Response to second round of reviewer comments

A. Lipp, P. Vermeesch

April 3, 2023

Associate Editor

I have read the updated manuscript and the referee's letters. I believe that accounting for the one remaining issue of referee 1 and fixing the few remaining issues will move the manuscript to a state where it will become a valuable contribution to the journal's portfolio.

We're pleased to hear this positive feedback. We have responded to all of the reviewers' remaining comments below and, where appropriate, modified the manuscript accordingly.

Reviewer 1

I feel that this is the only author's comment where a response is needed because my initial comment was oddly received. I did not contest or intend to diminish previous work. The presented work will add a feature to an existing toolbox, and the revised manuscript, with its discussion makes a justified contribution to the field. However, to me, it was and still is, more of a brief review of an existing metric one can use additionally under particular circumstances. There is nothing wrong with it, but we should also be realistic about the potential impact the presented manuscript will have.

We apologise for any miscommunication, and did not intend to be provide an odd response. We simply feel strongly that solid quantitative foundations are important to the field. Whilst the reviewer we are sure did not intend to diminish previous work, I'm sure they can understand our motivation for making sure that it is not neglected. We hope that the recasting the manuscript as a 'Short Communication' best suits the nature of this study, and does not seek to 'overstate' potential impacts.

Minor Comments

L8: Perhaps, '...as an additional and alternative metric...' to emphasise add-on character of the metric.

We have added this to the abstract as suggested.

L190 and later: please decide whether to use mono space letters of Python and R or not. I suggest using the mono space letters for source code only. The canonical writing capitalises the first letter of "Python", although the Python logo uses a small "p".

Thank you for raising this. We have removed mono spacing for referring to programming languages and consistently use a capital 'P' for Python.

L188: Please set URLs as (valid) links throughout (also, e.g., L194-195)

This has been amended as suggested.

Figures 1-6: Please unify the axis labels (e.g., 'Age, Ma' vs 'age [Ma]'). Ma should not read "ma"

Thank you for pointing this out, it has been revised as proposed.

Figure 5

- **Figure caption: add white space between "a)" and KDE.**
- This has been added.
- **Furthermore, it would be beneficial if the R code reflects the code used to create the figure (not all of it but the W2 part. Currently it looks a lot different from what you have in the manuscript (even if I play with the other graphical parameters); nothing related to the random state parameters. The online GUI shows the same result.**
- Unfortunately its unclear to us how this could be improved as, barring a rotation, the structure presented by the reviewer are nearly identical to those we present in Figure 5. We do not see how the results presented here result in different interpretations. We have clarified however that arbitrary rotations and translations (which do not affect the interpretations) are to be expected.
- **If you feel that the log conversion is useful (I think it is) you should add it to IsoplotR::mds() or IsoplotR::read.data().**
- It will be considered as an option for the MDS functions in future versions, thank you for raising this.

Please add more white space to the code snippets (as shown below), it is just a more reader friendly code formatting

We have added in white space to the code snippets as suggested.

L266-268: Please check the reference, it does not look right. The correct citation entry of R packages should be generated using utils:citation(...)

Thank you for raising this, we have corrected this.

L283: I believe that the series number is missing.

We have corrected the reference here.

Reviewer 2

This is now a second review of the manuscript submitted by Lipp and Vermeesch. The authors have addressed many issues raised by the other reviewer and myself. They have significantly expanded the scope of the discussion and included examples of where the Wasserstein distance both works well and fails. This version is significantly improved over the previous one. My remaining reservations with the manuscript address the underlying method of calculating the Wasserstein distance rather than its application as presented in the manuscript. With the exception of item 1 below, I think the manuscript could be published with minor revisions. I

would like to see the authors address item 1 however.

We are pleased to here the reviewer believes the manuscript has been significantly improved on the basis of their constructive feedback. We recognise the concern motivating their first item, and agree that Figure 2 could be potentially confusing if coming from a perspective of aligning peaks. Consequently we have addressed their concern by adding further clarifying text to the manuscript, as they suggest. This response, and others, are detailed below.

1. I maintain that the example in Figure 2 is problematic for both the KS and Wasserstein distances. As stated in the initial review, the translated distribution is most similar to the fixed distribution when their peaks align at 900 Ma or 1100 Ma. When the translated distribution is centered at 1000 Ma, it shares no ages with the fixed distribution (albeit the tails of the distributions overlap). Nevertheless, both the KS and Wasserstein distances indicate that at this point the distributions are most similar. The authors have not addressed this problem either in their response or in the text. Surely at the very least, some explanation for why the behaviour observed in the KS and Wasserstein distances is desirable or intuitive is in order to justify using this example (or these metrics).

We feel that *'the translated distribution is most similar to the fixed distribution when their peaks align at 900 Ma or 1100 Ma'* is not an objective fact, but a matter of subjectivity. As we now discuss in the main manuscript, there is no single distance function that could be applied in all scenarios. This logic extends to this figure too. As a result, we do not believe that this is really a 'problem' that needs to be addressed as suggested. Additionally, we believe the reviewer is neglecting the fact that both metrics do not suggest that, in absolute terms, the central value is a perfect fit. Both W_2 and KS are non-zero at the central point, indicating correctly that the unimodal and bimodal distributions are different. As such we don't feel that they 'fail' in this scenario. W_2 seeks *first* to identify average ages, and *additionally* match up age peaks, so to indicate that it neglects this information is wrong. Nonetheless, we recognise that it could be confusing, and as such we follow the reviewer's suggestion to add an explanation for the behaviour:

> We reiterate that at a translation of 0 Ma, W_2 (and the KS distance) is still non-zero, reflecting the fact that even when the average ages are aligned, the shapes of the uni-modal and bi-modal distributions do not match. This illustrates the tendency of W_2 in geochronological data to prioritise aligning the average ages of distributions *before* considering matching individual peaks. Such behaviour contrasts with approaches that seek to only match probability peaks neglecting any information of absolute ages (e.g., Saylor et al. 2016).

2. I maintain that the absolute distance along the x-axis does not encode geologically meaningful information absent some context (c.f., the geologically meaningful information encoded by KDEs that can be extracted without reference to the geological context and therefore can provide independent evaluation and verification of geological hypotheses (Sharman and Johnstone, 2017). See comment on line 168.

We agree with the reviewer that 'unmixing' approaches such as that proposed in Sharman et al. (2017) can be used to extract source region distributions from detrital distributions, and that these do not depend on absolute distances along the x-axis. Indeed, in Section 3.1 we already propose that such methods are better suited to analyse mixtures samples than MDS using the Wasserstein distance. Consequently, it is not clear how we can further address this comment. Additionally, we emphasise this scenario is just one particular scenario in which geochronological distributional data is analysed, and in many other scenarios absolute distances along the x-axis provides useful geologic information (see Section 3.2 – 3.4).

3. Due to the need to select a metric based on an expected outcome, as the authors suggest, I wonder if this whole enterprise does not descend into circular logic. In other words, the metric

is chosen because we expect a certain conclusion and (lo and behold!) the method confirms that conclusion. Does this return the detrital geochronology/thermochronology back to the realm of subjectively assessing each distribution, even if we then represent that the subjective analysis that we conducted using "objective" metrics?

We recognise the concerns of the reviewer here to an extent, although feel that such concerns are beyond the scope of the manuscript. Further, for hypothesis driven research there will always be expected outcomes to be tested. As a result, an appropriate metric can be chosen on this basis. We agree that the scenario the reviewer proposes of repeatedly re-analysing a dataset with a variety of metrics and selecting the 'best' results is undesirable and would be analogous to 'p-hacking'. We have slightly modified our language in the Discussion to emphasise that the appropriate metric ought be chosen on the basis of the scientific question being answered, rather than on the dataset itself, which we agree could be confusing: *'the most appropriate dissimilarity metric to use will depend on the scientific question being answered.'*.

4. Is there an internal check that would provide an assessment of whether the selected metric is successful? For example, for the data from Degraaf-Surpless et al. (2002), I calculated a stress of 0.14 for the MDS using the KS D value which is pretty high. Does the Wasserstein distance provide a better transformation (realizing that comparing between different metrics when conducting MDS is quite tricky)?

Unfortunately we do not believe that such a simple check exists as what 'success' entails entirely depends on the scientific question being analysed, which will only partially depend on the dissimilarity metric chosen. We argue that 'stress' should not be considered as a measure of the success of MDS either. Stress is just the reflection of whether a dataset can be projected onto a two-dimensional plane without distortion. This is not the same as whether an MDS projection has actually recovered useful latent variables.

For example, consider a scenario where each sample is simply assigned a random number. Then, we calculate distances between samples by calculating the difference between these random numbers. MDS on these distances would produce a map with very low stress because the original values are distributed along a 1D line. However, whilst the map has low stress, it obviously has no actual inherent meaning.

However, we do recognise the importance of stress as an indicate of projcetion fidelity, so we have now added stress values to all of our MDS maps to the figure captions.

5. If the method chosen is dependent on the specifics of the dataset and a prior analysis of the dataset, do the numerical metrics have any potential to illuminate (i.e., uncover latent features of) the dataset? See comment below on line 155.

As discussed above, we do not advocate choosing a metric on the basis of datasets, rather the scientific *question* being analysed. For example, questions related to absolute geologic timings are well suited for analysis using the Wasserstein distance, whereas questions related on quantifying *amounts* of material are best solved using explicit mixing models.

Minor Comments

Line 122: I recommend removing this discussion of linearity. This is really a function not only of the translation of the respective distributions, but also of the shape of the distributions themselves. In other words, translating a bimodal distribution past another bimodal distribution would produce non-linear results for the Wasserstein distance as well.

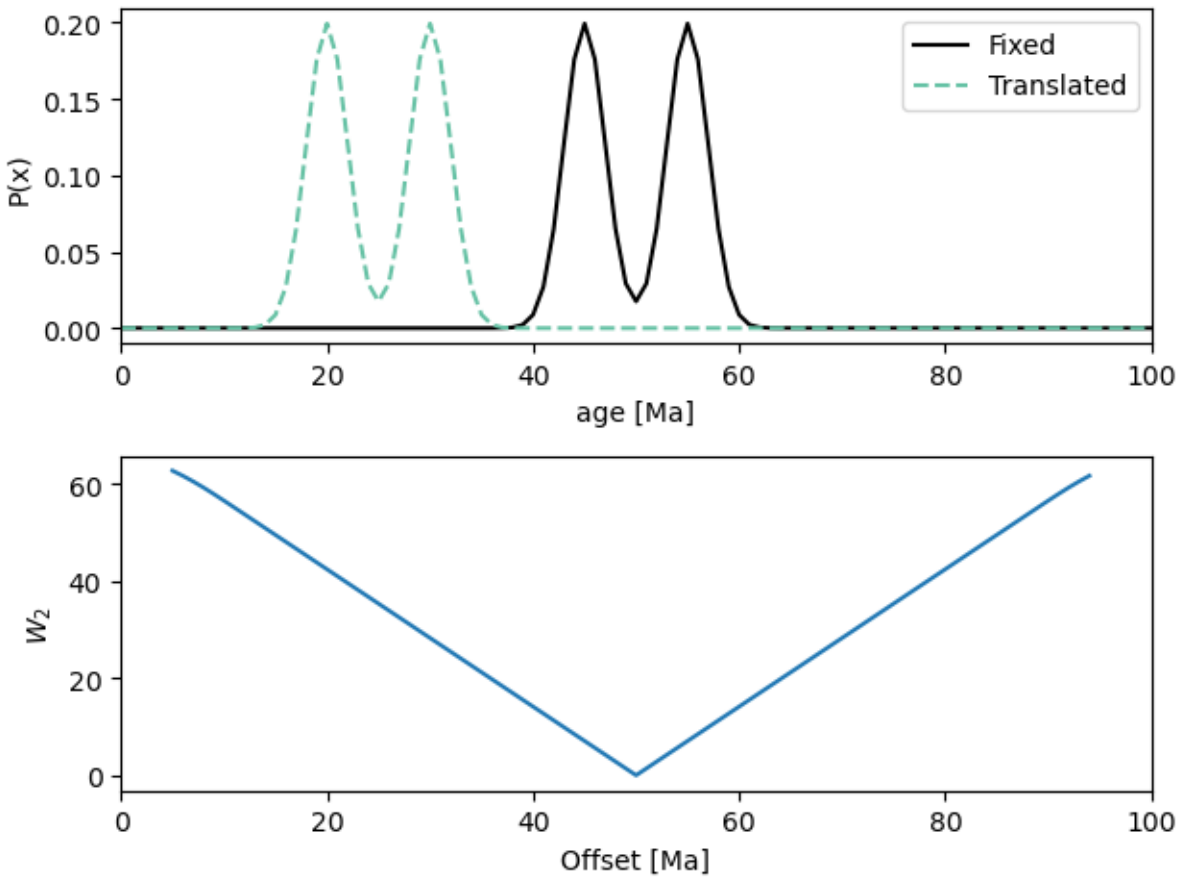

Figure 1: **Translating bimodal distributions past eachother.** Note the linearity of W_2 with respect to offset, contrary to the suggestions of the reviewer.

The reviewer's statement '*translating a bimodal distribution past another bimodal distribution would produce non-linear results for the Wasserstein distance as well*' is factually incorrect. Translating two bimodal distributions past eachother, much like the example in Figure 2 result in a linear change in W_2 as shown in Figure 1 in this document. This linearity is to be expected from Equation 3, where as only the means of the distributions are changing, the change in W_2 will linearly depend on the change in the means (i.e., the offset). Consequently we do not wish to modify our original (correct) language here.

Line 153: I think further caveats are warranted here. I don't think that the Wasserstein distance can be applied with certainty in all instances of identifying upsection trends in provenance changes. As one simple example, take the dataset below (from Smith et al. (2023)) where the Wasserstein distance fails to identify unimodal sample 1CCT3 as a unique population and instead lumps it in with the bimodal 1FCTC166. Counterintuitively the Wasserstein distance also places unimodes 2PCGT190 and 1DCGT243 closer to the bimodal cluster than other bimodal samples such as 1FCTC166, or even samples that share the same modes but in different proportions such as SJMT7. I realize that all of these metrics have their own caveats. I think the take-away for me is that the caveats need to be clear and that the basis for calculating the metric needs to be to be appropriate for the task (geochronology in this case). The authors can address the first of these issues by adding appropriate text to the

manuscript as it is. The second one forms the basis of my deeper concern as outlined in point 2 in the General comments.

We agree that it would be unlikely the Wasserstein distance (or any metric) to automatically identify upsection provenance trends in all possible scenarios. However, we feel that this is an unreasonably high bar to pass, and certainly we do not believe that any dissimilarity metric already in use by the geochronology would pass such a test. Instead, as we discuss in Section 3.2, W_2 is *better* suited for this task than other metrics due to its sensitivity to temporal trends in source age distributions. We also agree with the reviewer that all metrics have caveats, and have attempted to detail those of W_2 in the revised Discussion in the manuscript. We have addressed point 2 above.

Line 155: MDS of the cross-correlation coefficient provides something similar to the Wasserstein distance. There is less clustering between GV-42, -45, -40, and -64 than with the Wasserstein distance, but otherwise it is pretty close. I am not necessarily advocating that the authors present this approach. Rather, I am presenting alternative approaches to consider which may illuminate the data or a way forward.

We agree with the reviewer that this MDS map reasonably identifies the upsection trend of these samples. However, it is perhaps surprising that it struggles to cluster the four unimodal samples given that cross-correlation is so designed to cluster samples with shared peaks. We hope that this manuscript provides a useful discussion for the geochronological community on how dissimilarity metrics ought be chosen given various advantages and disadvantages.

Line 160, 170: What are the stresses associated with these MDS plots?

We have now added 'Stress' values to the captions of all MDS plots.

Line 168: I disagree that the absolute distance long the time-axis provides any useful information here. The interpretation that there is a break in exhumation between WBS7 and WBS8 would be the same if the older samples had modes at 100 Ma, instead of at 1,000 Ma. The relevant information is whether there is significant overlap in age distribution between the northern 4 samples and the southern 4 samples. Beyond this, the comparison between samples is not meaningful. (Obviously, it changes the geological interpretation whether the southern four samples have age modes at 100 Ma or 1,000 Ma, but that is not a function of the intersample comparison. The four southern samples reveal their own geological history without reference to any other samples.

We disagree with the reviewer here as there is an additional temporal trend within the southern samples from the most southern sample (WBS1) to the most northern (WBS8). This ordering of the southern samples is identified by W_2 due to its sensitivity to the time-axis. The KS map however both fails to discover the ordering of the southern samples. Further, we note that the gap between the two clusters on the W_2 MDS map of is proportional to the (significant) relative age gap between the two exhumation signals. The KS map however (which cannot 'see' the different ages of the clusters) suggests that the difference between the two clusters of exhumations is smaller than the variability within each group. From a thermochronological perspective this might erroneously indicate that the different exhumation signals across the physiographic divide are quite similar, and that variations in exhumation history within each cluster are equally as important. So whilst the reviewer is correct that to simply extract two clusters, the KS distance is adequate (although we note that these clusters are poorly defined), but there is more geological information that can be meaningfully extracted from the data if W_2 is used.

References

Saylor, J. E. and K. E. Sundell (2016). "Quantifying comparison of large detrital geochronology data sets". *Geosphere* 12.1, pp. 203–220.

Sharman, G. R. and S. A. Johnstone (2017). "Sediment unmixing using detrital geochronology". *Earth and Planetary Science Letters* 477, pp. 183–194.