn's test, Mann-

	Whitney U test or Kolmogorov-Smirnov test (LL. 108-109).
Line 100: While you do not use it, in this paragraph it is worth mentioning that other studies use PCA in place of DFA.	Thank you, we added a mention of the use of PCA (LL. 111-112).
Line 106: It should be noted whether the consensus method uses the range test first or if it ignores the range test all together.	As notice LL. 120-121, consensus method consists of two tests: conservativeness index and consensus ranking. Therefore, CM not used range test for tracer selection.
Line 128: It is unusual to divide research objectives into (a) and (b). Please re-assess this.	We modified objectives' numbering (LL. 138-141).
Line 138 etc: Where relevant, please ensure that percent totals = 100%	We added the bare soil category, which we did not mentioned previously as it covered a very small surface and was indicated on the map (LL. 148, Fig.1).
Figure 1. Given that FDNPP is not mentioned, please remove this from the caption.	FDNPP is mentioned in the box on the top right corner of the map (Fig.1)
Line 145: Does any of this precipitation fall as snow? If so, it may be worth mentioning this.	We mentioned the occurrence of snowfall in winter (LL. 159).
Line 161: There needs to be more explanation and justification why the 0-5 cm increments (i.e., most recent sediment) were not used for this study, and that the 6-36 cm depth range represents the most stable land use period. Was the core dated? If so, then please explain and provide this information.	The main idea of the 1 cm increments is to achieve a high-resolution study of the sediment core in order to reconstruct the strong and rapid land use changes that have affected the catchment (i.e. decontamination works). The core has been dated and interpreted, but the results will be presented in a separate paper focusing on the case study, however, we added some details about it (LL. 174-176).
Line 165: were any statistical tests run to show that these samples from the Niida River catchment were not different in the tested soil properties than those from the Hayama Lake catchment. Also, as others may disagree in principal with using samples from outside of the watershed, it may be worth citing the work by Williamson et al., (2023), which shows that source samples from elsewhere can be used in some circumstances Williamson TN, Fitzpatrick FA, Kreiling RM (2023) Building a library of source samples for sediment fingerprinting – Potential and proof of concept. Journal of Environmental Management 333:117254. https://doi.org/10.1016/j.jenvman.2023.117254	Indeed, as notice LL.182-183, a KS test was computed to ensure the similarity of soil sample properties from both catchments. Following your recommendation, we added some precisions in the manuscript (LL.179-183).
Line 175: were there properties where concentrations fell below the detection limit? If so, what was done with those values? If not, ignore this comment.	For elemental geochemistry properties analysed by XRF spectrometry, no properties were below detection limit.
Lines 177-179: This is confusing and needs clearer explanation.	We added an example, we hope that will make it clearer (LL. 193-196).

Line 250: does virtual tracers mean virtual properties (i.e., elements, reflectance, etc.)?	We changed "tracers" to "properties" as when running CI test tracers are not identified yet (L.270).
Line 266: Why was a score above 70 chosen by Lizaga et al. (2020a), and is it a hard line? Additionally, there seems to be no mention of the issue of underdetermined models, which is avoided by using n-1 tracers (n=sources). Does this matter using the CM/FingerPro?	Our initial idea was to compare existing approaches as they are described in the literature. In the CM, CR is defined with a threshold of 70, which could/should be discussed, but as we did not encounter the limitations of this test in our study, we did not include it in our discussion. We have added a mention of the underdetermined models (L.294).
Line 274: The issue of normality comes up frequently in using MixSIAR, but it does not seem a consensus has been reached. From the cited paper (Laceby et al. 2021b), it seems that there was no significant difference in untransformed data. This might be something that should be included in the discussion.	It was a mistake, thanks for catching this. We removed the sentence (LL.294-295).
Line 330: there should be more detail in this section as to which properties had values nearer to zero (and would have been kept) and which properties were far from zero. Based on my understanding from the explanation of the CM in the methods section, there would be a range of CI values. This is of particular interest because as you point out, there are many ways to implement the range test, but only one way to calculate the CI. Also, which of the four properties had the highest CR score? The lowest?	We added this precision about CI value in the manuscript (LL. 268-269) and expanded CI and CR results for properties close to thresholds (LL. 351-356).
Line 340: What does 'moderate' mean when refereeing the effect of the use of the DFA? I might be inclined to remove that part of the sentence and just write "The effect of the DFA stepwise selection was to mainly modify the prediction…"	Thanks for the suggestion, we modified the manuscript (LL. 363-366).
Line 342: I would add at the end of the sentence when the DFA was utilised.	We added mention of the DFA at the end of the sentence (L. 366).
Line 377: it gets a bit confusing as to what is the virtual mixture results vs. the sediment core results because of the use of 'predicted source contributions for the sediment core samples'. It might be simpler to just refer to those as source contributions.	To make the manuscript clearer, we have removed this part from Results section.
Line 379: this sentence is confusing. Does it mean that the contributions to the core samples were outside the range of the virtual mixture contributions?	To make the manuscript clearer, we have removed this part from Results section.
Line 388: Please clarify what "the DFA stepwise selection tended to reduce the number of	To make the manuscript clearer, we have removed this part from Results section.

matching sediment sample predicted contributions" means. What is matching sediment sample predicted contributions?	
Line 406: Was any dating done to determine the relative time period of sediment contributions? This would be interesting, and may explain changes in source contributions if there were any historic level flooding and/or land use changes.	We totally agree and this will be discussed in details in a specifically-dedicated article as an original method of relative dating based on typhoon occurrence reconstruction had to be developed to this end.
Line 409: same intrinsic information as what? The logic here is difficult to follow.	We removed the term intrinsic as it was not necessary (LL. 418).
Line 440: you mention the importance of grain size and in the methods note that you sieved to 63um, but did you also test to see if there was a difference in the D50 or SSA of the sediments vs. source samples?	As source and sediment were sieved to 63 µm, it minimises the potential impact of sorting. However, it should be interesting to compare D50 or SSA in further work.
Line 461: So the CI must be equal to 0 for the property to be used in modeling, but is that a choice made by the model developer or could that threshold be changed for other a priori conservative properties with a score close to 0, as you mention? For example, if more properties were needed to ensure that the model isn't underdetermined, could a property with a score of 0.1 be included? This may be outside the scope of this paper, but a few sentences to this effect might be useful, perhaps as an extension to the statement on line 467.	Thanks for the comment, we have added the CI threshold to 0 in the Materials and Methods (LL. 268-269). In addition, we have extended the discussion with the idea of modifying the CI threshold to include other properties (LL. 472-474).
Line 472: The correction factors were not necessarily considered useful, but understanding if there are significant particle size differences between source and sediment is important, particularly for certain geochemical properties, as you note in line 440.	Thanks for your comment, we have modified this part to be more precise (LL.478-480).
Line 495: relevant in what regard? My assumption is that you are stating that researchers should measure relevance based on the outcomes of virtual mixtures run simultaneously with field data, but it needs clarification.	To make the manuscript clearer, we deleted this sentence (LL.496-497).
Line 566/567: this sentence needs clarification, 'a greater or lesser number of sample predicted contributions fell outside the range ..." What is meant by a greater or lesser number?	We have rewritten this paragraph to make it clearer about the comparison between the space of prediction on the virtual mixtures and the actual sediment samples, and therefore we think it is more appropriate to talk about the transferability of the statistics (LL. 568-571).
Technical corrections:	

The English does require some improvement; below are some comments. For the reference list, please include all of the author's initials. I think some of the figures could be improved in terms of differentiating lines etc. Some of the colours are too close to each other, and some symbols are not legible (e.g., sediment sample value and measurement uncertainty on Fig 9).	We thank the reviewer for all his/her comments and suggestions, we corrected the typos and modified the figures accordingly.
Line 55: remove 'might'.	Please see L. 58.
Line 69: The last sentence needs some editing – "the third step of this approach consists of selecting optimal tracers…" or something similar.	Please see LL. 73-74.
Line 71: I would consider changing it to under different 'land usages or covers'.	Please see L. 76.
Line 108: if acronyms are introduced, they should be used throughput.	We were more careful about the use of acronyms, especially about CI and CR (LL. 120-121, 122-123….).
Line 134 and throughout: use "Hayama Lake".	Please see 144 and throughout
Line 138: move 'respectively' to after 1%	We remove "respectively", please see LL. 147-148.
Line 206: add 'and' after the comma after literature.	Please see L. 226.
Line 274: change to MixSIAR.	Please see L. 292.
Line 295: indicates 'that the mean'…	Please see L. 315.
Line 322: Remove 'only'. No test only identified Ti as conservative, but many of them did identify it as such.	Please see changes L.342.
Line 374: lowest?	Indeed, the Mean criterion get the lowest/poorest prediction quality statistics among the tracer selections (Fig. 4).
Line 366: This should be Fig 7?	Indeed, we corrected the miss numbering of references to figures (LL. 389).
Figure 6: The text in the caption and the axis should match. Predicted, theoretical and observed are all used. So I am assuming it is showing the virtual model mixtures (theoretical) vs. virtual model output (predicted).	Indeed, we corrected figures captions. See figures 6, 7 and 8.
Figure 7: Same issue, either use theoretical or observed.	We corrected figure 7 caption.
Line 401: remove 'really', perhaps change to "did not have a strong impact on the trends…"	Please see LL. 410-411.
Line 412: "For most of them", what is them referring to?	We were more precise about what we were referring to, please see L. 421.
Line 425: Remove 'a more or less'. Not clear what it means here.	We change the sentence to be clearer (LL. 433-434).
Line 439: remove the comma and 'and' after sheets.	Please see L.447.

Reviewer 2

General comment:

Despite the large number of fingerprinting studies identifying and quantifying sources of sediment under different conditions and scenarios, authors highlight the need of keep on working in the field as its use is still limited due to the complexity of the approach and their inherent limitations. This manuscript presents a detailed study and comparison of some of the steps followed in this kind of studies and could be an initial step to make this approach suitable by use for researchers but still far for managers and farmers.

We appreciate that the reviewer gave our manuscript a general favourable evaluation. Please find our replies to the specific comments below.

Specific comments:	Authors responses
Please, explain why only the 2000-63µm fraction is kept for evaluation in these studies (Introduction) and in this particular one (L 169). Have the authors made any kind of exploratory statistical analysis to evaluate possible differences between soil/sediment particle sizes before keeping only the 2000-63µm fraction? Could you provide any kind of information about the samples' particle size distribution (soil and sediment)?	Some additions have been made about the context of the study and the relevance of sieving material to 63µm (LL. 184-185). Particle size threshold choices including that at 63µm in other studies is also discussed now (LL. 84-87). No exploratory analysis of soil/particle size was undertaken prior to sieving at 63µm as samples were initially collected to investigate 137Cs transfers in Fukushima river systems, and 137Cs is known to be bound to the finest particles (LL. 151-153).
I miss how the sampling design (soil and sediment at spatial and temporal scales) is addressed in fingerprinting studies as another source of uncertainty/variability. Please, include general information (Introduction) but also a bit more of detail regarding this study in particular. In section 2.2 please explain better why these depth increments, what "stable land use period" means, etc.	In our paper we have focused on the comparison of tracer selection methods, the exact description of the full fingerprint sampling design may be beyond our scope. The main idea of the 1cm increments is to achieve a high-resolution study of the sediment core in order to reconstruct the strong and rapid land use changes that have affected the catchment (i.e. decontamination works). The core has been dated and interpreted, but the results are presented in a separate paper focusing on the case study, but we have added some details about it (LL. 173-177).
The term "theoretical" is frequently used but it is not clear to me what the authors mean in each case. Observed from virtual mixtures? Observed from the field samples? Please, clarify the meaning respectively.	Regarding the use of "theoretical", "observed" and "predicted" contribution terms, when generating virtual mixtures, a set of contributions is defined. These contributions can be referred to as "observed" or, as done in our manuscript, "theoretical". We prefer to use the term "theoretical" as the contributions are defined by the user, whereas – in our opinion – the use of the term "observed" would be more appropriate to refer to real observations. The "predicted" contribution refers to model outputs and can be associated with either virtual

	mixtures or field samples. In our opinion, making this distinction is needed when comparing "theoretical" and "predicted" contributions for virtual mixtures. We paid more attention to the definition of these terms.
Please delete in Fig. 1 “FDNPP: Fukushima Dai-ichi Nuclear power plant”. There is no mention to it anywhere.	FDNPP is mentioned in the box on the top right corner of the map (Fig.1)
L 203-204/L 424-475: “To be conservative, all the sample property values should lie within the source range” but how conservativity in time is addressed in this study?	When assessing conservativity by comparing the range of properties in each sediment core layer sample and in potential sources, we somehow provided an assessment of the conservativity of those properties throughout time. Indeed, if the properties had changed over time, they would show modifications along the core, and deeper sediment sample values may lie beyond the range of properties found in current sources and therefore be considered as no longer conservative.
L 250: What does virtual tracer mean?	We changed "tracers" to "properties" as when running the CI test, as tracers were not identified yet (L. 270).
L 377-379. Could you please rewrite this statement? It is unclear.	To make the manuscript clearer, we have removed this part from Results section. We add more details about transferability between virtual mixtures and real samples.
L 409-410. Please, clarify.	We modified the sentence to clarify it (LL. 418-419)
Conclusions section is a brief summary of the manuscript. However, I could not find much about recommendations for practitioners (L 575-578) and how to make fingerprinting studies more usable. Please, expound on.	One of our main results is that the fingerprinting technique may be too sensitive to tracer selection to be used with confidence (i.e. biased predictions and statistics) as a quantitative tool for landscape management without taking some precautions. We identified two range tests that provided reliable selections of tracers and realistic statistics based on the evaluation of virtual mixtures. However, further work remains needed to develop and implement reliable selections of conservative properties.
Please, improve the legibility of all figures.	Size of figure elements has been increased to improve the legibility.
Use the same format for the references list within the main text and also supplementary material. Please, change “Kanonika”.	We homogenised the fonts in both the main text and in supplementary materials, and corrected Kanonica spelling.

English could be checked and improved to make it more formal avoiding colloquial expressions and trying to be more precise.	English had been checked by a native speaker.