https://doi.org/10.5194/egusphere-2022-489
Preprint. Discussion started: 8 July 2022

Refining data-data and data-model biome comparisons using the Earth Movers' Distance (EMD)

Manuel Chevalier[1,2], Anne Dallmeyer[3], Nils Weitzel[4,5], Chenzhi Li[6,7], Jean-Philippe Baudouin[4,5], Ulrike Herzschuh[6,7,8], Xianyong Cao[9], Andreas Hense[1]

1 Institute of Geosciences, Sect. Meteorology, Rheinische Friedrich-Wilhelms-Universität Bonn, Germany
2 Institute of Earth Surface Dynamics, Géopolis, University of Lausanne, Switzerland
3 Max Planck Institute for Meteorology, Bundesstrasse 53, 20146 Hamburg, Germany
4 Institute of Environmental Physics, Heidelberg University, Im Neuenheimer Feld 229, 69120 Heidelberg, Germany
5 Department of Geoscience, University of Tübingen, Schnarrenbergstr. 94-96, 72076 Tübingen, Germany
6 Polar Terrestrial Environmental Systems, Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Telegrafenberg A45, 14473 Potsdam, Germany
7 Institute of Environmental Science and Geography, University of Potsdam, Karl-Liebknecht-Str. 24–25, 14476 Potsdam, Germany
8 Institute of Biochemistry and Biology, University of Potsdam, Karl-Liebknecht-Str. 24–25, 14476 Potsdam, Germany
9 Alpine Paleoecology and Human Adaptation Group (ALPHA), State Key Laboratory of Tibetan Plateau Earth System, Resources and Environment (TPESRE), Institute of Tibetan Plateau Research, Chinese Academy of Sciences, 100101 Beijing, China

Correspondence to: Manuel Chevalier (mchevali@uni-bonn.de)

Abstract. Biome reconstructions are commonly used in data-data and data-model comparison studies to understand past vegetation dynamics. However, most of these assessments are based on the direct comparison of dominant biomes inferred from pollen samples or vegetation simulations. Dominant biomes are deduced from pollen samples using biome affinity scores, which aggregate pollen percentages of taxa assigned to the different biomes. While this approach generates good results over a large range of temporal and spatial scales, reducing pollen assemblages to a single dominant biome can substantially simplify the vegetation signal preserved in pollen samples and even bias conclusions when, for instance, minimal changes in pollen percentages can change the inferred dominant biome. To resolve these issues, we propose to use the Earth Movers' distance (EMD) as a new metric to compare distributions of biome scores. The EMD has two main advantages: 1) the distributions of biome scores do not need to be reduced to their dominant biome, and the full breadth of the data is taken into account, and 2) different weights can be given to different types of disagreements to account for the ecological distance (*e.g.* reconstructing a temperate forest instead of a boreal forest is ecologically less wrong than reconstructing the temperate forest instead of a desert). We also introduce EMD-based statistical tests that determine if the similarity of two samples is significantly better than a random association. This paper illustrates the use of the EMD across a series of palaeoecological data-data and data-model case studies based on published data and simulations. These applications highlight the diverse types of analysis where the EMD adds value compared to analyses of the dominant biomes only. The EMD and the statistical tests are included in the paleotools R package (https://github.com/mchevalier2/paleotools).

1 Introduction

Fossil pollen records are commonly used to evaluate Earth System Model (ESM) palaeosimulations in the climate space (*i.e.* pollen data are converted into climate parameters using transfer functions, Birks et al. (2010) and Chevalier et al. (2020)) and the vegetation space (*i.e.* vegetation features are simulated using vegetation models, Prentice et al. (1998), Tian et al. (2018) and Wohlfahrt et al. (2008)). Both evaluations are necessary to explore the strengths and weaknesses of fossil pollen data, climate and vegetation models, and the modern observations used to link vegetation with climate. In both cases, a transformation of the pollen data is necessary. To perform data-model comparisons in the vegetation space, the pollen data are commonly translated into biomes, which correspond to broad vegetation classification units characteristic of regional- to global-scale features (e.g. Cao et al., 2019; Dallmeyer et al., 2017; Sato et al., 2021). 'Biomised' pollen data have several advantages over the raw pollen percentages: (*i*) they reduce the dimensionality of the vegetation space (*i.e.* reducing the few hundred pollen taxa usually observed across a continent to about 10-15 biomes), (*ii*) they summarise the main traits characterising the studied vegetation compositions (*i.e.* enabling a convergence of the traits and a spatial homogenisation of the data), and (*iii*) they improve the comparability of data between different data sources (*i.e.* pollen data, modern observations, and simulations).

Biomes are usually estimated from pollen data usually using a two-step algorithm that converts pollen percentages into biome scores (Prentice et al., 1996, 2001; Prentice and Webb III, 1998). Schematically, it involves designing two matrices, one to assign pollen taxa to one or more plant functional types (PFTs) and a second one to assign the PFTs to one or more biomes. Then, the pollen percentages are processed and distributed among the different PFTs and biomes following the rules defined by the two matrices to produce an array of biome scores. Additionally, minimum presence thresholds (*e.g.* 0.5 or 1%) and percentage scaling factors (*e.g. Larix* percentages are commonly multiplied by 15 to account for its low pollen production/preservation rates) can also be included to refine the estimates. From this, the biome with the highest score is often used to label the type of vegetation that dominates the immediate landscape at and around the sampling location (hereafter the "dominant biome"). Most data-data and data-model studies are based on the comparison of the dominant biome estimates, where the "agreeing" and "disagreeing" pairings (*i.e.* binary assessments of the compared estimates) are counted and summarised in contingency tables to determine the global accuracy between the compared datasets (e.g. Binney et al., 2017; Cao et al., 2019; Prentice et al., 1998).

However, this simplification can overly homogenise the data. When the highest biome score is much larger than the second-highest score, reducing the distribution of scores to a dominant biome is a reasonable simplification of the data. But when the difference between the highest and second-highest score is small (*e.g.* near ecotones), simplifying multidimensional biome data to one univariate dominant biome estimate often leads to ignoring a significant part of the information conveyed by the pollen data. As many distinct biome score distributions can lead to the same dominant biome, a large fraction of the fine-scale details of the vegetation structure gets lost. This is illustrated with the two samples represented in Fig. 1A. The 'Evergreen taiga' biome has the highest score in both samples, rendering them indistinguishable when reduced to their dominant biome. However, inspecting the distribution of biome scores informs us about significant differences, with the bottom sample most likely representing a well-forested environment and the top sample being closer to a mosaic of forest patches connected by open landscapes (characterised by more abundant Tundra and Grassland pollen taxa). In contrast, the two biome score distributions in Fig. 1B have distinct dominant biomes and are thus classified as different vegetation groups (one is classified as Evergreen Taiga and the other as Cool/Cold forest). Yet, their distributions only differ by minute changes between the different forested biomes, and the two distributions likely represent (very) similar environments.

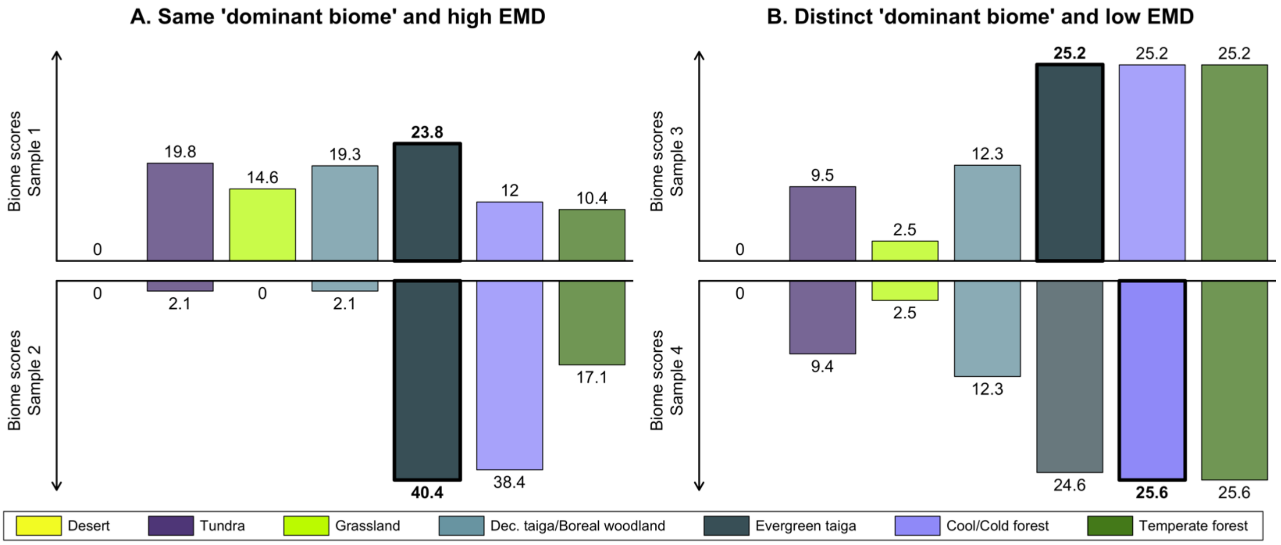


Figure 1. Illustration of limitation of using dominant biome estimates to compare samples. (A) The two samples have the same dominant biome (Evergreen taiga), but the distributions of biome scores have notable differences. (B) The samples have different dominant biomes (Evergreen taiga and Cool/Cold forest), but they only differ by minor differences. The data are reproduced from the study of Cao et al. (2019).

The approach of summarising an array of biome scores by its dominant biome is thus insufficiently sensitive, and it can lead to a loss of accuracy when comparing datasets (contrasting samples are assigned to the same category, while similar samples are assigned to different categories). Employing continuous metrics that consider the entire distribution of biome scores (as opposed to binary assessments) can thus refine the quality of such data-data and data-model biome comparisons. Many distances commonly used to compare pollen data – such as the Manhattan/Euclidean distance (*i.e.* calculating the absolute/squared differences between the scores of the same biomes) or the squared-chord distance (e.g. Overpeck et al., 1985) – could be used to measure the dissimilarity of two biome score distributions. However, these distances give the same importance to all the differences without accounting that all biome changes are not ecologically equivalent. For example, replacing a cool/cold forest with a temperate forest represents a smaller ecological/climatic shift than replacing a forest with a desert, even if the absolute differences in biome scores are the same.

To account for these two limitations, we propose the Earth Movers' Distance (EMD) metric as a new way to compare pollen-derived biome score distributions with one another and with vegetation simulations. The advantage of this distance metric compared to standard binary assessments is dual: 1) the EMD is continuous such that vegetation differences can be quantified in finer detail, and 2) the inclusion of ecologically-informed weights adds a level of refinement that takes into account different types of mismatches between samples. This paper first introduces the EMD and describes the many properties that make it ideal to capitalise on the rich information contained in biome scores distributions. Using a series of case studies based on already-published data and simulations, we then illustrate how the EMD can be used to perform ecologically-informed comparisons in the vegetation space. We finally discuss different research directions where the EMD could be used.

2 The Earth Movers' Distance (EMD)

2.1 Concept and formalisation

The EMD is a distance metric based on the underlying idea of measuring the minimal amount of work necessary to transform one entity into another. The general concept of the EMD algorithm can be most simply illustrated with the following everyday-life transportation problem: "What is the most cost-efficient way of transporting a fixed merchandise stock from *W* warehouses to *R* retailing shops?" (Levina and Bickel, 2001; Rubner et al., 2000). The problem can be reframed as: "How can the distribution of merchandise in the warehouses be transformed into the desired distribution of merchandise in the shops?". To solve this problem, we call $d_{i,j}$ the distance between warehouse W_i and retailing shop R_j, ω_i the stock of merchandise at W_i and ρ_j the stock of merchandise needed at R_j. The EMD algorithm searches for the optimal combination of flows $f_{i,j}$ of merchandise (*i.e.* the amounts) to be moved between the warehouses and shops in a way that minimises the total cost C (*i.e.* the sum of how much is moved between locations multiplied by their distance).

$$C = \min_{i,j}\left(\sum_{i=1}^{W}\sum_{j=1}^{R} f_{i,j}.d_{i,j}\right) \tag{1}$$

with the constraints:

1. $f_{i,j} \geq 0, 1 \leq i \leq W, 1 \leq j \leq R$, *i.e.* the flow of merchandise between locations W_i and R_j is positive or null. The merchandise is moved from the warehouses to the shops, and not the opposite.
2. $\sum_{j=1}^{W} f_{i,j} \leq \omega_i$, *i.e.* the total amount of merchandise leaving warehouse W_i to all the shops does not exceed its stock.
3. $\sum_{i=1}^{R} f_{i,j} \leq \rho_j$, *i.e.* the total amount of merchandise arriving at retailing shop R_j from all the warehouses does not exceed its need.
4. $\sum_{i=1}^{R}\sum_{j=1}^{W} f_{i,j} = \sum_{i=1}^{R} \omega_i = \sum_{j=1}^{W} \rho_j$, *i.e.* the total amount of merchandise transported between warehouses and shops is equal to the initial amount of merchandise in the warehouses and the final amount of merchandise in the retail shops. The overall amount of merchandise is conserved.

Once the optimal flows are estimated, the EMD is calculated as follows (the minimal cost normalised by the sum of all flows):

$$EMD = \frac{\sum_{i=1}^{W}\sum_{j=1}^{R} f_{i,j}.\, d_{i,j}}{\sum_{i=1}^{W}\sum_{j=1}^{R} f_{i,j}} \qquad (2)$$

Based on this formal definition, the transportation problem can be reframed in a broader context to become equivalent to finding the 'shortest' way of transforming one probability mass/density distribution into another (Levina and Bickel, 2001). With its flexibility, the EMD has been employed in a wide range of contexts, including, for instance, image retrieval algorithms (Rubner et al., 2000), the comparison of inorganic compositions (Hargreaves et al., 2020) or biomarker expression in cells (Orlova et al., 2016). To our knowledge, it has never been used to compare palaeoecological datasets.

2.2 The EMD applied to biomised data

2.2.1 Terminology

In this study, we propose to use the EMD to compare biome score distributions from vegetation simulations and reconstructions. The transportation of merchandise becomes a transport of biome scores between samples (*i.e.* a transformation of the vegetation composition). The concept of physical distance between entities (*e.g.* warehouses and shops) can be reframed as the ecological 'cost' of replacing a type of biome with another one. To ensure compatibility with constraint 4 of the previous section (the amount of 'merchandise' or 'biome scores' is the same between the two entities compared), the biome scores must be normalised (here rescaled to sum to 1). This step is essential because the most commonly used biomisation techniques do not ensure biome scores sum to a common target (*e.g.* 1 or 100). While alternative modelling techniques could be used to directly derive proper biome probabilities from pollen assemblages (e.g. Litt et al., 2012), we prefer focusing on the most commonly available data type (biome scores derived from biomisation schemes) to illustrate the advantages of the EMD for palaeoecological data-data and model-data comparison.

2.2.2 Definition of the weighting scheme (ecological distance)

We also use two different types of weighting schemes (the cost of replacing a biome with another one, the "d_{ij}" from Eq. 1) to illustrate how ecological knowledge can be introduced in such studies (Fig. 2). We use a "uniform" scheme where all the biome changes are given the same weight (EMD_{uni}) and an "ecologically informed" scheme, where differences are given a weight of 1, 2 or 3 depending on the nature of the ecological difference (EMD_{w}). We define the latter with the following rules:

1. Biomes are categorised in one of the three following macro-environments: 'forested environments', 'herbaceous/open landscapes', or 'deserts'.
2. Differences within one macro-environment are given a weight of 1 (*e.g.* replacing a forest with another forest)

3. Differences between forested environments and open landscapes and between open landscapes and deserts are given a weight of 2.
4. Differences between forested environments and deserts are given a weight of 3.

This simple weighting scheme demonstrates how an ecologically-informed weighting strategy can help refine interpretations compared to the uniform scheme. Different research questions or settings can lead to alternative weighting schemes (e.g. Sato et al., 2021).

2.2.3 Rescaling the EMD between 0 and 1

The EMD calculated with normalised biome scores and a weighting scheme is thus a dissimilarity metric that varies between 0 (the two distributions are identical, and nothing needs to be moved) and the highest cost of that weighting scheme (here defined as $\max_{i,j} d_{i,j}$). The highest distance can be reached when all the scores are transferred between the most different macro-environments. In our weighting scheme, this would correspond to the transformation of a pure forest composition into a desert, or *vice-versa*. To improve the comparability between studies and/or weighting schemes, the EMD can be normalised by the highest cost:

$$EMD_n = \frac{EMD}{\max_{i,j} d_{i,j}} \tag{3}$$

Unlike other metrics for which expert-elicited quality thresholds have been proposed (*e.g.* kappa statistics, Altman (1999) or Landis and Koch (1977)), no expert-based quality assessment exists for the EMD_n. In fact, defining such a quality scale could be counter-productive, as many study-dependent factors influence the range of values the EMD_n will take in a given study. These include: 1) the number of biomes to compare (more biomes usually lead to higher distances), 2) the definition of the weighting schemes (the EMD_n is inversely proportional to the highest cost), or 3) the data structure of the entities being compared (compare the EMD_n ranges in the data-data (same structure) and data-model (unary data compared to multidimensional data) comparison applications below for a concrete example). Comparing EMD values between studies should, therefore, always be done carefully.

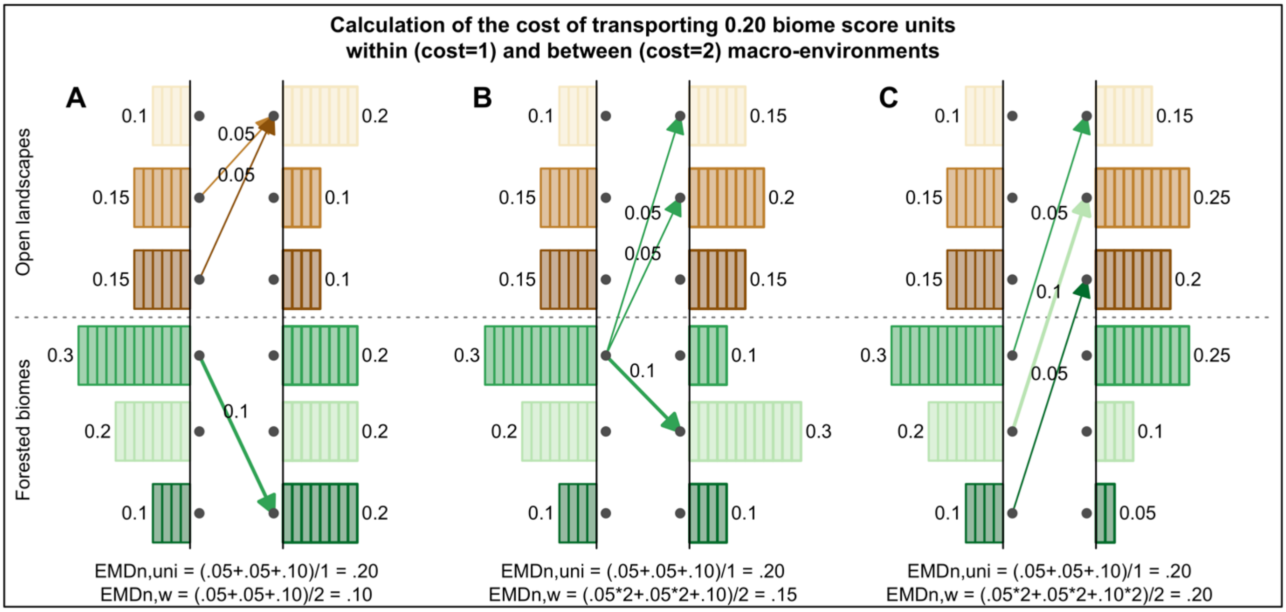

Calculation of the cost of transporting 0.20 biome score units
within (cost=1) and between (cost=2) macro-environments
A
B
C
Open landscapes
Forested biomes
0.1
0.15
0.15
0.3
0.2
0.1
0.2
0.1
0.1
0.2
0.2
0.2
0.05
0.05
0.1
0.15
0.2
0.15
0.1
0.3
0.1
0.05
0.05
0.1
0.15
0.25
0.2
0.25
0.1
0.05
0.05
0.1
0.05
EMDn,uni = (.05+.05+.10)/1 = .20
EMDn,w = (.05+.05+.10)/2 = .10
EMDn,uni = (.05+.05+.10)/1 = .20
EMDn,w = (.05*2+.05*2+.10)/2 = .15
EMDn,uni = (.05+.05+.10)/1 = .20
EMDn,w = (.05*2+.05*2+.10*2)/2 = .20

Figure 2. Calculation of the $\mathrm{EMD_n}$ for three different scenarios in which 0.20 units of normalised biome scores are transported: (A1) all changes are within the same macro-environments, (B) changes are both within and between macro-environments and (C) all changes are between macro-environments. The use of ecologically-informed costs (here only two options: 1 or 2) leads to three distinct values for $\mathrm{EMD_{n,w}}$ (0.1, 0.15, 0.2), while the uniform approach considers that all three changes are equivalent ($\mathrm{EMD_{n,uni}}$ = 0.2).

2.3 Implementation of the EMD in the 'paleotools' R package

To facilitate access to the EMD, we developed an R package called paleotools using R v4.0.2 (R Core Team, 2022) and the *devtools* package (Wickham et al., 2020). To calculate the EMD, *paleotools* includes a wrapper function of the *emd()* function from the *emdist* R package (Urbanek and Rubner, 2022). In addition, we also developed two statistical tests, *signif_threshold()* and *signif_struct()*, to overcome the absence of 'quality' thresholds.

Test 1: Considering the parameters of a study, can two samples be considered similar? This test is inspired by the Monte Carlo simulation designed by Sawada et al. (2004) to identify analogue samples from a large and heterogeneous collection of pollen samples. The underlying idea is to determine a distance threshold that is unlikely to have occurred by chance (Simpson, 2007). To do so, a large number of pairs of biome score distributions are randomly drawn from a data collection and their EMD calculated. This results in a distribution of EMD values derived from randomised comparison of biome score distributions. The EMD value corresponding to a certain percentile of that distribution (*e.g.* the 5th percentile) can be used as an empirical estimate of a similarity threshold. EMD values below/above that threshold correspond to comparisons of similar/different samples. Importantly, this test does not determine if two samples represent the same biome. Two samples can be statistically different and be characteristic of the same biome. Still, their statistical difference suggests they likely occupy a different position in that biome's vegetation and/or climate spaces. For instance, vegetation samples taken from the cold and warm ends of the temperature range experienced of the temperate forest biome are likely to be statistically different while being representative of temperate forests. This test can only be used if hundreds of biome score distributions representative of various environments are available to estimate the randomisation distribution. In addition, this type of threshold is only valid for a given study area and/or research question. The test is called *signif_threshold()* in *paleotools* and its use and interpretation are illustrated in Section 4.

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 collection of biome score distributions is smaller than expected when comparing two datasets exhibiting different spatial patterns. This test is performed in two steps. First, the data are shuffled (each biome score distribution is randomly assigned to one of the modelled values corresponding to a sample location), and the resulting EMD is calculated. This is repeated a few hundred times to estimate the distribution of EMD under the assumption that there is no spatial structure in the data (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 reconstructed and simulated data have similar spatial structures). Then, the uncertainty of the 'true' EMD value can be estimated by measuring the intra-sample variability. To do so, a second EMD distribution is estimated by 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 simulated vegetation map. Finally, the 95th percentile of the bootstrapped distribution is compared with the 5th percentile of the distribution of the null hypothesis. 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.

3 Data

3.1 Pollen and biome reconstructions

To illustrate the use and the strength of the EMD for palaeoecological studies, we use the pollen-based biome reconstructions presented in Cao et al. (2019). The dataset covers the entire Northern Hemisphere extratropics. Here, we restricted it to the Euro-Mediterranean Basin, where the quality and quantity of pollen records are ideal for testing the EMD in various conditions (Fig. 3). The pollen data were extracted from the European Pollen Database in June 2017, and a total of 1347 records fall within our study area. The biomisation strategy employed by Cao et al. (2019) follows the biomisation tables presented in Binney et al. (2009) and Bigelow et al. (2003), and the algorithm of Prentice et al. (1996). 13 distinct biomes can be theoretically reconstructed across the study area (Table 1).

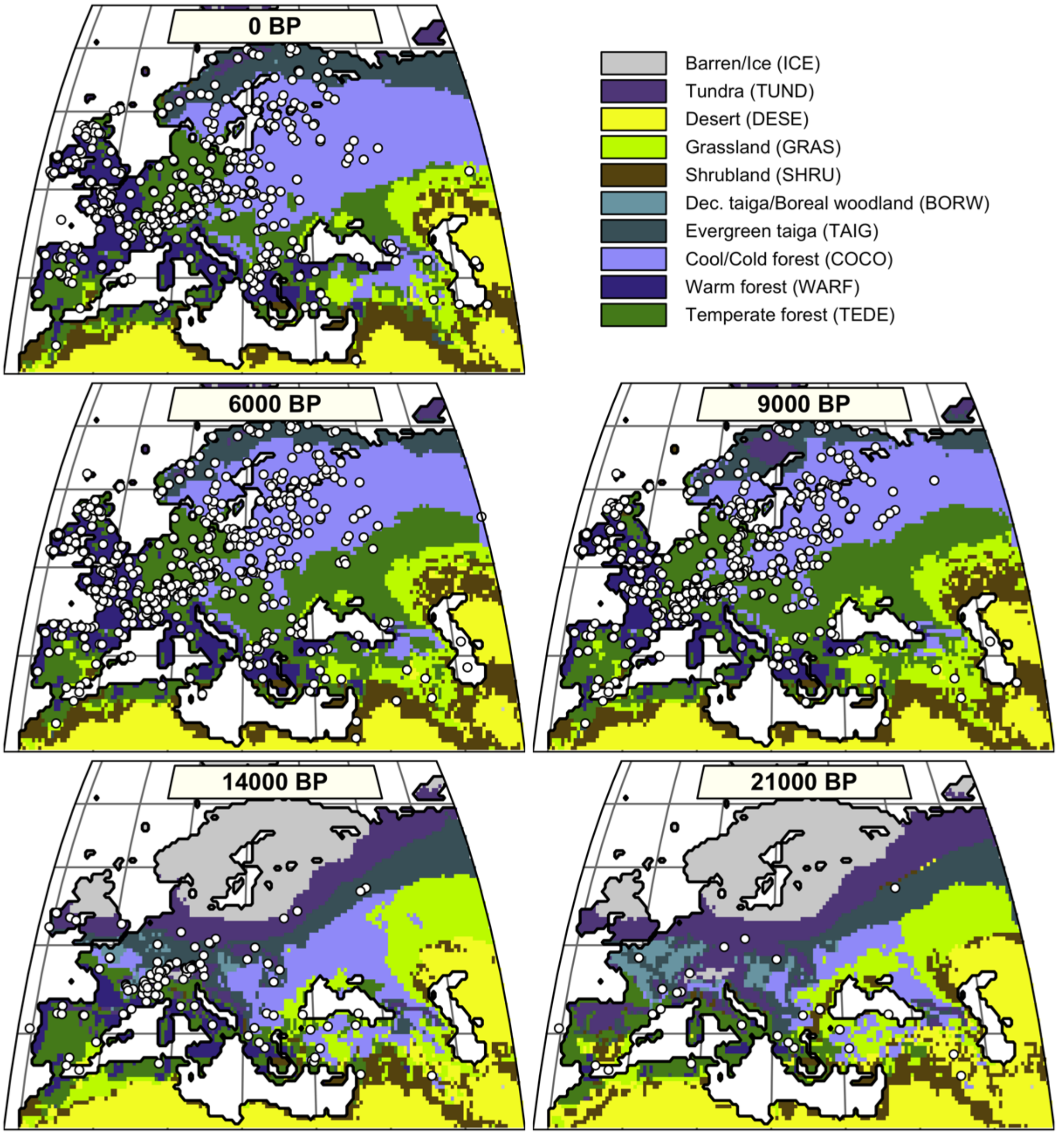


Figure 3. Distribution of the simulated mega-biomes at 0, 6, 9, 14 and 21 ka. Each pollen-based biome estimate is represented as a white dot and corresponds to a distribution of biome scores, as illustrated in Fig. 1.

3.2 Climate and vegetation simulations

We use the vegetation simulations presented in Cao et al. (2019). These simulations were derived from the biome model BIOME4 (Kaplan et al., 2003) in the version adapted by Dallmeyer et al. (2017). BIOME4 calculates the equilibrium biome distribution for 28 potential biomes using a prescribed climate and taking biogeographical and biogeochemical processes into account (Kaplan, 2001; Kaplan et al., 2003). Of these 28 biomes, 21 were observed in our study area for at least one time period (Table 1). Input variables are climatological monthly mean temperature, cloud cover and precipitation, the climatological mean absolute minimum temperature of the year, atmospheric CO_2 concentration, and physical properties of the soil such as water-holding capacity and percolation rates. The results are outputted as the 'dominant biome' only.

Macro-environments	Mega-biome (Dallmeyer et al., 2017)	Biomes from BIOME4	Euro-Mediterranean biomes from pollen (Cao et al., 2019)
Forests	Temperate forest / Woodland (TEDE)	Temperate deciduous forest Temperate conifer forest Temperate sclerophyll woodland	Temperate deciduous forest
	Warm forest (WARF)	Warm mixed forest	Warm-temperate evergreen broadleaved and mixed forest
	Cold/Cool forest (COCO)	Cool mixed forest Cool conifer forest Cold mixed forests	Cool evergreen needle-leaved forest Cool-temperate evergreen needle-leaved forest Cool mixed forest
	Evergreen Taiga (TAIG)	Evergreen taiga / montane forest	Cold evergreen needle-leaved forest
	Deciduous Taiga / Boreal woodland (BORW)	Deciduous taiga / Montane forest Open conifer woodland Boreal parkland	Cold deciduous forest
Herbaceous / Open landscapes	Shrubland (SHRU)	Temperate xerophytic shrubland Tropical xerophytic shrubland	Temperate xerophytic shrubland
	Grassland (GRAS)	Tropical grassland Temperate grassland	Temperate grassland
	Tundra (TUND)	Steppe tundra Shrub tundra Dwarf shrub tundra Prostrate shrub tundra Cushion forb lichen moss tundra	Graminoid and forb tundra Low and high shrub Erect dwarf-shrub tundra Prostrate dwarf-shrub tundra Cushion-forb tundra
Deserts	Desert (DESE)	Desert	Desert

Table 1. Biome to mega-biome to macro-environments assignments following Dallmeyer et al. (2017) for the simulated biomes and Cao et al. (2019) for the reconstructed biomes.

In the simulations used here, BIOME4 has been forced by climate simulations conducted in the coupled general circulation model Community Earth System Models (COSMOS) in the spatial resolution T31 (~4°x4° on a gaussian grid). COSMOS has been developed by the Max Planck Institute for Meteorology. It consists of the general circulation model for the atmosphere ECHAM5 (Roeckner et al., 2003) coupled with the land-surface model JSBACH (Brovkin et al., 2009) and the ocean model MPIOM (Marsland et al., 2003). An anomaly approach has been used to prepare the climate input data for the biome model and reduce systematic model biases, for instance, due to the coarse spatial resolution of the model in which the orography is strongly smoothed. For this purpose, the difference between the climate simulated for a particular time slice and the pre-industrial reference climate has been calculated, bilinearly interpolated to a regular 0.5°x0.5°grid and added to observations (here: CRU-TS3.1 data, Harris et al., 2014). Five timeslices are available, i.e. 21ka and 14ka (Zhang et al., 2013), 9ka and 6ka (Wei and Lohmann, 2012), and 0ka (Wei et al., 2012). Further details and global boundary conditions of the climate simulations are described in Dallmeyer et al. (2017) and Tian et al. (2018).

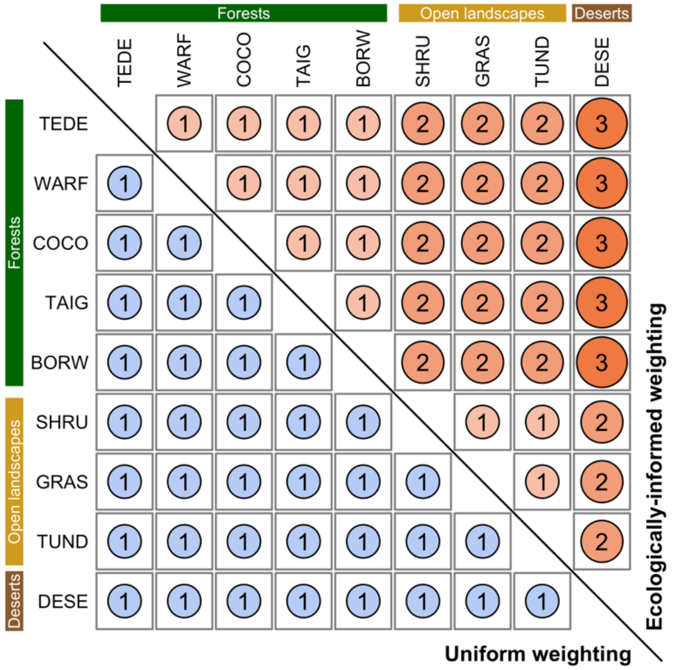
3.3 Harmonisation of the biome reconstructions and simulations

Since the definition of biomes was slightly different in the two datasets, the biome reconstructions and simulations were harmonised with the 'mega-biome' scheme of Dallmeyer et al. (2017) to enable direct comparison. This scheme is a classification tool composed of 12 levels, which allows the grouping of biomes into higher-order vegetation classes. Nine mega-biomes were observed across the study area (see Fig. 3 and Table 1). Because the model results were only available as the dominant biome of the grid cell, the harmonisation at the mega-biome level was straightforward. Each grid cell was assigned to the mega-biome corresponding to its biome (Table 1). Harmonising the pollen data was more challenging because the data were available as arrays of biome scores. Since many taxa are part of multiple biome, adding the scores of the different biomes belonging to the same mega-biomes would lead to overestimating the mega-biome scores (*i.e.* the weight of some taxa would be accounted for several times). Re-running the biomisation algorithm would have thus been necessary to obtain exact mega-biome scores (replacing the 'plant functional type to biome' table with a 'plant functional type to mega-biome' table in the biomisation algorithm). However, not all the required data were available. For simplicity, we assumed the mega-biome scores could be defined by the highest score of all their composing biomes (see Table 1 for the detailed biome composition of each mega-biome). This solution is imperfect and underestimates the actual scores. Still, we believe this simplification is sufficient for the purpose of this study, which is to illustrate how the EMD can be used in data-data and data-model comparison studies and not generate new data. Finally, the mega-biomes were also grouped into three macro-environments ('forested environments', 'herbaceous/open landscapes', or 'deserts'; Table 1) to define the weights used to calculate the $EMD_{n,w}$ (Fig. 4).

4 Data-data comparison: EMD vs dominant mega-biome estimates

4.1 Discrimination between mega-biomes

We perform a series of data-data comparison case studies to evaluate the performance of the $EMD_{n,w}$ compared to analyses based on dominant mega-biome reconstructions only. First, we analyse the $EMD_{n,w}$ values calculated between 2500 randomly selected mega-biome samples, irrespective of their ages (past and modern samples were pooled together). The resulting 3,123,750 pairwise comparisons are grouped according to the agreeing or disagreeing status of their 'dominant mega-biomes'. If two samples have the same dominant mega-biome X, such as in Fig. 1A, the pair is labelled as 'Mega-biome X'. If they differ (Mega-biome X and Mega-biome Y, such as in Fig. 1B), the pair is labelled as 'Mega-biome X with other biomes' and as 'Mega-biome Y with other biomes'. The two resulting $EMD_{n,w}$ collections for Mega-biome X (*i.e.* 'Mega-biome X' and 'Mega-biome X with other mega-biomes') are interpreted as the intra-mega-biome and inter-mega-biome $EMD_{n,w}$ variability distribution of Mega-biome X. The results for the five most abundant mega-biomes across the study area and the $EMD_{n,w}$ are summarised in Fig. 5 (see Appendix 1 for the same figure with the $EMD_{n,uni}$).



	TEDE	WARF	COCO	TAIG	BORW	SHRU	GRAS	TUND	DESE
TEDE		1	1	1	1	2	2	2	3
WARF	1		1	1	1	2	2	2	3
COCO	1	1		1	1	2	2	2	3
TAIG	1	1	1		1	2	2	2	3
BORW	1	1	1	1		2	2	2	3
SHRU	1	1	1	1	1		1	1	2
GRAS	1	1	1	1	1	1		1	2
TUND	1	1	1	1	1	1	1		2
DESE	1	1	1	1	1	1	1	1	

Figure 4. The two penalty matrices used in this study. The lower (upper) triangle in blue (orange) represents the uniform (ecologically-informed) weighting, respectively. In both cases, the diagonal of the matrix contains 0s. The biome acronyms are defined in Table 1.

These data are also used to explore how the proposed statistical similarity test performs (Test 1, Section 2.3). We test the ‘stability’ of the significance threshold as a function of the number of $EMD_{n,w}$ values available. At the scale of Europe, the $EMD_{n,w}$ threshold for a significant similarity at 5% is ~0.031 (Fig. 5B). While this value can be correctly estimated on average from a few samples, its variability can, however, be high when only a limited number of samples are selected. Undersampling the data (or small-sized datasets) can thus lead to an increased risk of mistakenly rejecting or accepting the null hypothesis (H0: The two samples are dissimilar). Here, our results suggest that considering about 10,000 $EMD_{n,w}$ values, which corresponds to all the pairwise comparisons between about 140-150 independent samples, is necessary to obtain stable thresholds. The results of this similarity test are always relative to the size of the study area, wherein small-scale studies will have smaller EMD thresholds because the samples will be more similar on average. Each threshold is thus study-specific and should not be employed in a different context.

For the five biomes selected here, the mean $EMD_{n,w}$ of the intra-mega-biome distributions are smaller than the mean $EMD_{n,w}$ of the corresponding inter-mega-biome distributions and large intra-mega-biome $EMD_{n,w}$ values are not observed for most mega-biomes, except perhaps for tundra (TUND). This result is coherent and expected, as the dominant mega-biome estimate is a summary measure that extracts the dominant signal from the data. However, comparisons of the intra-mega-biome $EMD_{n,w}$ distribution with the inter-mega-biome $EMD_{n,w}$ distribution highlight a substantial overlap. Many pairs of

https://doi.org/10.5194/egusphere-2022-489
Preprint. Discussion started: 8 July 2022

samples from the inter-biome distributions have very low EMD$_{n,w}$, suggesting strong similarities in their relative mega-biome compositions despite having different dominant mega-biome estimates (similar to the example in Fig. 1B). In the 'extreme' case of temperate deciduous forests (TEDE), about one-third of the pairs from TEDE's inter-mega-biome EMD$_{n,w}$ distribution has a smaller EMD$_{n,w}$ than pairs from TEDE's intra-mega-biome EMD$_{n,w}$ distribution.

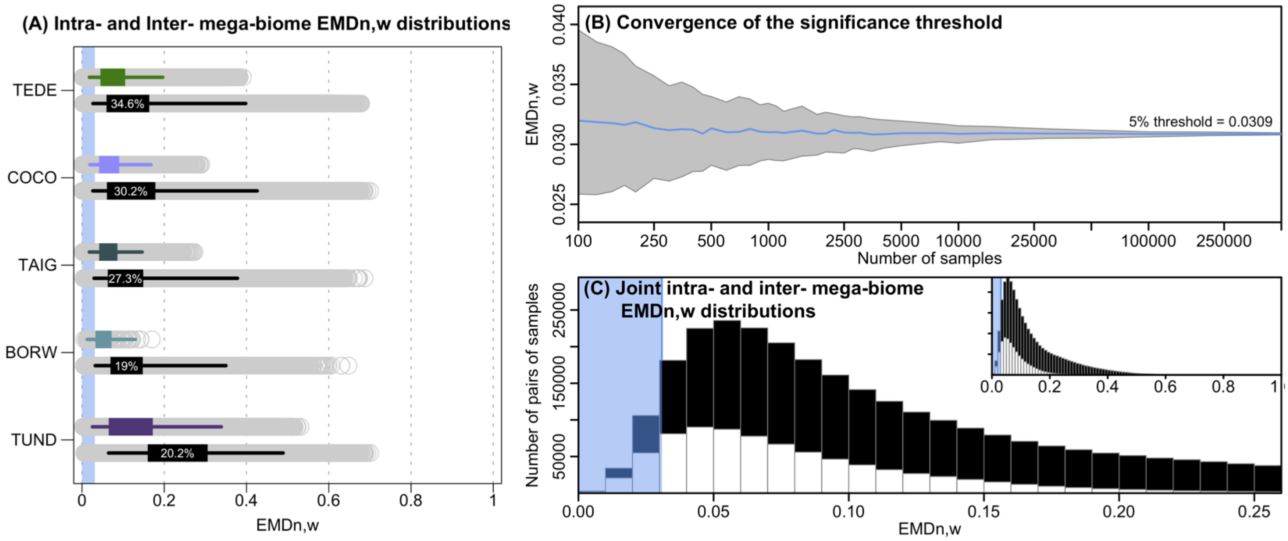


Figure 5. (A) Distribution of intra- (coloured) and inter- (black) mega-biome EMD$_{n,w}$ distributions. For all mega-biomes, the top/coloured boxplot represents the distribution of the pairwise distances of all the samples with the same dominant mega-biome, and the bottom/black boxplot represents the EMD$_{n,w}$ distributions of these samples with different dominant mega-biomes. The box of the boxplot represents the 25-75% interval (interquartile range), and the whiskers represent the 2.5-97.5% interval. The percentages indicate the proportion of samples where the EMD$_{n,w}$ of the inter-mega-biome distribution is lower than the intra-mega-biome distribution (estimated from 10,000 bootstrapped pairs of samples drawn from the intra- and inter-mega-biome EMD$_{n,w}$ distributions). The higher the percentage, the higher the overlap of the two distributions is. (B) Evolution of the estimation of the EMD$_{n,w}$ threshold as a function of the number of samples selected. Note the log scale of the x-axis. (C) Distribution on the intra- (white) and inter- (black) mega-biome EMD$_{n,w}$ (all mega-biomes pooled together). The blue band in (A) and (C) represents the range of EMD$_{n,w}$ values that characterise statistically similar samples, based on our first statistical test (estimated in (B)).

This large overlap between the inter- and intra-mega-biome distributions can be further illustrated with the statistical test we designed to determine if two samples can be considered similar (Test 1). Of all the pairwise comparisons that are deemed significant (all mega-biomes included), only a bit more than half (54%) corresponds to comparisons of samples with identical dominant mega-biome estimates (Fig. 5C). Therefore, these results demonstrate that while the dominant mega-biome approach produces good results on average, fine-scale details of the vegetation structure are missed in some comparisons when samples are solely labelled by their dominant mega-biome estimates.

4.2 Characterising mega-biome changes in space and time

With the second data-data comparison study, we show how the more gradual response of the EMD$_{n,w}$ to changes in mega-biome score distributions can be used to refine vegetation change interpretations through time and space (Fig. 6). When

https://doi.org/10.5194/egusphere-2022-489
Preprint. Discussion started: 8 July 2022

mega-biome reconstructions are only represented by their dominant mega-biome estimate, oscillations between different mega-biomes can be frequent. However, this apparent variability can be an artefact caused by the simplification of the multidimensional data to univariate estimates. This is illustrated by the mega-biome reconstruction from the pollen record Lago Piccolo di Avigliana (Finsinger et al., 2011; Finsinger and Tinner, 2006; Fig. 6A). For this record, we calculate 1) the $EMD_{n,w}$ of all the samples with the top sample to measure the broad trends of mega-biome divergence over time relative to modern-day and 2) the sample-to-sample $EMD_{n,w}$ to measure the high-frequency vegetation variability in the data. We also used the similarity threshold defined in the previous section from the 2500 $EMD_{n,w}$ values.

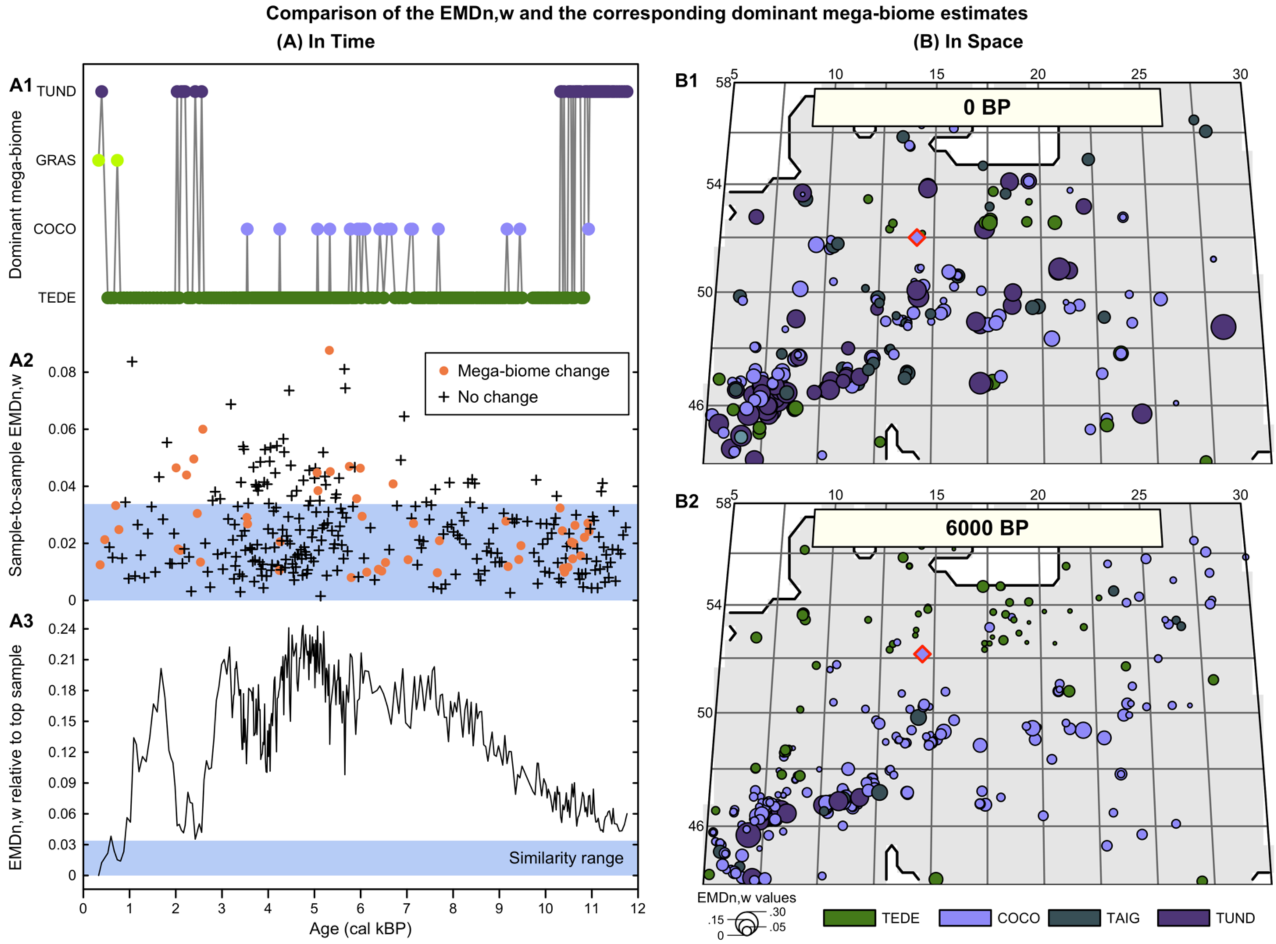


Figure 6. Comparison of the $EMD_{n,w}$ and the corresponding dominant mega-biome estimates (A) in time and (B) in space. (A1) 'dominant mega-biome' reconstruction for a pollen record from northern Italy (Cao et al., 2019; Finsinger et al., 2011; Finsinger and Tinner, 2006). (A2) $EMD_{n,w}$ calculated between neighbouring pairs of samples, highlighting that vegetation changes that trigger a change in dominant mega-biome are not different from the changes that do not. (A3) $EMD_{n,w}$ of the biome scores compared to the top sample, highlighting significant vegetation changes across time. The significance threshold at 5% (blue dashed lines) was derived from the random sampling of 2500 pairs of Holocene samples across Europe. (B) Mapping of the EMD_w of all the regional samples compared to the mega-biome reconstruction at the location indicated with a red diamond at 0 BP (B1) and 6000 BP (B2).

Large vegetation changes are evident in the record, with all the samples older than 1000 BP being dissimilar to the top sample. The representation of the data by the dominant mega-biome estimates suggests high vegetation instability over time (52 changes for 321 samples). However, if these mega-biome shifts are analysed with the EMD, most sample-to-sample changes are associated with statistically similar samples. In particular, the mean differences between samples that trigger a change in the dominant mega-biome estimate ($\overline{EMD_{n,w}}$ = 0.026, sigma = 0.016, n = 51) are not statistically different than the mean changes between samples that do not ($\overline{EMD_{n,w}}$ = 0.024, sigma = 0.015, n = 269; t-test p-value = 0.39). As opposed to the representation based on dominant mega-biomes that suggests a constant vegetation variability across the record, the $EMD_{n,w}$ trends suggest that vegetation changes were rather slow before ~7,000 BP and since ~2000 BP, and more intense in between. This example illustrates how the type of representation chosen for the data can influence interpretations. In this case, the oscillations visible in the dominant biome estimates are only a visual artefact resulting from the simplification of the data to single estimates instead of looking at the entire distribution of mega-biome scores.

Similar smooth transitions can be observed for the variability across space, where the spatial granularity of the data is much lower than what is suggested by dominant mega-biome estimates (Fig. 6B). Many neighbouring samples characterised by distinct dominant biomes are, in fact, rather similar according to the $EMD_{n,w}$ (*e.g.* the small dots of different colours near the target sample in Fig. 6B). In general, the size of the dots (*i.e.* the EMD_{w}) increases with distance to the target sample or with higher elevation, such as in the Alpine and Carpathian regions. The mean $EMD_{n,w}$ values at 6,000 BP is much lower than modern values, and their distribution in space is much more regular. Determining the reasons why these differences exist is beyond the scope of this paper, but it could be related to the influence of humans on modern environments (*e.g.* deforestation and opening of the landscapes) or different climate conditions.

5 Data-Model comparison: Evaluation of vegetation simulations

Data-model biome comparisons are commonly based on comparisons of dominant biome estimates from models, pollen assemblages or modern observations. In such cases, the number of agreeing and is used to measure the accuracy, and the results are reported in contingency tables. These tables are ultimately analysed using different indices (e.g. the Kappa statistics, Cohen (1960)). As shown in the previous section, this type of data simplification is suboptimal, because the information of the non-dominant biomes cannot be taken into account. If the matching biome is the second-best biome, the strength of the mismatch is not the same as if it were the biome with the lowest score. To illustrate the advantage of the EMD and of the ecologically-informed weights in such a data-model comparison context, we reproduced the data-model comparison of Cao et al. (2019), who used the Kappa statistic (among other metrics) to evaluate the similarity of the patterns displayed by the data and models. The results are summarised in Table 2 and represented in Fig. 7.

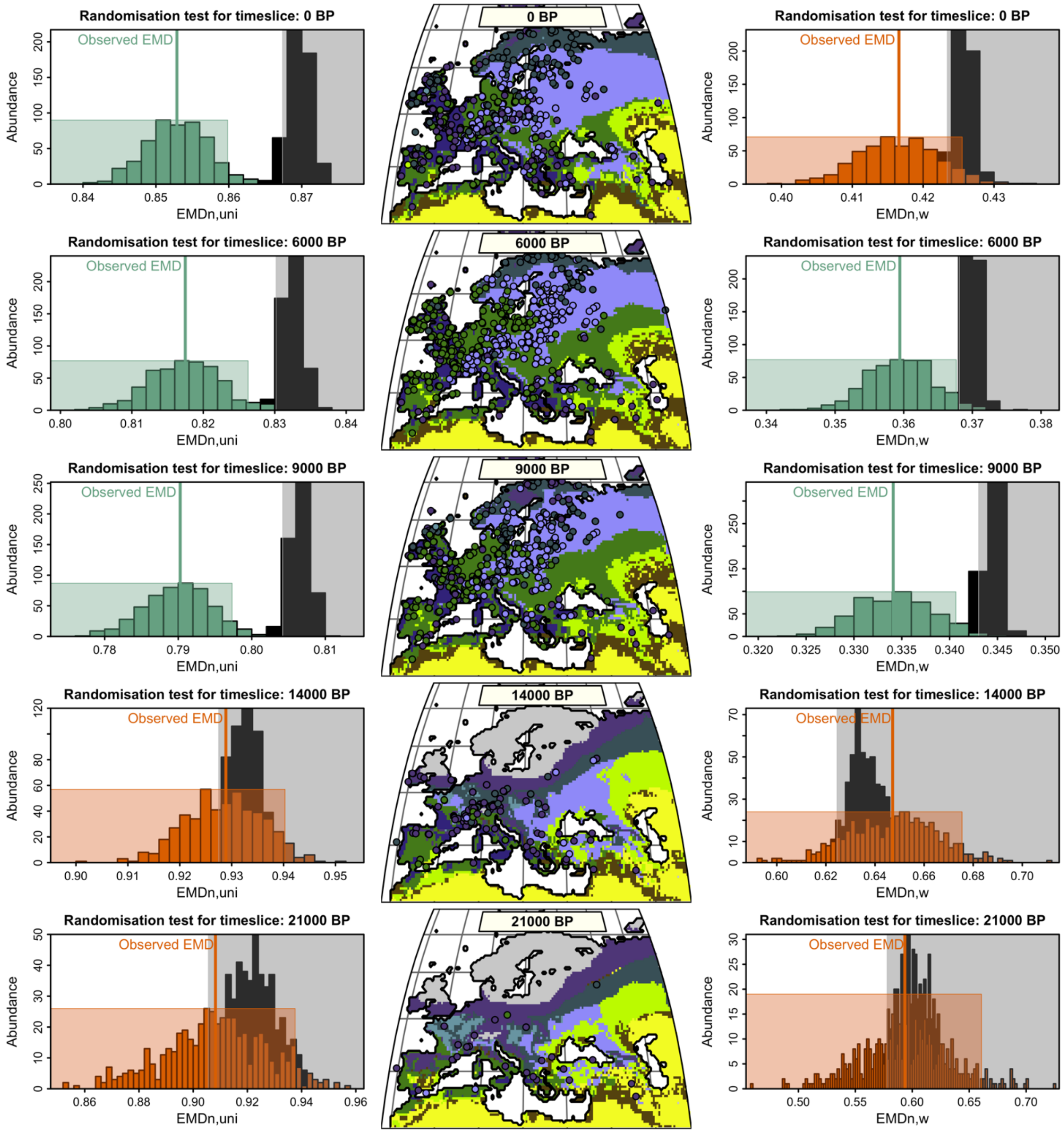


Figure 7. Data-model comparisons for the five simulated timeslices using either EMD$_{w,uni}$ (left) or EMD$_{n,w}$ (left). (centre) The dominant mega-biome estimates derived from the pollen data are plotted over the simulated mega-biomes. (left/right) Statistical test evaluating the degree of similarity between the reconstructed and simulated mega-biomes. The black histogram represents the distribution of EMDs under the null hypothesis (the spatial distributions of the two datasets are different). The coloured histogram represents the uncertainty distribution of the observed EMD. The null hypothesis is rejected when the black and coloured rectangles do not overlap (the rectangles are defined based on a 5% significance threshold and 500 repetitions). Green/Orange means that the null hypothesis is rejected/accepted and that the two datasets have a similar/different spatial structure.

One limitation of this study is that we compare pollen-derived multidimensional biome score distributions with model-derived unary biome distributions (all scores are concentrated on one single biome). This fundamental difference in the data structure of the two entities being compared means that reaching an EMD of 0 is highly unlikely because obtaining such a concentrated distribution of biome scores from pollen data is nearly impossible. In general, dominant biomes have an affinity score in the 0.2-0.3 range, even if this can be quite variable (Fig. 1). This implies that even if the dominant biome of a pollen sample matches the simulated biome of the corresponding grid cell, the $EMD_{n,uni}$ will have, in general, a value of about 0.7-0.8 because all the non-dominant biome scores have to be “moved” to the dominant biome category. The same principle applies to the $EMD_{n,w}$, but calculations of a ‘best case scenario’ range are less direct due to the penalty matrix. This explains why the absolute EMD values of this data-model comparison are much higher than the EMD values calculated in the previous data-data comparison based on the comparisons of datasets with similar structures. Still, this technical limitation does not impede using the EMD for data-model comparisons.

Time Period	Accuracy (correct/total)	Kappa statistic	$EMD_{n,uni}$	$EMD_{n,w}$
0 BP	77 / 368	0.04	0.853	0.417 (NS)
6000 BP	106 / 422	0.06	0.817	0.359
9000 BP	85 / 292	0.08	0.790	0.334
14000 BP	21 / 100	0.00	0.929 (NS)	0.647 (NS)
21000 BP	4 / 28	0.00	0.908 (NS)	0.593 (NS)

Table 2: Summary statistics of the data-model comparison. The accuracy and Kappa statistics are reported by Cao et al. (2019), and the EMD values result from our statistical test. Both are reported as the mean of all values across the study area. The Kappa statistics and the EMD are similarity and dissimilarity measures, respectively. A value of 1 (0) is the best (worst) score for the Kappa statistics and the worst (best) score for the EMD. Non-significant tests are labelled with (NS).

Among all timeslices, models and data are most consistent at 9 ka according to the three evaluation indices (Table 2). The overall ranking of the five data-model comparisons based on the $EMD_{n,w}$ and $EMD_{n,uni}$ is also consistent with the Kappa statistics and accuracy of Cao et al. (2019) (Table 2). We used the statistical test defined in Section 2.3 (Test 2) with both the $EMD_{n,w}$ and $EMD_{n,uni}$ to determine if the spatial patterns of the simulated and reconstructed biomes for the five timeslices (0, 6, 9, 14 and 21 ka) were similar. The results indicate that the data-model comparisons for timeslices at 6K and 9K are significant in both cases, while those for timeslices at 14 ka and 21 ka are not. Interestingly, the simulation at 0 ka is significant with the $EMD_{n,uni}$ and not significant with the $EMD_{n,w}$. These contrasting results can be explained by the type of ecological differences between the data and model at 0 ka. Most of the mismatches correspond to pollen samples with tundra as the dominant biome (TUND, Fig. 7) and where forested environments (either TEDE or COCO) are simulated by the model. By definition of the penalty matrix and the ‘ecologically-informed’ ranking of errors (Fig. 4), the replacement of forests with more open landscapes is strongly penalised in $EMD_{n,w}$. This difference tips the test result from significant without the weights (the two datasets have a similar spatial structure if all differences are considered equally) to non-significant when the weights are included (their spatial structure is different if we consider that replacing a forest with more open landscapes is a large ecological change).

6 Perspectives

The examples presented in the previous sections illustrate how using a continuous metric, as opposed to a binary assessment of similarity, can help refine interpretations of data-data comparisons and facilitate a better understanding of vegetation dynamics through time (Fig. 6A) and space (Fig. 6B). The EMD proved to be a powerful tool for performing statistically robust data-model comparisons, despite the unary distributions of the simulated data (Fig. 7). These examples demonstrated that 1) while interpretations based on dominant mega-biome estimates tend to be correct on average, they miss fine-scale details of the data, and 2) the simplification can even add noise to reconstructions (*e.g.* temporal oscillation of the dominant mega-biomes, Fig. 6), even if the underlying data changes smoothly. These examples represent only a fraction of the type of applications where the use of the EMD could be recommended. For example, the statistical test could also compare if the data-model agreement at a given timeslice is statistically more robust than at another one.

While the EMD was used with already biomised pollen data, the EMD could also be used to optimise the biomisation schemes themselves. Creating biomisation schemes often requires tuning several parameters in parallel while evaluating the results with modern vegetation maps. Due to the sensitive nature of the dominant mega-biome estimates, small changes can easily change one dominant mega-biome into another (Fig. 1B), which can strongly impact Kappa statistics and other binary indices. The EMD offers a smoother alternative and could be used to optimise biomisation schemes and/or compare existing schemes more cohesively. Introducing different weighting schemes for the EMD could also become a variable to integrate into such biomisation studies. The one used in this study is based on simple ecological considerations, but the data-model comparison example nevertheless highlighted its advantages (Fig. 7). Yet, more complex, data-informed distance matrices could be designed by, for instance, calculating (some form of) inter-mega-biome distance in the climate and/or vegetation spaces, integrating plant traits ecological distance (e.g. Sato et al., 2022) or modelling the probability of mistaking one biome for another using independent calibration data. Developing these alternative matrices is, however, complex as their stability in time and space should be assessed before being used.

Despite its simplicity, our categorical penalty matrix already adds a level of refinement that is absent from most other biome comparison techniques. Similar matrices could also be defined at different taxonomical resolutions of the pollen data (*e.g.* taxa or PFT levels). For example, many pollen-based reconstruction techniques are based on the comparison of pollen samples (Chevalier et al., 2020). The EMD could be introduced as metric toa to support the definition of analogues, and thus climate reconstructions could benefit from the refinements brought by the penalty matrix. However, 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, raw data counts with a different total number of fossils/artefacts cannot be directly compared with the EMD, but their percentages always can because they sum to 100. Said differently, any two samples can be compared with the EMD provided that they have the same total mass.

7 Conclusion

Comparisons of discrete palaeoclimatic vegetation data are often based on the co-evaluation of their best estimates. While based on sound principles, this approach has limitations, particularly regarding the impossibility of accounting for the multidimensionality of the data. This paper proposes to replace the binary metrics commonly used to perform data-data or model-data biome comparisons with the Earth Movers' Distance (EMD). The EMD is a valuable alternative because it considers the complete distributions of biome scores and can assign specific weights to different types of errors. Since this metric integrates more information, EMD-based studies allow for more refined interpretations. The versatility of the EMD enables performing various types of data-data and data-model comparisons with biome data (as presented here) and with other palaeoenvironmental, palaeoclimatic or archaeological proxies. To complement the use of the EMD, we propose a statistical framework to test the robustness of comparisons (*i.e.* testing if the different elements being compared share similar features). The EMD and the EMD-related significance tests have been integrated into an R package, 'paleotools', to facilitate access and reuse.

Code and data availability. The 'paleotools' R package is available from https://github.com/mchevalier2/paleotools, and the scripts and data used in this study are available from zenodo at XXX [Available for publication].

Financial support. This research has been supported by the PalMod Initiative, funded by the German Federal Ministry of Education and Research (BMBF), Research for Sustainability initiative (FONA, http://www.fona.de, last access: 6th June 2022). In particular, MC is funded by PalMod subproject 01LP1926D, AD by subproject 01LP1920A, JPB by subproject 01LP1926C and UH by subproject 01LP1510C. This research has been supported by the European Research Council (ERC Glacial Legacy 772852 to UH). CL holds a scholarship from the Chinese Scholarship Council (grant no. 201908130165). NW acknowledges funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project no. 395588486. XC acknowledges funding from the Basic Science Center for Tibetan Plateau Earth System (BSCTPES, NSFC project No. 41988101) and the Sino-German Mobility programme (grant no. M-0359).

Author contribution. MC designed the study, performed the experiments and wrote the original draft. All authors contributed ideas from the earliest stages and commented on the different iterations of the manuscript.

https://doi.org/10.5194/egusphere-2022-489
Preprint. Discussion started: 8 July 2022

Acknowledgements. MC thanks the Max Planck Institute für Meteorology (MPI-M) in Hamburg and the Climate Vegetation Dynamics group, in particular.

8 Appendix 1

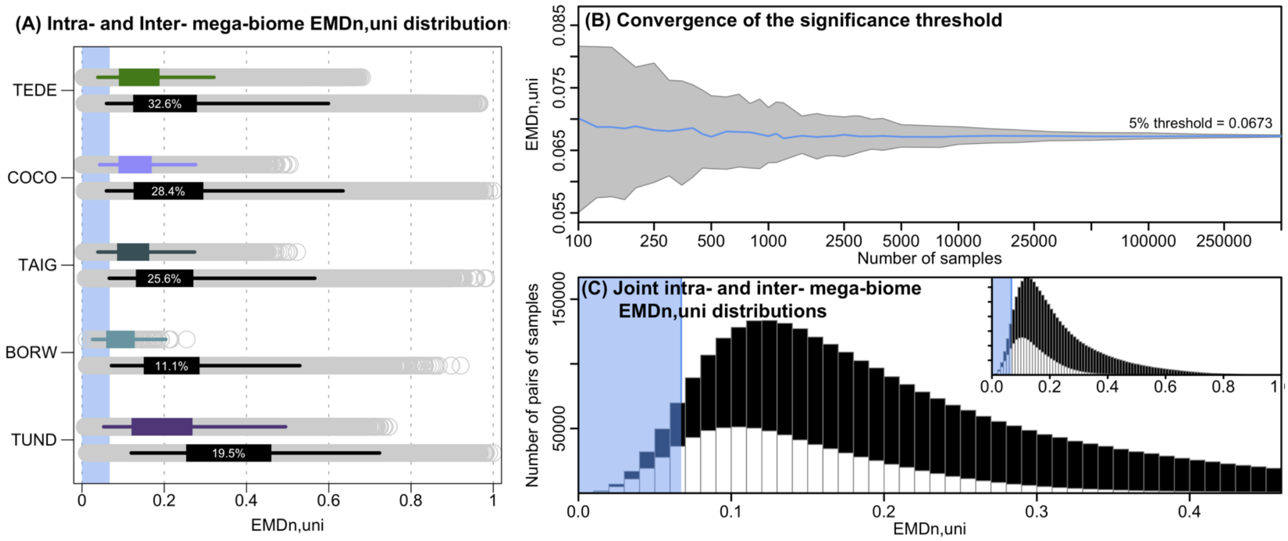


Figure A1. Distribution of intra- (coloured) and inter- (black) mega-biome $EMD_{n,\,uni}$ distributions across the study area. In all five panels, the top/coloured boxplot represents the distribution of the pairwise distances of all the samples with the same dominant mega-biome, and the bottom/black boxplot represents the $EMD_{n,\,uni}$ distributions of these samples with different dominant mega-biomes. The box of the boxplot represents the 25-75% interval (interquartile range), and the whiskers represent the 2.5-97.5% interval. The percentages indicate the proportion of samples where the $EMD_{n,\,uni}$ of the inter-biome distribution is lower (estimated from 10,000 bootstrapped pairs of samples drawn from the intra- and inter-mega-biome $EMD_{n,\,uni}$ distributions). The higher the percentage, the higher the overlap of the two distributions.

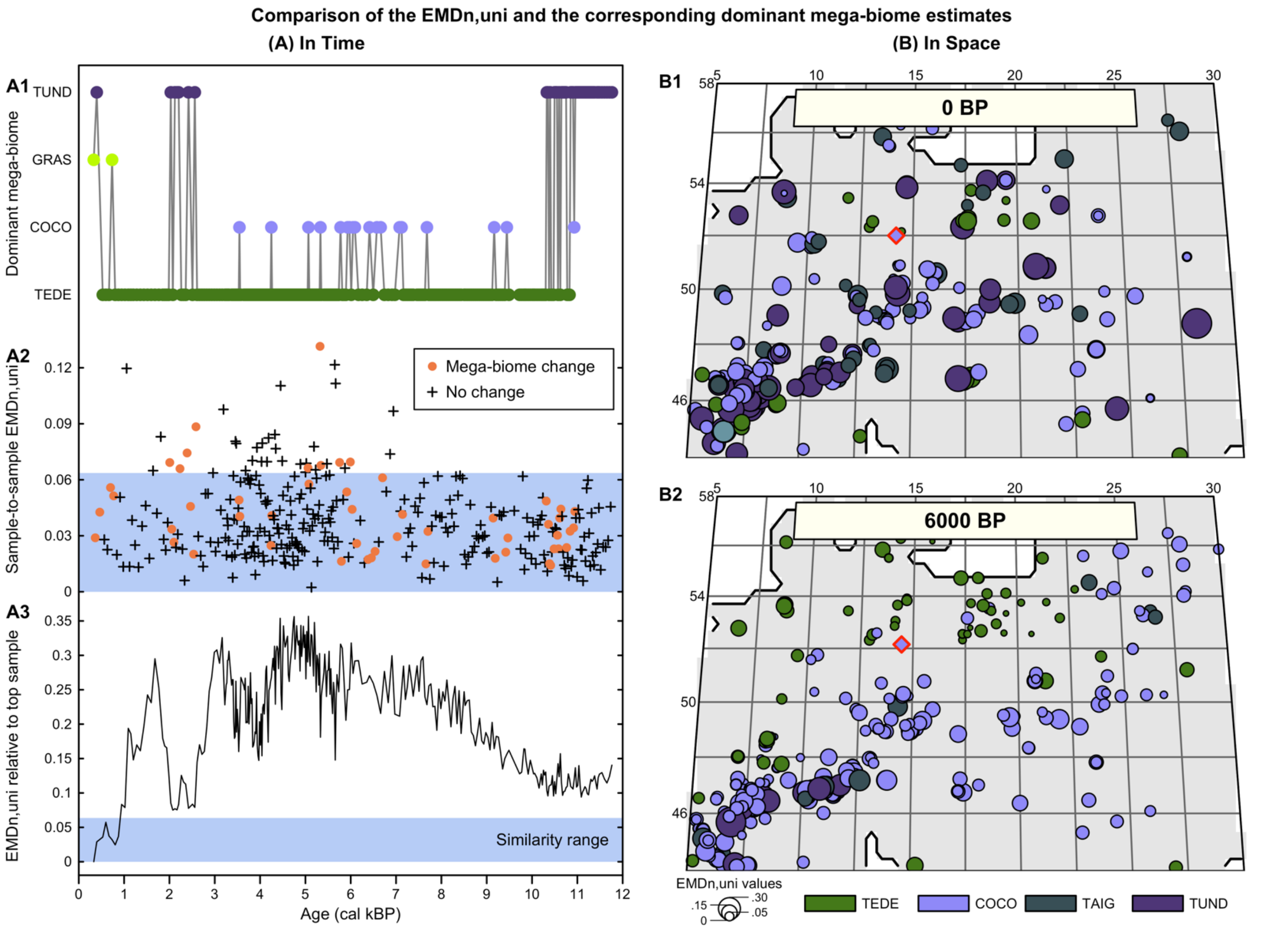


Figure A2. Comparison of the EMD$_{n,uni}$ and the corresponding dominant mega-biome estimates (A) in time and (B) in space. (A1) 'dominant mega-biome' reconstruction for a pollen record from northern Italy (Cao et al., 2019; Finsinger et al., 2011; Finsinger and Tinner, 2006). (A2) EMD$_{n,uni}$ calculated between neighbouring pairs of samples, highlighting that vegetation changes that trigger a change in dominant mega-biome are not different from the changes that do not. (A3) EMD$_{n,uni}$ of the biome scores compared to the top sample, highlighting significant vegetation changes across time. The significance threshold at 5% (blue dashed lines) was derived from the random sampling of 2500 pairs of Holocene samples across Europe. (B) Mapping of the EMD$_{n,uni}$ of all the regional samples compared to the mega-biome reconstruction at the location indicated with a red diamond at 0 BP (B1) and 6000 BP (B2).

References

Bigelow, N. H., Brubaker, L. B., Edwards, M. E., Harrison, S. P., Prentice, I. C., Anderson, P. M., Andreev, A. A., Bartlein, P. J., Christensen, T. R., Cramer, W., Kaplan, J. O., Lozhkin, A. V., Matveyeva, N. V., Murray, D. F., McGuire, A. D., Razzhivin, V. Y., Ritchie, J. C., Smith, B., Walker, D. A., Gajewski, K., Wolf, V., Holmqvist, B. H., Igarashi, Y., Kremenetskii, K., Paus, A., Pisaric, M. F. J. and Volkova, V. S.: Climate change and Arctic ecosystems: 1. Vegetation changes north of 55°N between the last glacial maximum, mid-Holocene, and present, Journal of Geophysical Research:

Atmospheres, 108(19), doi:10.1029/2002jd002558, 2003.

Binney, H. A., Willis, K. J., Edwards, M. E., Bhagwat, S. A., Anderson, P. M., Andreev, A. A., Blaauw, M., Damblon, F., Haesaerts, P., Kienast, F., Kremenetski, K. V., Krivonogov, S. K., Lozhkin, A. V., MacDonald, G. M., Novenko, E. Y., Oksanen, P., Sapelko, T. V., Väliranta, M. and Vazhenina, L.: The distribution of late-Quaternary woody taxa in northern Eurasia: evidence from a new macrofossil database, Quaternary Science Reviews, 28(23–24), 2445–2464, doi:10.1016/j.quascirev.2009.04.016, 2009.

Binney, H. A., Edwards, M. E., Macias-Fauria, M., Lozhkin, A., Anderson, P., Kaplan, J. O., Andreev, A. A., Bezrukova, E., Blyakharchuk, T. A., Jankovska, V., Khazina, I., Krivonogov, S., Kremenetski, K. V., Nield, J., Novenko, E. Y., Ryabogina, N., Solovieva, N., Willis, K. J., Zernitskaya, V. P. and Jankovská, V.: Vegetation of Eurasia from the last glacial maximum to present: Key biogeographic patterns, Quaternary Science Reviews, 157, 80–97, doi:10.1016/j.quascirev.2016.11.022, 2017.

Birks, H. J. B., Heiri, O., Seppä, H. and Bjune, A. E.: Strengths and weaknesses of quantitative climate reconstructions based on Late-Quaternary biological proxies, The Open Ecology Journal, 3, 68–110, doi:10.2174/1874213001003020068, 2010.

Brovkin, V., Raddatz, T., Reick, C. H., Claussen, M. and Gayler, V.: Global biogeophysical interactions between forest and climate, Geophysical Research Letters, 36(7), L07405, doi:10.1029/2009GL037543, 2009.

Cao, X., Tian, F., Li, F., Gaillard, M., Rudaya, N., Xu, Q. and Herzschuh, U.: Pollen-based quantitative land-cover reconstruction for northern Asia covering the last 40 ka cal BP, Climate of the Past, 15(4), 1503–1536, doi:10.5194/cp-15-1503-2019, 2019.

Chevalier, M., Davis, B. A. S., Heiri, O., Seppä, H., Chase, B. M., Gajewski, K., Lacourse, T., Telford, R. J., Finsinger, W., Guiot, J., Kühl, N., Maezumi, S. Y., Tipton, J. R., Carter, V. A., Brussel, T., Phelps, L. N., Dawson, A., Zanon, M., Vallé, F., Nolan, C., Mauri, A., de Vernal, A., Izumi, K., Holmström, L., Marsicek, J., Goring, S., Sommer, P. S., Chaput, M. and Kupriyanov, D.: Pollen-based climate reconstruction techniques for late Quaternary studies, Earth-Science Reviews, 210, 103384, doi:10.1016/j.earscirev.2020.103384, 2020.

Dallmeyer, A., Claussen, M., Ni, J., Cao, X., Wang, Y., Fischer, N., Pfeiffer, M., Jin, L., Khon, V., Wagner, S., Haberkorn, K. and Herzschuh, U.: Biome changes in Asia since the mid-Holocene - An analysis of different transient Earth system model simulations, Climate of the Past, 13(2), 107–134, doi:10.5194/cp-13-107-2017, 2017.

Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, CRC press., 1994.

Finsinger, W. and Tinner, W.: Holocene vegetation and land-use changes in response to climatic changes in the forelands of the southwestern Alps, Italy, Journal of Quaternary Science, 21(3), 243–258, doi:10.1002/jqs.971, 2006.

Finsinger, W., Lane, C. S., van Den Brand, G. J., Wagner-Cremer, F., Blockley, S. P. E. and Lotter, A. F.: The lateglacial Quercus expansion in the southern European Alps: Rapid vegetation response to a late Allerød climate warming?, Journal of Quaternary Science, 26(7), 694–702, doi:10.1002/jqs.1493, 2011.

Hargreaves, C. J., Dyer, M. S., Gaultois, M. W., Kurlin, V. A. and Rosseinsky, M. J.: The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions, Chemistry of Materials, 32(24), 10610–10620, doi:10.1021/acs.chemmater.0c03381, 2020.

Harris, I., Jones, P. D., Osborn, T. J. and Lister, D. H.: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset, International Journal of Climatology, 34(3), 623–642, doi:10.1002/joc.3711, 2014.

Kaplan, J. O.: Geophysical Applications of Vegetation Modeling., 2001.

Kaplan, J. O., Bigelow, N. H., Prentice, I. C., Harrison, S. P., Bartlein, P. J., Christensen, T. R., Cramer, W., Matveyeva, N. V., McGuire, A. D., Murray, D. F., Razzhivin, V. Y., Smith, B., Walker, D. A., Anderson, P. M., Andreev, A. A., Brubaker, L. B., Edwards, M. E. and Lozhkin, A. V.: Climate change and Arctic ecosystems: 2. Modeling, paleodata-model comparisons, and future projections, Journal of Geophysical Research: Atmospheres, 108(19), doi:10.1029/2002jd002559, 2003.

Levina, E. and Bickel, P.: The Earth Mover's distance is the Mallows distance: Some insights from statistics, Proceedings of the IEEE International Conference on Computer Vision, 2, 251–256, doi:10.1109/ICCV.2001.937632, 2001.

Litt, T., Ohlwein, C., Neumann, F. H., Hense, A. and Stein, M.: Holocene climate variability in the Levant from the Dead Sea pollen record, Quaternary Science Reviews, 49, 95–105, doi:10.1016/j.quascirev.2012.06.012, 2012.

Marsland, S. J., Haak, H., Jungclaus, J. H., Latif, M. and Röske, F.: The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates, Ocean Modelling, 5(2), 91–127, doi:10.1016/S1463-5003(02)00015-X, 2003.

Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E. B., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., Moore, W., Moss, R. B., Herzenberg, L. A. and Walther, G.: Earth Mover's Distance (EMD): A true metric for comparing biomarker expression levels in cell populations, PLoS ONE, 11(3), 1–14, doi:10.1371/journal.pone.0151859, 2016.

Overpeck, J. T., Webb III, T. and Prentice, I. C.: Quantitative interpretation of fossil pollen spectra: Dissimilarity coefficients and the method of modern analogs, Quaternary Research, 23(1), 87–108, doi:10.1016/0033-5894(85)90074-2, 1985.

Prentice, I. C. and Webb III, T.: BIOME 6000: reconstructing global mid-Holocene vegetation patterns from palaeoecological records, Journal of Biogeography, 25, 997–1005 [online] Available from: http://onlinelibrary.wiley.com/doi/10.1046/j.1365-2699.1998.00235.x/abstract (Accessed 30 January 2013), 1998.

Prentice, I. C., Guiot, J., Huntley, B., Jolly, D. and Cheddadi, R.: Reconstructing biomes from palaeoecological data: A general method and its application to European pollen data at 0 and 6 ka, Climate Dynamics, 12(3), 185–194, doi:10.1007/BF00211617, 1996.

Prentice, I. C., Harrison, S. P., Jolly, D. and Guiot, J.: The climate and biomes of Europe at 6000 yr BP: Comparison of model simulations and pollen-based reconstructions, Quaternary Science Reviews, 17(6–7), 659–668, doi:10.1016/S0277-3791(98)00016-X, 1998.

Prentice, I. C., Jolly, D. and Participants, B. 6000: Mid-Holocene and glacial maximum vegetation geography of the northern continents and Africa, Journal of Biogeography, 27(3), 507–519, doi:10.1046/j.1365-2699.2000.00425.x, 2001.

R Core Team: R: A Language and Environment for Statistical Computing, [online] Available from: http://www.r-project.org/, 2022.

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M. A., Hagemann, S., Kirchner, I., Kornblueh, L., Rhodin, A., Schlese, U., Schulzweida, U. and Tompkins, A.: The atmospheric general circulation model ECHAM5: Part 1: Model description, Report / MPI für Meteorologie, 349(November 2003), 1–140 [online] Available from: http://en.scientificcommons.org/8586047, 2003.

Rubner, Y., Tomasi, C. and Guibas, L. J.: Earth mover's distance as a metric for image retrieval, International Journal of Computer Vision, 50(2), 99–121, doi:10.1023/A:1026543900054, 2000.

Sato, H., Kelley, D. I., Mayor, S. J., Martin Calvo, M., Cowling, S. A. and Prentice, I. C.: Dry corridors opened by fire and low CO2 in Amazonian rainforest during the Last Glacial Maximum, Nature Geoscience, 14(8), 578–585, doi:10.1038/s41561-021-00777-2, 2021.

Sawada, M., Viau, A. E., Vettoretti, G., Peltier, W. R. and Gajewski, K.: Comparison of North-American pollen-based temperature and global lake-status with CCCma AGCM2 output at 6 ka, Quaternary Science Reviews, 23(3–4), 225–244, doi:10.1016/j.quascirev.2003.08.005, 2004.

Simpson, G. L.: Analogue Methods in Palaeoecology: Using the analogue Package, Journal of Statistical Software, 22(2), 1–29, doi:10.18637/jss.v022.i02, 2007.

Tian, F., Cao, X., Dallmeyer, A., Lohmann, G., Zhang, X., Ni, J., Andreev, A., Anderson, P. M., Lozhkin, A. V., Bezrukova, E., Rudaya, N., Xu, Q. and Herzschuh, U.: Biome changes and their inferred climatic drivers in northern and eastern continental Asia at selected times since 40 cal ka bp, Vegetation History and Archaeobotany, 27(2), 365–379, doi:10.1007/s00334-017-0653-8, 2018.

Urbanek, S. and Rubner, Y.: emdist: Earth Mover's Distance v0.3-2, [online] Available from: https://cran.r-project.org/package=emdist, 2022.

Wei, W. and Lohmann, G.: Simulated Atlantic Multidecadal Oscillation during the Holocene, Journal of Climate, 25(20), 6989–7002, doi:10.1175/JCLI-D-11-00667.1, 2012.

Wei, W., Lohmann, G. and Dima, M.: Distinct Modes of Internal Variability in the Global Meridional Overturning Circulation Associated with the Southern Hemisphere Westerly Winds, Journal of Physical Oceanography, 42(5), 785–801, doi:10.1175/JPO-D-11-038.1, 2012.

Wickham, H., Hester, J. and Chang, W.: devtools: Tools to Make Developing R Packages Easier (R package version 2.3.2), [online] Available from: https://cran.r-project.org/package=devtools, 2020.

Wohlfahrt, J., Harrison, S. P., Braconnot, P., Hewitt, C. D., Kitoh, A., Mikolajewicz, U., Otto-Bliesner, B. L. and Weber, S. L.: Evaluation of coupled ocean-atmosphere simulations of the mid-Holocene using palaeovegetation data from the northern hemisphere extratropics, Climate Dynamics, 31(7–8), 871–890, doi:10.1007/s00382-008-0415-5, 2008.

Zhang, X., Lohmann, G., Knorr, G. and Xu, X.: Different ocean states and transient characteristics in last glacial maximum simulations and implications for deglaciation, Climate of the Past, 9(5), 2319–2333, doi:10.5194/cp-9-2319-2013, 2013.