Response to Reviewer 2's comments on manuscript egusphere-2022-1200 entitled 'Comparing detrital age spectra, and other geological distributions, using the Wasserstein distance.'

A. Lipp, P. Vermeesch

January 30, 2023

Summary

The authors have not demonstrated that the WS Distance produces geologically meaningful results. At the very least they need to address the concerns raised below before publication. However, the comments below suggest that the WS Distance may not be an appropriate metric to compare geochronological distributions, because of the unique geological implications of minorly distinct age modes (distinct sources) versus multimodal distributions which include some shared age modes (potentially shared sources).

We thank the reviewer for their critical review, which has prompted us to re-evaluate the pros and cons of the W_2 distance compared to the KS-statistic. We agree with the reviewer (below) that '*whether [a dissimilarity measure] is more 'sensible' depends on the application*'. In some cases, this is the W_2 distance, but in other cases, the KS statistic may be better. To help the reader make this judgement, we will expand the paper with a thorough discussion of a number of specific scenarios.

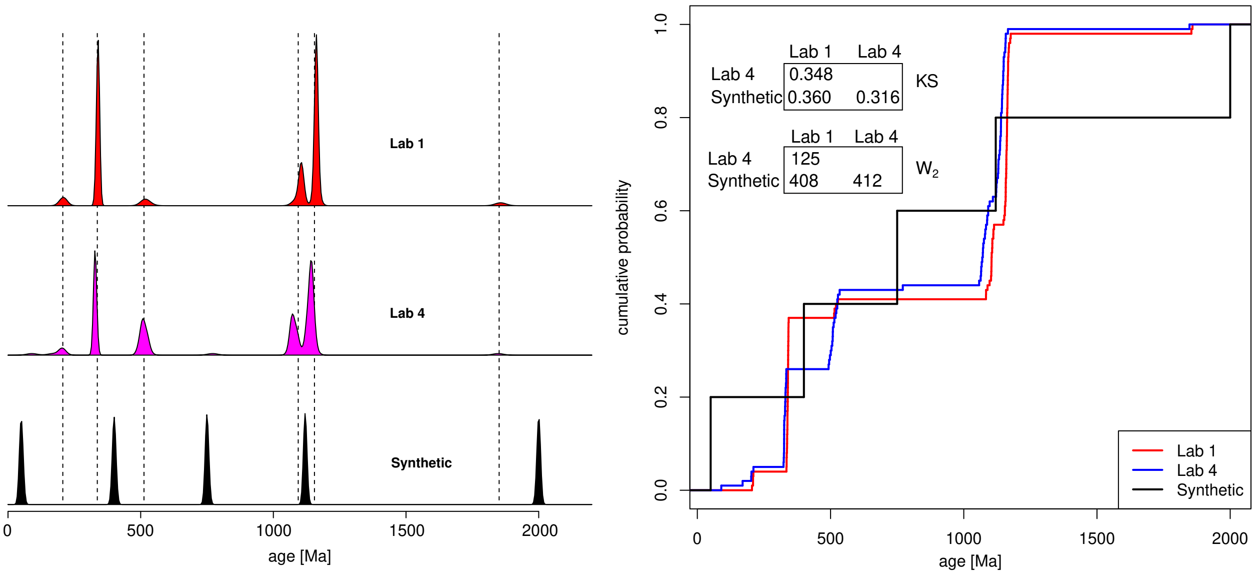
The KS-statistic is very good for detecting the *presence* or absence of specific discrete age components. If one does not care about the *size* of the age differences then the KS statistic is generally the best metric. However, even then there are cases where caution must be exercised and the W_2 distance is preferred:

1. Studies that combine measurements from several laboratories, which are affected by inter-laboratory *biases* (Košler et al. 2013, Figure 1).

Other common geological scenarios in which the magnitude of the time (/horizontal) axis differences *does* matter include:

2. Studies that combine samples of different depositional age, causing the shape of the source age distributions to change as a function of time (Figure 2).
3. Detrital thermochronology, in which age distributions shift in response to thermal signals (Figure 3).
4. Fitting of detrital age distributions such as mass balance calculations for sediment mixing in a river network (Amidon et al. 2005).

In all of the above cases, the W_2 distance is preferred over the KS statistic. Similar considerations apply to other geological distributional data of variables where absolute values (not just theoretical endmembers) are of meaning, notably, grainsize distributions (Weltje et al. 2007).



	Lab 1	Lab 4
Lab 4	0.348	
Synthetic	0.360	0.316

	Lab 1	Lab 4
Lab 4	125	
Synthetic	408	412

Figure 1: KDEs (left) and ECDFs (right) of two samples from the inter-laboratory comparison study of Košler et al. (2013), plus a synthetic sample. Dashed lines mark the true ages of the detrital mixture. According to the KS-statistic, the age distribution produced by Lab 4 is more similar to the synthetic distribution than it is to the distribution produced by Lab 1, despite the absence of any shared age components. The W_2 distance correctly deems the distribution produced by Lab 4 to be closer to that of Lab 1 than to the synthetic mixture.

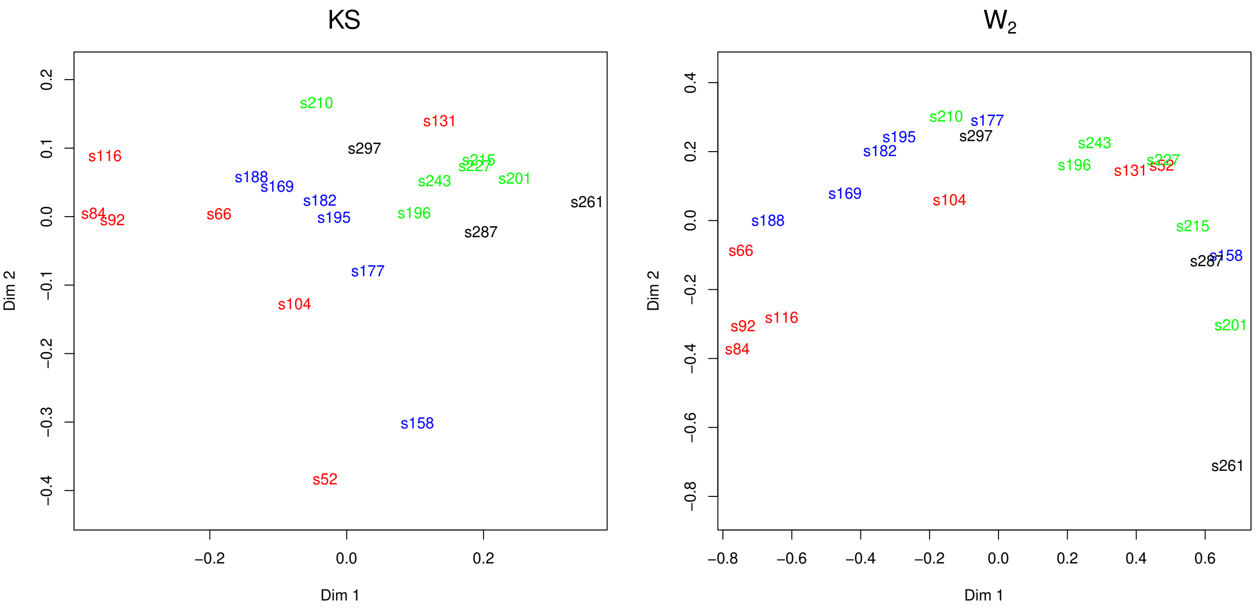


Figure 2: MDS configurations using the KS statistic (left) and the W_2 distance, following a log transform, (right) of DZ U-Pb data for the Coconino drill core of Gehrels et al. (2020). This core samples, from bottom to top, the following Members of the Chinle Formation: Blue Mesa (black); Lower Sonsela (green); Upper Sonsela (blue); and Petrified Forest (red). Both MDS configurations correctly separate these main subdivisions. However, whereas the KS distance implies that the amount of chronological diversity within the Petrified Forest Member equals that between the Petrified Forest Member and the Blue Mesa Member, the W_2 clarifies that the Petrified Forest samples are all quite similar and distinct from the Blue Mesa samples.

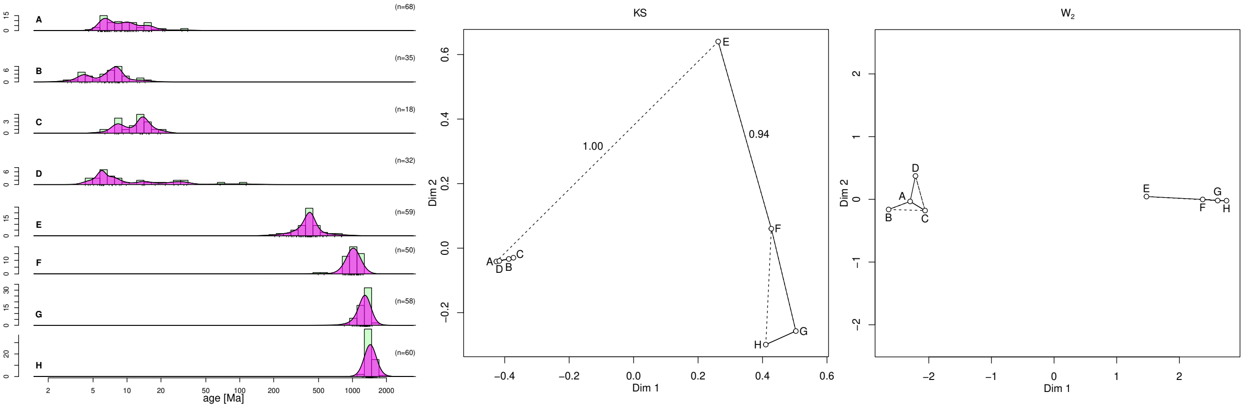


Figure 3: KDEs (left) and MDS configurations (middle and right) for a detrital mica ^{40}Ar/^{39}Ar dataset of Wobus et al. (2003). Samples A–G have been relabelled from the original publication and arranged from north to south across a physiographic transition of the central Himalaya in Nepal. The MDS configuration using the KS statistic (middle) is poorly constrained due to the complete lack of overlap between groups A-D and E-H, respectively. The nearest neighbour lines connect the two groups based on a KS-distance of 1 between samples A and E. The W_2 distance (following a log transform) fares much better: it correctly identifies the two groups, which are separated by the physiographic transition. Note that the W_2 distance also recovers the correct geographical order of the samples (except for A–D, which are statistically indistinguishable from each other).

General Comments

The WS Distance may be more intuitive than the KS distance in many cases, but whether it is more 'sensible' depends on the application. While the toy dataset clearly shows the advantage of WS over KS for assessing simple dissimilarity between sample ages, in many DZ studies it is the degree to which samples share the same sources that is of interest; the absolute difference in age is not directly relevant. For example, if we assume the sources for the samples in the toy dataset each have distinct ages, A and C are no more similar in terms of their sources than A and D, and the equal KS value of 1 (i.e., complete dissimilarity) for (A, C) and (A, D) is actually more informative than the WS values W(A, C) = 1 and W(A, D) = 10.

We refer the reviewer/reader to our comments above about scenario 1 in which the W_2 is and KS is not the most appropriate dissimilarity measure. We argue that the concern the reviewer raises here (and below) of samples being described as mixtures of very well-defined discrete sources of largely irrelevant absolute ages, is just *one* scenario where measures such as the KS distance would be more appropriate. There are many other scenarios (see above) where it is not correct to discount absolute differences in age and thus where the W_2 distance is more appropriate.

The authors need to explore the behavior of WS Distance for multi-modal data sets. For example, what is the interpretation of high versus low WS Distances when comparing multimodal data sets? This may seem intuitive, however, I suspect that the results will not be intuitive given how the WS Distance is calculated.

First, we would like to point out that the KS distance of multimodal datasets does not always yield intuitive results either (see Figure 1). Second, the interpretation of the W_2 distance is actually very intuitive. Assuming the two multimodal datasets are aligned, the W_2 is simply the cost of 'moving' the grains from one distribution into the shape of another. For example, assuming that the two datasets have the same number of grains, we simply sum the (squared) distance that each grain travels when rearranging them from one distribution along the time axis, to the other. When there are different numbers of grains, some of them are 'split', but conceptually it is the same. If the datasets are misaligned, the W_2 is simply an addition of the cost of translating the two datasets to be aligned, and the cost of 'rearranging' the grains. That the W_2 is an additive with respect to a) translating distributions and b) changing their shape is intuitive and helpful.

We explore this more fully using a simple synthetic example (Figure 4) where we mix two unimodal distributions at 500 and 800 Ma in proportions between 0 and 1. This generates a suite of mixture distributions with changing levels of bimodality. We then calculate the W_2 distance between each of these and the 'central' bimodal distribution. We find that W_2 changes nearly linearly with respect to the mixing proportion. We would happily include a similar example to this in the published manuscript.

The geological context of the samples needs to be provided. Even a reproduction of the Morton et al. (2008) Figure 1 would be helpful in terms of understanding which samples would be predicted to be derived from similar sources.

Whilst we are happy to provide further context, we note that Reviewer 1 in fact suggests that this table is superfluous. As both reviewers suggest mutually exclusive modifications to this table we suggest that it is adequate in its current form.

I am fairly surprised that the authors chose samples with n$\leq$60 grains to showcase a new statistical comparison metric. Multiple studies (including one by the co-author) have shown that n$\gg$100 (ideally n>300) are needed for robust statistical comparison. Although it can be argued that samples of that size are unnecessary for simple age distributions such as those presented here, this raises the question of why choose the simplest possible scenario to showcase

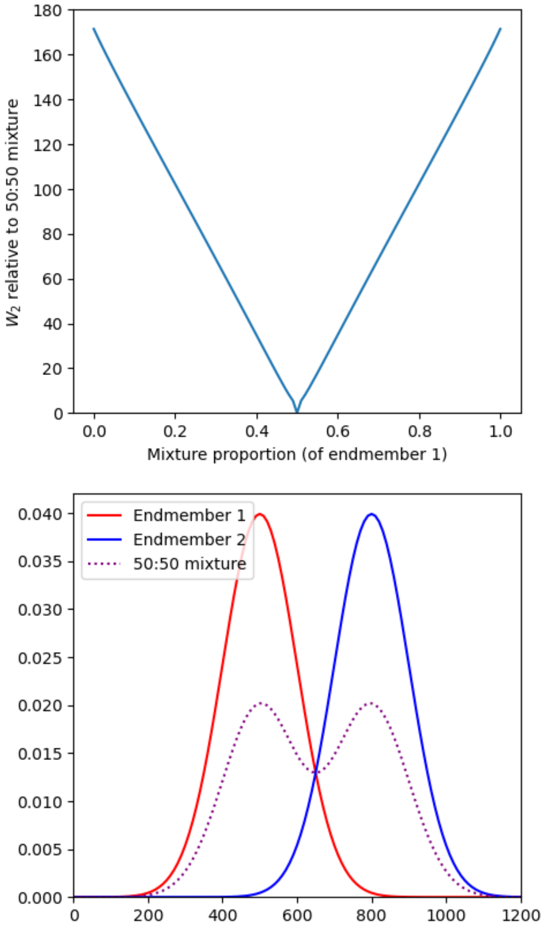


Figure 4: The impact of multi-modal mixing on W_2

a new application of a statistical metric. Rather, it seems that a true breakthrough would be demonstrating that the WS Distance can deal with a previously unresolved problem or elegantly deals with an intensely complex data set.

First, whilst we recognise the general importance of considering the number of grains in analyses, it is not relevant in this particular study for the reasons the reviewer already states. Incidentally, we note that, *because it penalises horizontal distances* the Wasserstein distance is actually very robust to potential variations introduced to small sample sizes.

We stress that the Wasserstein distance is indeed useful for comparing ‘complex’ datasets, as illustrated by the examples provided at the beginning of this response. One interpretation of a dataset’s complexity is its dimensionality. As we discuss in the Introduction, the Wasserstein distance in fact extends easily (and elegantly) to distributional data of multiple dimensions. This topic is further discussed in a separate paper that was accepted by another journal (JGR-ES) pending minor revisions (Vermeesch et al. 2023).

Figure 2 demonstrates problems with both the KS and WS distances. First, the WS distance increases linearly with displacement away from 1000 Ma. However, once the two distributions no longer overlap, they are no longer any more or less similar because the x-axis is age, not distance. Two age distributions that share no age modes are equally dissimilar regardless of the age difference between their modes due to the geological implications of sharing versus not sharing age modes. Hence, the increasing WS Distance with displacement beyond zero overlap is an undesirable trait. Second, both the WS and KS distances indicate that the distributions are most alike (minimum WS and KS distance) when both are centered at 1000 Ma. However, at that point they share no age modes. The green distribution would have an age mode at 1000 Ma, but the black distribution would have age modes at 900 and 1100 Ma. Rather, they should be most alike when the green distribution overlaps with one of the age modes of the black distribution. As an aside, this is the behaviour that cross-correlation of KDEs of these distributions provides (below).

As discussed above, we agree with the reviewer to an extent. In *some* cases (e.g., when discrete sources are well defined and samples are mixtures of them) non-overlapping samples can all be described as equally dissimilar. As a result, in such a scenario, the KS distance may be preferable. However, as previously discussed, this is just one scenario for which comparing detrital age distributions is used. In many other scenarios, described above, absolute distance along the time/x axis ought be considered for comparing distributions.

Figure 3 seems problematic for application of the WS distance to detrital geochronology. For example, Ranealven, Lainioalven, and Bysealven all share peaks at 1800 Ma. The KS distance correctly locates these samples closest to each other. The WS distance in contrast locates the Ljusnan closer to the Ranealven than the Lainioalven even though the major age modes between the Ranealven and Lainioalven (100 Myr offset) are closer than those of the Ranealvan and Ljusnan (200 Myr offset). The WS Distance also locates the Vindelalven and Lainioalven equidistant from the Ranealven. The problem is that distinct source areas may have similar but non-overlapping age modes, and the WS distance is insensitive to these minor differences and therefore unable to discriminate them as distinct sources. A detrital geochronology distribution that has a mode at 1800 Ma and 2800 Ma is more likely to share a source with a sample with a distribution at 1800 Ma (i.e., one overlapping age mode) than one at 1700–1750 Ma (i.e., no overlapping age modes). The authors may be able to address this concern, but it seems to me to be a fatal flaw in this metric.

The reviewer makes two incorrect statements here. First, ‘*distinct source areas may have similar but non-overlapping age modes, and the WS distance is insensitive to these minor differences and therefore unable* *to discriminate them as distinct sources*'. The W_2 distance is perfectly sensitive to minor differences in non-overlapping age modes. In fact the W_2 between two non-overlapping age modes is *exactly* equal to the age offset between them (as is also shown in Figure 2 and described in Equation 3). We argue that this is in fact, as sensitive (a linear, 1 to 1 relationship) as possible. Thus, not only is the W_2 able to discriminate them as distinct sources it *additionally* provides useful geological information on how similar in time those sources are.

Second, the reviewer implies that in the described scenario ('A detrital geochronology distribution that has a mode at 1800 Ma and 2800 Ma is more likely to share a source with a sample with a distribution at 1800 Ma than one at 1700–1750 Ma.'), the Wasserstein distance would be smaller between the 1700 modal sample and the bimodal sample than the 1800 modal sample. This is not correct. We recreate this scenario in Figure 5. The distance matrix between these 3 samples is given in Table 1. As is clear, the distance between samples a & b is smaller than the distance between samples c & b. Consequently, we believe that it is incorrect to suggest that this is a fatal flaw.

However, it is possible to conceive of more extreme scenarios, in which the age difference between the two modes is more extreme (e.g., Cenozoic and Archean). In that case, it is possible that the old age component has an excessive influence on the W_2-distance. This problem can be mitigated by log-transforming the data. Incidentally, that is how Figures 2 and 3 were generated. The scale-dependency of the W_2 can be rightly considered as a weakness of the method relative to the KS-distance, which is scale invariant. The revised manuscript will be more clear about this limitation.

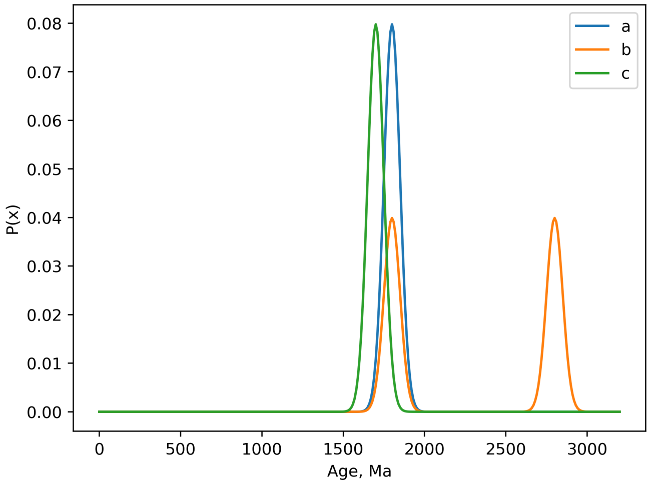


Figure 5: **Three synthetic distributions.** Two unimodal distributions (gaussian with σ=100 Ma centred on 1800 Ma (a) and 1700 Ma (b) respectively. One bimodal distribution with peaks at 1800 and 2800 Ma.

	a	b	c
a	0	680	100
b	680	0	757
c	100	757	0

Table 1: W_2 distances between the distributions displayed in Figure 5

The problems with this metric can be seen when comparing some more complex age distributions. Simple visual inspection indicates that Samples 1, 2, and 7 should be quite similar. This is confirmed by a KS-based MDS (above left), or Kuiper-based MDS (above right). It is also confirmed by a cross-correlation-based MDS (above). However, a WS Distance-based MDS (above) non-intuitively locates samples 1 and 6 quite close, and locates 1, 6, and 7 closer to 5 than to 2.

It is unclear how best to respond to this comment as what the reviewer states as fact ('Simple visual inspection indicates that Samples 1, 2, and 7 should be quite similar') is actually subjective. Moreover, suggesting that there is one 'correct' sample mapping that a metric can fail to reproduce contradicts with the reviewer's own statement that: 'whether [a distance] is more 'sensible' depends on the application'. Additionally, it is not clear to us whether the desired clustering of 1, 2 & 7 is even reproduced on the MDS maps that the reviewer provides. Whilst in the cross-correlation map, the samples 1,2, & 7 *are* clustered, this is not the case in the KS and Kuiper maps. In the KS map, the distance between samples 1 and 2 is over half the x-range. The same is true for samples 1 and 7 on the Kuiper map.

Notwithstanding the above, we also disagree that the W_2 map (Figure 6) is behaving unintuitively. Again we emphasise that there is no one 'correct' mapping of samples. We argue that, however, the mapping produce by W_2 is one particularly intuitive mapping that can be deconstructed sequentially in terms of the means and shapes of the distributions. In the MDS shown on Figure 6 we can see that the samples are primarily distributed on a trend from the upper-left to the lower-right (red arrow), giving the ordering: 2, 7, 1, 6, 5, 4 & 3. We now visualise the KDEs of these samples sequentially in this order on the left-hand (red) column on Figure 6. We observe that this ordering coincides with the ordering from youngest to oldest of the average age of the grains (highlighted with vertical grey dashed lines). We next investigate the ordering of samples on the perpendicular direction from the top-right to the lower-left (blue arrow). Visualising samples in this order (right-hand column, note that these samples have been mean-shifted) we can see that this corresponds to a change in shape from bimodal to unimodal distributions. As a result, we see that, counter to the Reviewer's suggestion, W_2 MDS maps often contain latent directions which are easily interpreted. A similar analysis using the KS distance does not produce readily interpretable latent directions (Figure 7).

A second example highlights the disproportionate impact of misalignment of detrital age modes on the WS Distance. P1 and P3 below share age modes at 1800 Ma. P3 has an additional mode at 500 Ma. P2 shares no age modes with either P1 or P3 and rather has an age mode at 1700 Ma. This is reflected in their KS MDS plot (above left) and an MDS plot based on cross-correlation (above right, note difference in x- and y-axis scale). However, the WS Distance MDS (above) does not reflect the true relationship between age modes, but rather reflects the anomalous area between P1 and P3 (i.e., the horizontal distance between 500 Ma and 1800 Ma in the ECDF plot).

We disagree that misalignment of detrital modes has a **dis**proportionate impact on W_2. As discussed above, the misalignment of modes is in fact (literally) proportional to the W_2 distance. In this particular example we again emphasise that there is no 'correct' mapping for a set of samples. As discussed above, this particular analysis is one where the KS distance may be appropriate, as the absolute ages are not of major importance. However, we reiterate that the mapping produced by W_2 still follows a consistent logic, which considers the absolute ages of the samples. As P3 is the only sample with any grains derived from 'young' sources, it is identified as an outlier. Such a grouping could be intuitive for many (but not all) uses such as identifying approximate ages of deposition/formation. Secondly, we note again that this mapping still distinguishes samples P1 and P2, with a W_2 between them of 100, corresponding to the offset of their peaks. Such an interpretation may not be applicable for *every* study, but it is up to the user to determine which metric is most appropriate.

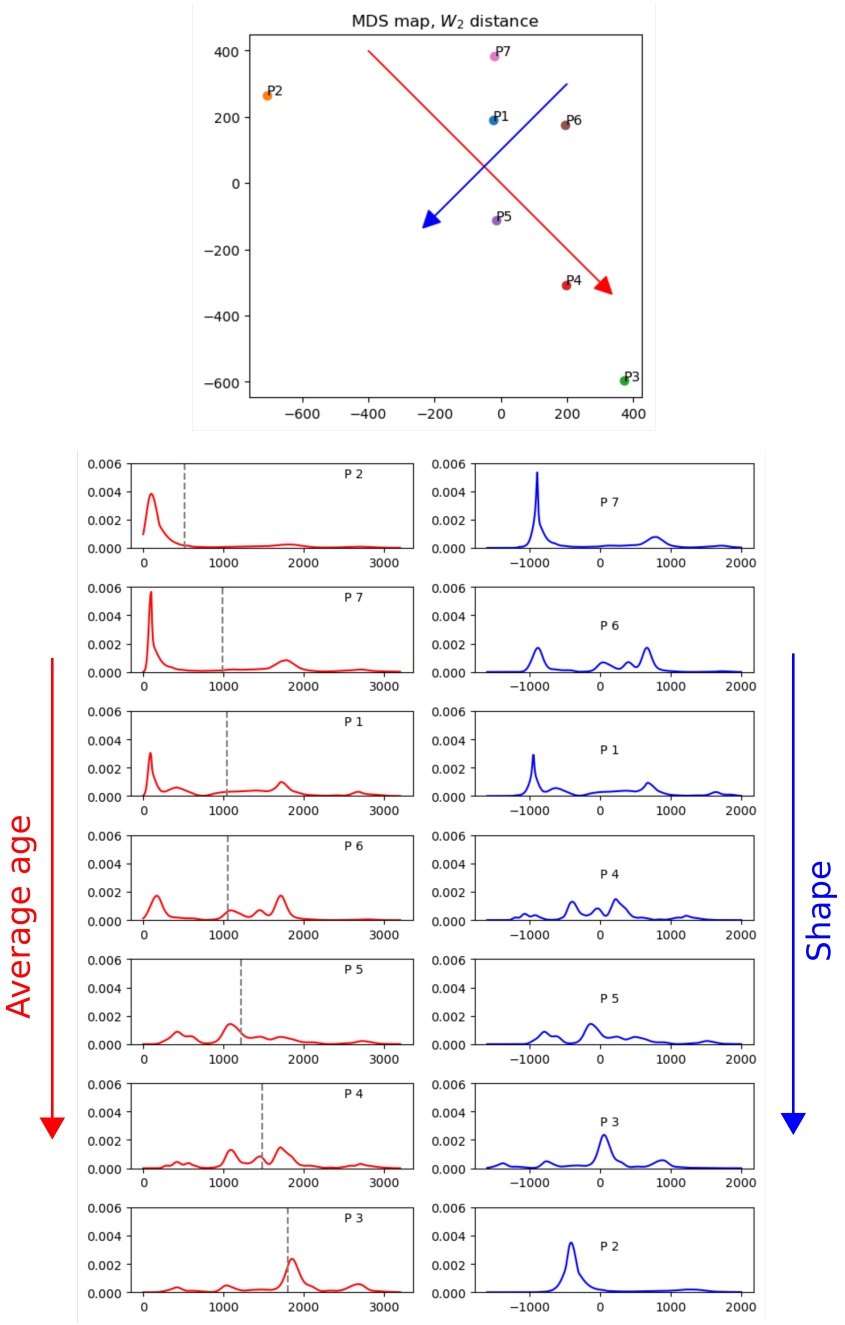


Figure 6: Top figure shows MDS map using W_2 distance as metric of samples provided by reviewer. Arrows indicate the ‘order’ of samples which are visualised in KDEs in the columns below. Left column shows order of samples from top-left to bottom-right. Right column (mean shifted) shows order from top-right to bottom-left.

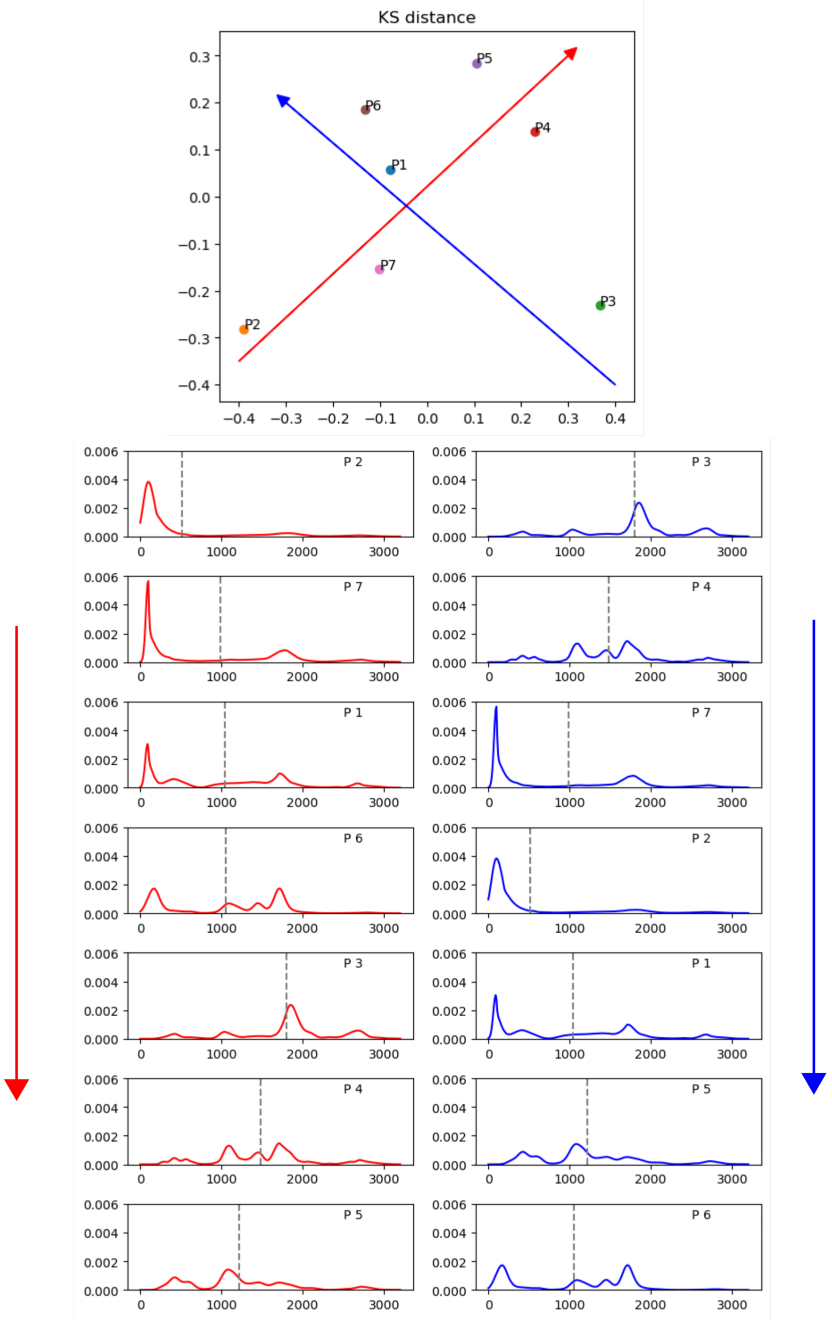


Figure 7: Same as Figure 6 but using KS distance

Detailed comments

The authors assert that cross-correlation is an ad hoc method, yet the Pearson coefficient which is the basis for the cross-correlation coefficient is widely used in seismic analysis, waveform analysis, and image analysis. As such, it is unclear what is meant by "ad hoc" in this context. Although Vermeesch (2018) correctly points out limitations of cross-correlation as applied to probability density plots of extremely precise data, these caveats do not apply to its application to kernel density estimates or kernel functional estimates. Similarly, the charge that Likeness is ad hoc is unfounded. Likeness is an adaptation of the L1 norm applied, usually, to a 1D geochronology distribution. (However, see Sundell et al. (2021) for application of the L1 norm to 2D distributions.) Finally, like the cross-correlation coefficient and Likeness, the Sircombe-Hazelton distance (L2 norm, Sircombe and Hazelton, 2004; Vermeesch, 2018) also requires discretization of continuous functions for its calculation.

The application of cross-correlation to age distributions looks superficially similar to the applications in seismology and image analysis, but is fundamentally different. In seismology, acoustic amplitude is recorded at regular time intervals. In digital image analysis, pixels in a CCD are regularly spaced in a grid. However, the U-Pb ages that constitute a detrital spectrum are not regularly spaced. Unlike the Kolmogorov-Smirnov and Wasserstein distances, which readily accept these irregular data, the cross-correlation and likeness coefficients require that the data are 'shoehorned' into an evenly spaced set of intervals.

The proponents of these methods argue that this can be done using Probability Density Plots (PDPs) or Kernel Density Estimates (KDEs). The shortcomings of PDPs were pointed out by Vermeesch (2012) and are acknowledged by the reviewer. KDEs are also problematic, as they require the selection of a bandwidth. Different bandwidths result in different cross-correlation coefficients and likeness factors. We do not see any justification for this arbitrary intermediate step, which is not required by the KS and W_2 distances.

The cross-correlation and likeness metrics are 'ad hoc' methods for the same reason why the PDP is an ad hoc method. PDPs superficially look like KDEs (Brandon, 1996) but are fundamentally different. Similarly, the cross-correlation coefficient superficially looks like its equivalent in waveform analysis and image recognition, but it is fundamentally different. And the Likeness factor looks a bit like Sircombe et al. (2004)'s L2 norm, but serves a completely different purpose. The S-H metric is used in a niche application, whereby high precision and low precision datasets are combined. In this specific case, some smoothing is justified (and even necessary). In most situation, this is not the case and smoothing should be avoided.

I guess the take away from this is that it may be useful to compare the performance of cross-correlation and Likeness in addition to the KS Distance to newly applied metrics like the WS Distance.

Vermeesch (2018) argued that cross-correlation and likeness should be abandoned, so we will not discuss them further in our paper.

It is ironic that after railing against use of the cross-correlation coefficient, the authors reintroduce it in section 2.2.

$\rho^{\mu\nu}$ is *not* a cross-correlation coefficient. It compares quantiles, not density estimates.

I am confused by the "Unimodal" vs "Multimodal" and "Older" vs "Younger" labels in Figure 3. To my eye the Ljungan is more prominently bimodal than the Byskealven, yet it plots closer to the Unimodal side of the figure than the Byskealven. Similarly, what portion of the distribution is "Younger" or "Older" when comparing multimodal distributions?

'Younger' and 'older' in this Figure refers to the mean age of the samples (c.f., Figure 6). In a revised manuscript we will clarify the description of this figure, in a style similar to those presented above (e.g., Figure 6)

What are the stress values for the MDS plots in figure 3?

We are happy to add the stress values to the MDS plots in a revised manuscript. In nearly all cases that we have seen, the W_2 distance leads to lower stress values than the KS-statistic. However this does not, in itself, mean that W_2 is better than KS. It just means that W_2-distances are more easily captured by two-dimensional MDS configurations than KS-distances are. The reason for this is apparent in the examples of Figures 2 and 3. The W_2-based MDS configurations are almost one-dimensional patterns, whereas the KS-based configurations fill the 2D-space.

References

Amidon, W. H., D. W. Burbank, and G. E. Gehrels (2005). "Construction of detrital mineral populations: insights from mixing of U-Pb zircon ages in Himalayan rivers". *Basin Research* 17, pp. 463–485.

Gehrels, G., D. Giesler, P. Olsen, D. Kent, A. Marsh, W. Parker, C. Rasmussen, R. Mundil, R. Irmis, J. Geissman, et al. (2020). "LA-ICPMS U–Pb geochronology of detrital zircon grains from the Coconino, Moenkopi, and Chinle Formations in the Petrified Forest National Park (Arizona)". *Geochronology.*

Košler, J., J. Sláma, E. Belousova, F. Corfu, G. E. Gehrels, A. Gerdes, M. S. A. Horstwood, K. N. Sircombe, P. J. Sylvester, M. Tiepolo, M. J. Whitehouse, and J. D. Woodhead (2013). "U-Pb Detrital Zircon Analysis – Results of an Inter-laboratory Comparison". *Geostandards and Geoanalytical Research* 37.3, pp. 243–259.

Sircombe, K. N. and M. L. Hazelton (2004). "Comparison of detrital zircon age distributions by kernel functional estimation". *Sedimentary Geology* 171, pp. 91–111.

Vermeesch, P., A. Lipp, D. Hatzenbühler, L. Caracciolo, and D. Chew (2023). "Multidimensional scaling of varietal data in sedimentary provenance analysis". *Journal of Geophysical Research – Earth Surface.*

Vermeesch, P. (2012). "On the visualisation of detrital age distributions". *Chemical Geology* 312-313, pp. 190–194.

– (2018). "Dissimilarity measures in detrital geochronology". *Earth-Science Reviews* 178, pp. 310–321.

Weltje, G. J. and M. A. Prins (2007). "Genetically meaningful decomposition of grain-size distributions". *Sedimentary Geology.* From Particle Size to Sediment Dynamics 202.3, pp. 409–424.

Wobus, C. W., K. V. Hodges, and K. X. Whipple (2003). "Has focused denudation sustained active thrusting at the Himalayan topographic front?" *Geology* 31, pp. 861–864.