The comments of Reviewer 2 are in black, and our responses are in blue.

We really appreciate the positive feedback provided by Reviewer 2, and the reviewer's suggestions allowed us to clarify some important elements regarding what the EMD can and cannot do. We hope Reviewer 2 will find our responses accordingly adequate.

Review of "Refining data-data and data-model biome comparisons using the Earth Movers' Distance (EMD)" by Chevalier et al.

Does the paper address relevant scientific questions within the scope of CP?

Yes, this contribution looks promising as a new method to analyse pollen data for the past and compare them with vegetation model reconstructions. It is thus directly connected with key aspects of paleoclimate reconstruction and, hence, falls fully within the scope of CP.

Does the paper present novel concepts, ideas, tools, or data?

Yes. To my knowledge, the application of the EMD to the analysis of pollen data is completely new. This concept is quite interesting.

Are substantial conclusions reached?

Yes. The authors prove the applicability of the EMD for analysing/comparing pollen data/vegetation model reconstructions. In particular, they show that the EMD has the advantage of conserving more information from the original pollen data. Reconstructions look to be more stable through time compared to classical methods based on the biome with highest affinity score, because these latter methods do not offer a continuous measure of vegetation state (i.e., they are discrete classes).

Are the scientific methods and assumptions valid and clearly outlined?

Yes, the main method used, the EMD, is well explained and references are provided. Other methods are statistical analyses, which relatively well explained.

Are the results sufficient to support the interpretations and conclusions?

Yes.

Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?

Yes, probably. The authors have even developed a R package that certainly helps for such a reproduction of their results, as well for analysis of other pollen data.

Do the authors give proper credit to related work and clearly indicate their own new/original contribution?

Yes.

Does the title clearly reflect the contents of the paper?

Yes.

Does the abstract provide a concise and complete summary?

Yes.

Is the overall presentation well structured and clear?

Yes, it is very well-organised and generally clear.

Is the language fluent and precise?

I am not native English-speaker, but language looks fine to me.

Are mathematical formulae, symbols, abbreviations, and units correctly defined and used?

Yes.

Should any parts of the paper (text, formulae, figures, tables) be clarified, reduced, combined, or eliminated?

The paper has a reasonable length. All parts look necessary. Description of Test 2 (lines 201-216) could be improved. It is difficult to read, especially the method used to establish the second EMD distribution.

We fully agree with Reviewer 2 that the second part of the second test was not as clear as it should have been. We have changed the description to the following, and we believe this makes the test more intelligible.

> *Test 2: Considering the parameters of a study, are the data and the simulation (or modern observations) displaying similar spatial patterns? This second test aims to determine if the mean EMD obtained when comparing a simulated (or observed) vegetation map with a large collection of biome affinity score distributions is smaller than expected when comparing two datasets with different spatial patterns. This test is performed in two steps. First, the data are shuffled (each biome affinity score distribution is randomly assigned to one of the modelled values corresponding to a sample location), and the resulting mean EMD across all locations (i.e. spatial mean) is calculated. This is repeated several times to estimate the distribution of spatial mean EMD values under the assumption that the spatial structure in the data differs from the spatial structure of the simulation (null hypothesis). The 5th percentile of that distribution (any other significance threshold could be used) represents the threshold to reject the null hypothesis (alternative hypothesis: the data and the simulation have similar spatial structures). Then, the uncertainty of the observed EMD value can be estimated by measuring the intra-sample variability. To do so, a second EMD distribution is estimated by bootstrapping, i.e. randomly sampling the same number of biome samples with replacement (some samples are selected many times and others excluded) and calculating the EMD of this bootstrapped dataset with the observed/simulated vegetation map. To determine if the data and the simulation display the same spatial pattern, the 95th percentile of the bootstrapped distribution is compared with the 5th percentile of the distribution of the null hypothesis (one-sided test). If the former is larger than the latter, the null hypothesis is rejected, and the spatial structure of the simulated and reconstructed biomes is considered similar. Efron and Tibshirani (1994) recommend performing at least 200 repetitions to estimate the bootstrapped and null hypothesis distributions. This test is called signif_struct() in paleotools, and its use and interpretation are illustrated in Section 5.*

Are the number and quality of references appropriate?

Yes.

Is the amount and quality of supplementary material appropriate?

Yes.

Comments

In their paper, Chevalier et al. establish a new method based on the EMD to analyse pollen data (change in space and time) or compare them to model vegetation reconstructions. This method is novel, quite interesting and promise to be of broad applicability. The paper is well structured and very well written. It can be published after very minor revision. I have only a few remarks and suggestions that the authors could consider in their revision:

The concept of biome is integrative, i.e., it is used to represent (within classes) the overall vegetation present at a given location. So, only one biome should exist at a given location. Thus, the words "dominant biome" should be avoided, and replaced by something like "biome with the highest affinity score" (with the pollen data).

We originally opted for "dominant biome" because we wanted to avoid repeating "biome with the highest affinity score" multiple times across the paper. However, since both reviewers find this term inappropriate, we now follow their recommendation and opt for "biome with the highest affinity score".

The authors claim that the biomes are discrete quantities, and for that reason, the methods based on the biome with highest affinity score is presented as less robust than the use of the EMD, which is more continuous. However, with the EMD, the authors use the biome concept and their affinity scores. So, some of the "discontinuities" associated to the discrete definition of biomes still remain, especially when mega-biomes are used as done here. Actually, the EMD method developed here could equivalently be applied using plant functional types (PFTs) and PFT scores rather than biomes/biome scores. For instance, Henrot et al. (2017) (Palaeogeogr . Palaeoclim. Palaeoecol. 467, 95-119, 2017) compared model reconstructions with vegetation data at the level of PFT, i.e., using PFT scores. This allows to keep more information from the original pollen data, that are provided at the genus level. In this case, biome maps are just created to illustrate/capture vegetation distribution in a single map.

We believe Reviewer 2 conflated two ideas here: the description of biomes (or mega biomes) as categorical data with the continuity of the metric used to compare them. Usual comparisons assess whether the biomes with the highest affinity scores are the same or not; hence the result is 1 or 0 (*i.e.* discrete). In contrast, comparing distributions of biome scores with the EMD leads to a real number (*i.e.* continuous). The EMD does not aim to fix the problems associated with the categorical data themselves.

We agree that other types of data could be used for the comparison. As discussed in our response to Reviewer 1, we selected biomes because they are simple data that – we believe – are great for illustrating the concepts we are presenting regarding how to use the EMD. Our goal is not to say that biomes are the best way to compare data and models; they probably are not, at least not for all types of applications. While already present in the discussion, we have tried to make this point clearer earlier in the manuscript.

Commented [1]: I think simplicity is not really a good argument....I think a better argument would be the availability of continental/global datasets. We have maps of global biome distributions for the PMIP key time-slices, that's why the palaeo-modelling community is using biomes for model evaluation.... And even if the EMD can be used with other kind of vegetation data, the "original" idea was to find a better method to deal with biomes in model-data comparison studies. I don't think it is the best method to compare PFT distributions. If they are available as quantitative cover fractions, it will work, but not if you just assign the taxa to PFTs. As far as I understood, in the cited study by Henrot et al, they just compare the occurrence of different PFTs, i.e. agreement = (number of sites at which model and data show the same PFT + number of sites at which they both don't show the PFT) / total number of sites.....
That is probably not a good example....

> *Then, using a series of illustrative case studies based on the already-published biomised data and simulations of Cao et al. (2019), we show how the EMD can perform ecologically-informed comparisons in the vegetation space. While more and more quantitative reconstructions of PFT distributions at regional scales have been published in recent years (e.g. the REVEALS-based studies by Githumbi et al. (2022) or Marquer et al. (2017)), we preferred using biomised data because biomes are currently the most-widespread format of publicly available continental-to-global scale syntheses of past vegetation changes (e.g. Binney et al. (2017) and Cao et al. (2019) for Eurasia during the last 40kyr, Prentice et al. (2000) for the Northern Hemisphere and Africa, and Marchant et al. (2009) in South America studying the mid-Holocene and Last Glacial Maximum, or Dowsett et al. (2016) for the mid-Pliocene). They also have a lower dimensionality than PFT data and provide, as such, a simpler context to explore the advantages of the EMD. Despite our focus on biomised data, it is important to stress that other categorical vegetation formats, such as pollen-based quantitative reconstructions as computed by REVEALS (Sugita, 2007) and Earth System Models, PFT affinity scores (e.g. Huntley et al., 2003; Allen et al., 2010; Henrot et al., 2017), or even the comparison of pollen percentages at the taxa level could have been used for our case studies.*

The EMD method allows to measure a (continuous) distance in the multidimensional space phase with the scores of the different biomes. You can thus show for instance (as in Figure 6) how the distance to a present-day biome has varied in the past. However, the biome phase space is multidimensional and the distance does not tell you in which direction you move. Are you moving towards more forests (and if yes towards which type of forests) or towards more grasslands or deserts? This information is quite important to characterize past vegetation. So, the EMD alone is not sufficient to reconstruct precisely past vegetations. It must be combined with a measure of directions in the phase space. In the method presented here, this role is played by the change in the biome with highest affinity score (so the classical method). But could it be improved to achieve a continuous method to evaluate such directions?

It is true that the EMD only measures how dissimilar two samples are and does not say anything about the types of differences. This is similar to the binary evaluations of biomes with the highest affinity score (they only assess if the biomes are the same or not), which is the metric we aim to replace. Both metrics are insufficient to characterise the direction of past vegetation change. Additional analyses, interpretations and/or external knowledge are necessary to characterise the types of changes. We do not foresee any obvious way to include a direction of change into the EMD itself since mathematical metrics by definition do not contain information on the direction of change. Such direction could be determined "a posteriori" by closely looking at the biome affinity score distributions and EMDs, using the biome with the highest affinity score only or the overall amount of affinity scores with forested biomes, for instance.

However, the computation of the EMD could offer some additional insights into the direction of changes through the "optimal flows" that "transport" the affinity scores of sample A to the affinity scores of sample B (see Sect. 2.1.). The optimal flows, which minimise the transport cost, can be written as a transport matrix. As this matrix contains information on how much affinity score needs to minimally be transported to transform the affinity score distribution from sample A to sample B, it does also contain information on the direction of the mismatch between samples. However, quantifying and interpreting this information is not trivial, because (a) the optimal flows are not necessarily unique (in fact, they will rarely be unique in the case of uniform weights since in this case all flows share the same costs), and (b) the form of the transport matrix depends on the weighting scheme. In principle, the directions of the mismatch will be ecologically more meaningful, the more meaningful the weighting scheme is. Incorporating methods to interpret the transport metrics should be explored in future research and the ecological interpretability of transport matrices could be another major advantage of the EMD compared to other metrics that do not contain information on the direction of change. However, it is beyond the scope of this paper where we are not aiming at a metric for the "vegetation phase shift". Here we only introduce a continuous and ecologically-informed metric to replace a binary one. In addition, and as discussed in length in our response to Reviewer 1, we only used biomes as an illustration of the potential of the EMD to compare multivariate datasets. We do not want to "over-specialise" the metric to biomised data, and we prefer to keep it more generic and usable with other datasets, including PFT scores, as suggested above. This element of discussion is nevertheless very interesting, and we have expanded on it in the discussion of the manuscript as well as a discussion on using the optimal flows in the EMD computation to quantify the direction of vegetation changes/mismatches.

"As with most distance metrics, the EMD only measures how dissimilar two samples are and does not provide direct information on the type of (multidimensional) direction of differences. For example, the EMD cannot tell whether sample A is more forested than sample B. It can only quantify how different samples A and B are. This is similar to the binary evaluations of biomes with the highest affinity score. While it is common practice to characterise the direction of change by analysing the properties of the compared datasets separately, the computation of the EMD could offer more direct insights through the "optimal flows" that "transport" the affinity score distribution of sample A to the affinity score distribution of sample B (see Sect. 2.1). The optimal flows, which minimise the transport cost, could be written as a transport matrix. Therefore, this transport matrix would contain information on the (multidimensional) direction of the mismatch between samples. However, quantifying and interpreting these flows is challenging because (a) the optimal flows are not necessarily unique (in fact, they will rarely be unique in the case of uniform weights since, in this case, all flows have the same cost), and (b) the form of the transport matrix depends on the penalty matrix and thus the level of ecological complexity implemented in the penalty matrix. As such, the ecological interpretability of transport matrices could be another

advantage of the EMD compared to other metrics. Therefore, we believe that methods to interpret the "optimal flows" should be explored in future research."

"Finally, it is essential to emphasise that the EMD, as presented in this study, is not limited to vegetation studies. It can be used with any form of discrete palaeodata (i.e. ordinal and categorical) from different disciplines, including but without being limited to, all palaeoecological datasets (e.g. chironomids, foraminifera, rodents, etc.), geochemical datasets (e.g. n-alkane distribution from terrestrial or marine sediments), or archaeological datasets (e.g. lithics and tools from archaeological deposits). More generally, while raw data counts with a different total number of fossils/artefacts cannot be directly compared with the EMD, their percentages always can because they sum to 100. Said differently, any two samples can be compared with the EMD, provided they have the same total mass."

I have not found many typos. Only on line 436, "toa" does not make sense. I guess the authors want to say: "... as a metric to support ..."

Typo corrected.

The authors may want to discuss topics (2) and (3) in the discussion section.

We absolutely agree. See our responses to topics (2) and (3) above and our improved discussion in the manuscript.