An improved statistical approach for reconstructing past climates from biotic assemblages: improved palaeoclimate reconstruction

Article

Accepted Version

To link to this article DOI: http://dx.doi.org/10.1098/rspa.2020.0346

Publisher: Royal Society Publishing

An improved statistical approach for reconstructing past climates from biotic assemblages

Target journal: Proceedings of the Royal Society A (Mathematics)

Article type: Research article

Authors:

Mengmeng Liu[1,*], Iain Colin Prentice[1,2,3], Cajo J. F. ter Braak[4], Sandy P. Harrison[3,5]

1: Department of Life Sciences, Imperial College London, Silwood Park Campus, Buckhurst Road, Ascot SL5 7PY, UK

2: Department of Biological Sciences, Macquarie University, North Ryde, NSW 2109, Australia

3: Ministry of Education Key Laboratory for Earth System Modelling, Department of Earth System Science, Tsinghua University, Beijing 100084, China

4: Biometris (Applied Mathematics and Applied Statistics Centre), Wageningen University & Research，6708 PB Wageningen, The Netherlands

5: Department of Geography and Environmental Science, University of Reading, Reading, RG6 6AB, UK

*** Corresponding author: m.liu18@imperial.ac.uk**

Abstract. Quantitative reconstructions of past climates are an important resource for evaluating how well climate models reproduce climate changes. One widely used statistical approach for making such reconstructions from fossil biotic assemblages is weighted averaging partial least squares regression (WA-PLS). There is however a known tendency for WA-PLS to yield reconstructions compressed towards the centre of the climate range used for calibration, potentially biasing the reconstructed past climates. We present an improvement of WA-PLS by assuming that: (a) the theoretical abundance of each taxon is unimodal with respect to the climate variable considered; (b) observed taxon abundances follow a multinomial distribution in which the total abundance of a sample is climatically uninformative; and (c) the estimate of the climate value at a given site and time makes the observation most probable, i.e. it maximizes the log-likelihood function. This climate estimate is approximated by weighting taxon abundances in WA-PLS by the inverse square of their climate tolerances. We further improve the approach by considering the frequency (*fx*) of the climate variable in the training data set. TWA-PLS with *fx* correction greatly reduces the compression bias, compared to WA-PLS, and improves model performance in reconstructions based on an extensive modern pollen data set.

Keywords: climate reconstruction, palaeoclimate, WA-PLS, bias reduction, model calibration, pollen data

1 Background

Past climate states allow tests of the models that are used to project climate responses to changes in atmospheric composition and land-use [1–4]. Direct measurements of climate only extend back to the 17th century [5] and in many regions are not available before the 20th century [6]. Reconstructions for earlier, and more different, palaeoclimate states have to be inferred from indicators that respond to climate. Most reconstructions of terrestrial palaeoclimates are based on biotic assemblages, including pollen, chironomids and diatoms preserved in sedimentary archives. The relationships between taxon abundances in these assemblages and a specific climate variable is derived using modern climate data and modern assemblages as a training data set. The inferred relationship is then used to reconstruct past climate from fossil assemblage data, assuming that the environmental space occupied by different taxa has remained the same through time.

Many different methods are used to obtain this indicator-climate relationship. Weighted averaging partial least squares regression (WA-PLS) [7,8] is one of the most widely used methods and has been applied to biotic indicators including pollen [9,10], chironomids [11,12] and diatoms [13,14]. However, one feature common to many WA-PLS reconstructions is that values reconstructed from the training data set tend to be higher than observed values at the low end, and lower at the high end, of the climate range. This artificial "compression" towards the central part of the range occurs whatever biotic indicator is being used [9–11,13,15,16]. Compression could result in the amplitude of climate changes being underestimated.

In this paper, we motivate an improved version of WA-PLS making use of information about the climatic tolerances of taxa, which vary considerably – taxa with narrow climatic ranges having greater indicator value than taxa with wide climatic ranges. Whereas tolerance down-weighting has been applied in simple two-way WA [17–20], it has not been used in WA-PLS and there has been no demonstration of its value for alleviating the compression issue. Climate values that occur frequently in the training data set might also cause bias, so we further improve the model by taking the frequency of climate values into account. Using a large modern pollen data set from Europe, the Middle East and northern Eurasia, we show that the new method reduces the compression bias, decreases root mean squared error of prediction (RMSEP) and increases R^2. Using two Holocene pollen records from the Iberian peninsula as examples, we show that the new method can sometimes produce significantly different results from the standard WA-PLS, which may possibly explain some known discrepancies between existing palaeo-reconstructions and model-simulated climates [21].

2 Methods

2.1 Theoretical basis

In counting pollen, the analyst determines how many pollen are counted and assigned to a taxon. In consequence, pollen data are compositional data, and the sample total does not convey information of interest. Counts are often transformed to percentages so that the sample total is 100, but the original counts can also be used if they sum to equal sample totals. In other words, counts are transformed to the proportions to the total counts at a site. The true abundances are not known; only the proportions can be observed.

Our approach as developed here is based on three assumptions:

(a) The theoretical abundance of each taxon follows a Gaussian (unimodal) curve with respect to each climate variable considered [17,22] as shown in Equation (1).

$$p_{ik}^{*} = e^{a_k - \frac{(x_i - u_k)^2}{2t_k^2}} \tag{1}$$

where p_{ik}^{*} is the theoretical abundance of the k^{th} taxon at the i^{th} site, a_k is the log-value of the theoretical maximum abundance of the k^{th} taxon, x_i is the value of the climate variable at the i^{th} site, u_k is the optimum (the ideal climate value) of the k^{th} taxon, and t_k is the tolerance (a measure of the breadth of the climatic distribution) of the k^{th} taxon.

(b) The observed abundances of taxa ($y_{i1}, y_{i2}, \ldots, y_{ik}, \ldots, y_{im}$) follow a multinomial distribution (Equation (2)) [23], in which the total abundance of a sample is climatically uninformative, with likelihood

$$f = \frac{\sum_{k=1}^{m} y_{ik}}{\prod_{k=1}^{m} y_{ik}!} \prod_{k=1}^{m} {p_{ik}}^{y_{ik}} \tag{2}$$

where f is the probability function of the multinomial distribution, y_{ik} is the observed abundance of the k^{th} taxon at the i^{th} site, m is the total number of taxa, and p_{ik} indicates the probability of observation y_{ik}, which is equal to the proportion of the theoretical abundance of the k^{th} taxon to the theoretical abundance of all taxa at the i^{th} site. p_{ik} can be expressed by Equation (3):

$$p_{ik} = \frac{p_{ik}^{*}}{\sum_{k'=1}^{m} p_{ik'}^{*}} \tag{3}$$

(c) The estimate of the climate value at a given site and time makes the observation most probable, i.e. it maximizes the log-likelihood function. Combining Equation (2) and (3), the log-likelihood at the i^{th} site [24,25] can be expressed as:

$$l = \log f = log \sum_{k=1}^{m} y_{ik} - \sum_{k=1}^{m} \log y_{ik}! + \sum_{k=1}^{m} y_{ik}\, log p_{ik}^{*} - \sum_{k=1}^{m} y_{ik}\, log \sum_{k'=1}^{m} p_{ik'}^{*} \tag{4}$$

The last term in Equation (4) is ignored for simplicity of derivation, a strategy supported by previous research [17,26] and in the Supplementary Material 1, so that:

$$l \approx log \sum_{k=1}^{m} y_{ik} - \sum_{k=1}^{m} \log y_{ik}! + \sum_{k=1}^{m} y_{ik}\, log p_{ik}^{*} \tag{5}$$

According to assumption (a), p_{ik}^{*} can be replaced by a function of x_i, so the log-likelihood function can be written as:

$$l \approx log \sum_{k=1}^{m} y_{ik} - \sum_{k=1}^{m} \log y_{ik}! + \sum_{k=1}^{m} y_{ik} \left(a_k - \frac{(x_i - u_k)^2}{2t_k^2} \right) \tag{6}$$

The derivative of the log-likelihood function to x_i is then given by:

$$\frac{\partial\, l}{\partial\, x_i} \approx -\sum_{k=1}^{m} \frac{y_{ik}}{t_k^2}(x_i - u_k) \tag{7}$$

The estimate of the climate value at the i^{th} site $\hat{x}_i$ is obtained by setting Equation (7) to zero [24,25]. The solution is

$$\hat{x}_i = \frac{\sum_{k=1}^{m} \frac{y_{ik} u_k}{t_k^2}}{\sum_{k=1}^{m} \frac{y_{ik}}{t_k^2}} \tag{8}$$

which is thus the approximate maximiser of equation (4). Here, y_{ik}/t_k^2 provides a weighting for u_k to provide a weighted average. This equation results in taxa with a more limited climate range being given more weight, which can be incorporated into the `WAPLS` and `predict` functions in the R package `rioja` [27] (Table 1).We have developed a package to do this, please see Supplementary Material 2 for a brief description of this package.

The estimated optimum ($\hat{u}_k$) and unbiased tolerance ($\hat{t}_k$) [28] of each taxon used in Equation (8) are calculated from the modern training data set [22] as follows:

$$\hat{u}_k = \frac{\sum_{i=1}^{n} y_{ik} x_i}{\sum_{i=1}^{n} y_{ik}} \tag{9}$$

$$\hat{t}_k = \sqrt{\frac{\sum_{i=1}^{n} y_{ik}(x_i - \hat{u}_k)^2}{(1 - 1/N_{2k})\ \sum_{i=1}^{n} y_{ik}}} \tag{10}$$

where

$$N_{2k} = \frac{1}{\sum_{i=1}^{n} \left(\frac{y_{ik}}{\sum_{i'=1}^{n} y_{i'k}}\right)^2} \tag{11}$$

where n is the total number of sites; y_{ik} is the observed abundance of the k^{th} taxon at the i^{th} site; x_i is the observed climate value at the i^{th} site; N_{2k} is the effective number of occurrences for the k^{th} taxon [28]. For binary abundance data, Equation (10) is precisely the sample (instead of: population) standard deviation.

Tolerance should be included in WA-PLS, as shown in Equation (8), thus the new approach can be called tolerance-weighted WA-PLS (TWA-PLS). The regression part of WA-PLS can also be improved. In the WA-PLS paper [8], step 7 is to regress the environmental variable x_i on the components obtained so far using weights $\frac{\sum_{k=1}^{m} y_{ik}}{\sum_{i=1}^{n}(\sum_{k=1}^{m} y_{ik})}$ (= constant 1/n) in the regression and take the fitted values as current estimates. This means that sites are given equal weights. However, modern pollen sites are often not sampled evenly, so their corresponding modern climate values do not follow a uniform distribution (Figure 4h-j). Frequent climate values might bias the regression and thus the current estimates using the components obtained so far. Therefore, the frequencies of the climate values at the modern sampling sites, *fx*, should also be taken into account. Because weighted averages are taken twice (the first time is to use weighted averaging of the climate values to calculate optima and tolerances of taxa, the second time is to use weighted averaging of optima and tolerances to estimate the climate values) [8], frequent values bias the calculation twice. Therefore, $1/fx^2$ should be used as weights in the regression, to reduce the bias brought by frequent climate values. Algorithms for WA-PLS and TWA-PLS, with and without *fx* correction, are shown in Table 1. The orthogonalization and standardization procedures are the same as the ones used in WA-PLS [29]. We use robust fitting of linear models (rlm) [30] in the regression step (Step 7 in Table 1) whereas WA-PLS in `rioja` uses least-squares fitting (lm); the difference in fit was found very minor for the data in this paper and we report results using rlm only.

The standard error of $\hat{x}_i$ can be obtained from the second derivative of the log-likelihood function, which can be calculated from Equation (7).

$$\frac{\partial^2 l}{\partial {x_i}^2} \approx -\sum_{k=1}^{m} \frac{y_{ik}}{t_k^2} \tag{12}$$

The Fisher information [24,25] is

$$I(\hat{x}_i) = -E\left(\frac{\partial^2 l}{\partial {x_i}^2}\right) \approx \sum_{k=1}^{m} \frac{y_{ik}}{t_k^2} \tag{13}$$

When the sample size is large, as is the case here, the standard error of the likelihood estimation can then be approximated [24,25] by:

$$se(\hat{x}_i) \approx \frac{1}{\sqrt{I(\hat{x}_i)}} \approx \frac{1}{\sqrt{\sum_{k=1}^{m} \frac{y_{ik}}{t_k^2}}} \tag{14}$$

This standard error corresponds to the maximum likelihood standard error given by ter Braak & Barendregt [17] when t_k is constant. Equation (14) has limited practical value as the pollen counts may show overdispersion compared to the multinomial distribution and also because the original pollen counts are often unavailable, but are being used in the equation. Bootstrap estimates of the standard error [20,31] are to be preferred.

2.2 Implementation

Modern pollen data were obtained from the SMPDS data set (the SPECIAL modern pollen data set) [32], which contains pollen assemblages from 6458 terrestrial sites from Europe, the Middle East and northern Eurasia (Figure 1a). The SMPDS data were derived from the European Modern Pollen Database (EMPD) v3.0 [33] and the EMBSeCBIO (Eastern Mediterranean-Black Sea-Caspian corridor BIOmes) Initiative [34], individual published records [35–44] obtained from the European Pollen Database (http://www.europeanpollendatabase.net/) or Pangaea (https://www.pangaea.de/), and 73 modern surface samples from northern Spain. Counts for obligate aquatics, insectivorous plants, non-native species and cultivated plants are not included in the SMPDS since their abundance is assumed not to be primarily controlled by climate. Some pollen types have been combined to a higher taxonomic level because they are not routinely identified across all the sites. There are 247 taxa included in the SMPDS, but some of these only occur in a small number of sites. For the current analysis, we used the 195 taxa that occur at > 10 sites.

Three bioclimatic variables at the locations of the SMPDS pollen sites (Figure 1b-d) were also obtained from the SMPDS data set [32]. This data set provides mean temperature of the coldest month (MTCO), growing degree days above a baseline of 0 °C (GDD_0) and a moisture index (MI), defined as an estimate of the ratio of annual precipitation to annual potential evapotranspiration, at each of the SMPDS pollen sites. These three variables reflect ecophysiological controls on plant distribution [32,45] that have been shown to influence the distribution and abundance of plant species independently of one another [46–48]. The individual and joint effects of these three variables were tested explicitly [49] for the SMPDS data set using Canonical Correspondence Analysis [26], and a strong correlation between species abundance and each of the three bioclimate variables was shown, with correlations of 0.83, 0.61 and 0.47 respectively for the first three CCA axes and VIF scores of < 6 for each bioclimatic variable, well within the range considered suitable for the application of regression methods in general.

Values of MTCO, GDD_0 and MI were obtained using a geographically-weighted regression of climatological values (1961-1990) of mean monthly temperature, precipitation, and fractional sunshine hours from the CRU CL v2.0 gridded data set [50] in order to correct for elevation differences between the CRU grid cells and the pollen sites. The climate of each pollen site was then estimated based on its longitude, latitude, and elevation. MTCO (Figure 1b) was taken directly from the GWR regression. GDD_0 (Figure 1c) was estimated from daily data using a mean-conserving interpolation [51] of the monthly mean temperatures. MI was calculated for each pollen site using SPLASH v1.0 [52] based on daily values of precipitation, temperature and sunshine hours obtained using a mean-conserving interpolation of the monthly values of each. We further transformed MI to an alternative measure of available moisture, α (Figure 1d), defined as the ratio of actual evapotranspiration to equilibrium evapotranspiration. The α index emphasises differences at the dry end of the climate range, which have a more pronounced effect on vegetation distribution than differences at the wet end [53]. We use the parametric Fu-Zhang formulation of the Budyko relationship to make this transformation:

$$\alpha = 1.26 \cdot MI \cdot \left(1 + \frac{1}{MI} - \left(1 + \left(\frac{1}{MI}\right)^{\omega}\right)^{\frac{1}{\omega}}\right) \qquad (15)$$

using $\omega = 3$ [54]. The derivation of this equation is given in Supplementary Material 3.

We use WA-PLS, TWA-PLS, WA-PLS with *fx* correction, and TWA-PLS with *fx* correction to reconstruct modern climates and compare them to observations at the modern pollen sites. When including frequency (*fx*) into the models in step 7 (Table 1), bins of 0.02, 20, 0.002 are used for MTCO, GDD_0 and α, respectively.

2.3 Evaluation

Comparison of these reconstructions against the observed climate at the training sites provides a test of model performance. The initial estimates of the climate optima and tolerances are refined based on the regression residuals until the incremental change in the values ceases to create predictive improvement [7,8], where the identification of the optimal number of components has often been based on whether there is significant improvement of the leave-out root-mean-square error (RMSEP). To reduce the risk of pseudo-replication, when using one site as the test site in the cross-validation, not only this site but also sites that are both geographically close (within 50 km horizontal distance from the site) and climatically close (within 2% of the full range of each climate variable in the data set) to this test site are removed from the training set. By doing this, multiple sites that provide the same information are not included and thus will not inflate the cross-validation statistics. The criterion used here to select the number of components in all cases was an abrupt increase in *p*-value, where *p* assesses whether using the current number of components is significantly different from using one component less.

We assess the degree of overall compression by fitting a linear regression line to the result. The closer the slope is to unity, the less the overall compression. We assess the degree of local compression by locally estimated scatterplot smoothing of the residuals. To compare the methods, we use the last significant number of components, because this would be the number used to make palaeoclimate reconstructions in practice.

To examine the implications of the new method for palaeoclimate reconstructions, we use fossil pollen data covering the Holocene (past ca 11,700 years) from Basa de la Mora [55] and Estanya [56,57]. Basa de la Mora (42.54527° N, 0.3255° E) is a high elevation lake site (1906m) in the central Pyrenees. Estanya (42.02826° N, 0.52905° E) is a lower-elevation lake site (677m) in the pre-Pyrenean foothills. We compare the reconstructions at the site made using the last significant number of components for each method. We obtain bootstrap estimates of the sample-specific errors by resampling the training set 1000 times [20,31], and then calculate 95 % confidence interval using 1.96 times sample-specific errors, to see if the confidence intervals of reconstructions using different methods overlap with each other. If they do not overlap, then the reconstructions show significant difference.

3 Results

3.1 Modern training results

Comparisons below are made using the last significant number of components (indicated in bold in Table 2) for each method. Comparisons using the same number of components can be found in Table 2 and Supplementary Material 4.

TWA-PLS has RMSEP of 4.58, 863, 0.153 for MTCO, GDD_0 and α, respectively, while WA-PLS has RMSEP of 5.05, 950, 0.165 for MTCO, GDD_0 and α, respectively; TWA-PLS with *fx* correction has RMSEP of 4.58, 869, 0.156 for MTCO, GDD_0 and α, respectively, while WA-PLS with *fx* correction has RMSEP of 5.20, 999, 0.172 for MTCO, GDD_0 and α, respectively (Table 2). Therefore, including tolerance (*t*) reduces RMSEP, while including frequency (*fx*) slightly increases RMSEP.

TWA-PLS has R^2 of 0.72, 0.69, 0.69 for MTCO, GDD_0 and α, respectively, while WA-PLS has R^2 of 0.66, 0.63, 0.64 for MTCO, GDD_0 and α, respectively; TWA-PLS with *fx* correction has R^2 of 0.73, 0.71, 0.69 for MTCO, GDD_0 and α, respectively, while WA-PLS with *fx* correction has an R^2 of 0.66, 0.63, 0.63 for MTCO, GDD_0 and α, respectively (Table 2). In general, both including tolerance (*t*) and frequency (*fx*) increase R^2.

The degree of overall compression is assessed by the slope of the linear regression; the degree of local compression is assessed by whether the residuals are around zero across the climate range in locally estimated scatterplot smoothing. The slope of TWA-PLS is 0.74, 0.70, 0.72 for MTCO, GDD_0 and α, respectively, while the slope of WA-PLS is 0.68, 0.65, 0.67 for MTCO, GDD_0 and α, respectively; the slope of TWA-PLS with *fx* correction is 0.82, 0.83, 0.79 for MTCO, GDD_0 and α, respectively, while the slope of WA-PLS with *fx* correction is 0.77, 0.78, 0.73 for MTCO, GDD_0 and α, respectively (Table 2). Including either tolerance (*t*) or frequency (*fx*) makes the slope closer to 1 and the residuals closer to 0, in other words, reduces the overall and local compression, and including both reduces the compression further (Table 2, Figure 2, Figure 3).

Of all the four methods, TWA-PLS with *fx* correction has the lowest compression, highest R^2 and second lowest RMSEP (its RMSEP is only slightly larger than that obtained using TWA-PLS). TWA-PLS with *fx* correction is therefore our recommended method. The abbreviation fxTWA-PLS will be used in the following text.

There is still a wide scatter for MTCO < −20 °C and compression bias at GDD_0 > 6000. This reflects the fact that few taxa occur either at extreme low winter temperatures or extreme high growing degree days, and thus there are too few taxa to constrain the model well at low MTCO and high GDD_0.

3.2 The causes of "compression" in WA-PLS

The equation used in WA-PLS [17,22] is:

$$\hat{x}_{iWA} = \frac{\sum_{k=1}^{m} y_{ik} u_k}{\sum_{k=1}^{m} y_{ik}} \tag{16}$$

Compared to Equation (8), WA-PLS corresponds to the special case when all taxon tolerances (t_k) are equal. However, this is far from reality generally (Figure 4e-g).

In the simple case of two taxa (n = 2), we have

$$\hat{x}_i = \frac{\frac{y_{i1} u_1}{t_1^2} + \frac{y_{i2} u_2}{t_2^2}}{\frac{y_{i1}}{t_1^2} + \frac{y_{i2}}{t_2^2}} \tag{17}$$

$$\hat{x}_{iWA} = \frac{y_{i1} u_1 + y_{i2} u_2}{y_{i1} + y_{i2}} \tag{18}$$

Taking $\hat{x}_i$ from $\hat{x}_{iWA}$ gives:

$$\hat{x}_{iWA} - \hat{x}_i = \frac{y_{i1} y_{i2}}{t_1^2 t_2^2\,(y_{i1} + y_{i2})\left(\frac{y_{i1}}{t_1^2} + \frac{y_{i2}}{t_2^2}\right)}(u_1 - u_2)(t_1^2 - t_2^2) \tag{19}$$

When $u_1 > u_2$, $t_1 > t_2$, $\hat{x}_{iWA} > \hat{x}_i$; when $u_1 < u_2$, $t_1 < t_2$, $\hat{x}_{iWA} > \hat{x}_i$; when $u_1 > u_2$, $t_1 < t_2$, $\hat{x}_{iWA} < \hat{x}_i$; when $u_1 < u_2$, $t_1 > t_2$, $\hat{x}_{iWA} < \hat{x}_i$. In all the four cases, $\hat{x}_{iWA}$ is always closer to the optimum of wide-spread taxon than $\hat{x}_i$ (Figure 4a-d). Furthermore, taxa with wide climate ranges are more abundantly represented in the centre of the climate range (Figure 4e-g), so that $\hat{x}_{iWA}$ is closer to the center than $\hat{x}_i$. This explains why reconstructions that do not take account of the different tolerances of different taxa will tend to show compression.

Another cause of compression is the non-uniformly sampled modern climates. More points are in the centre of the climate range (Figure 4h-j). If points are given the same weights, as in WA-PLS, the centre of the climate range has too much weight in the linear regression in step

7 in Table 1. This will make the fitted line flatter, so the fitted values will be compressed towards the centre.

3.3 Estimation of past climate states

All three climate variables at both sites show larger ranges using fxTWA-PLS (TWA-PLS with *fx* correction) than using WA-PLS (Figure 5d-f, 6d-f), as might be expected from the reduction in compression.

The reconstructions of GDD_0 and α over the Holocene at Basa de la Mora (Figure 5b, 5c) using the two methods are similar but there is a large difference in the reconstructed MTCO (Figure 5a). The MTCO reconstructions using fxTWA-PLS do not overlap with WA-PLS reconstructions and are on average ca 3 °C warmer. At Estanya, the two methods also produce significant differences in the reconstructions of MTCO (Figure 6a). GDD_0 and α reconstructions show more difference (Figure 6b, 6c) compared to Basa de la Mora.

Differences between the two reconstruction techniques reflect how far the reconstructed climate is from the centre of the climate range, where compression of WA-PLS is much lower than at the two ends. This centre point can be calculated from the slope (b_1) and intercept (b_0) of WA-PLS in Table 2, by setting $b_0 + b_1x - x$ to zero. Centre points for the three climate variables are shown in dashed horizontal lines in Figure 5 and Figure 6. When above the dashed line, WA-PLS reconstructions tend to be lower than fxTWA-PLS reconstructions (Figure 5a, 6a, 6b); when below this line, WA-PLS reconstructions tend to be higher than fxTWA-PLS reconstructions (Figure 6c); when roughly around the dashed line, WA-PLS reconstructions tend to be similar to fxTWA-PLS reconstructions (Figure 5b,5c). In other words, WA-PLS reconstructions tend to be closer to the centre, biasing the reconstructed past climates.

4 Discussion

4.1 Comparison with a Bayesian approach

The new approach, fxTWA-PLS , offers an improvement compared to the standard WA-PLS method in the sense that it shows lower RMSEP, higher R^2 and less compression towards the centre of the climate range.

Another promising approach compared to WA-PLS is Bayesian climate reconstruction [10,58,59]. We run the Bayesian User-friendly Model for Palaeo-Environmental Reconstruction (BUMPER) [59] using the same training data set. BUMPER does not provide the leave-out (multiple sites) cross validation used in this paper, so instead we compare leave-one-out cross validation results between our methods and BUMPER (Supplementary Material 5). BUMPER standard model with full taxa has the lowest RMSEP and highest R^2 among the four BUMPER models (standard model including all taxa, standard model only including taxa with more than 2% abundance, presence-absence model including all taxa, presence-absence model only including taxa with more than 2% abundance) (Table S5.1). This best BUMPER model has RMSEP of 4.42, 882, 0.166 and R^2 of 0.74, 0.72, 0.71 for MTCO, GDD_0 and α, respectively (Table S5.1). It shows better performance than WA-PLS, which has RMSEP of 4.85, 905, 0.158 and R^2 of 0.69, 0.67, 0.67 for MTCO, GDD_0 and α, respectively (Table S5.1, S5.2). However, it is not as good as fxTWA-PLS which has RMSEP of 4.37, 830, 0.148 and R^2 of 0.76, 0.73, 0.72 for MTCO, GDD_0 and α, respectively (Table S5.1, S5.2). The overall compression of BUMPER is better than fxTWA-PLS (Table S5.1, Figure S5.1, Figure 2), however, the local compression is much worse, with skewed residuals for all the three climate variables (Figure S5.2, Figure 3).

4.2 Palaeo-reconstructions

We have shown that WA-PLS and fxTWA-PLS produce different estimates of past climates when the climate to be reconstructed is not in the centre of the climate range used for model training. By reducing the compression bias, fxTWA-PLS allows for reconstructions of more extreme climate changes. The reduction of compression bias may help to explain some known discrepancies between existing palaeo-reconstructions and model-simulated climates [21].

Reconstructed MTCO and GDD_0 at 0 cal yr BP at Basa de la Mora are warmer than the observed modern climate, while α is drier (Figure 5a-c). In mountain regions, pollen from lower sites is transported upward by daytime orographic winds [60,61]. Pollen assemblages are therefore biased towards lowland taxa, and pollen-based climate reconstructions at high-elevation sites tend to show warmer and drier conditions than those in the immediate surroundings of the site, which are reflected in the modern training set (see Figure S6.1). This explains why the reconstructed temperatures are higher, and the reconstructed moisture lower, than observed 0 cal yr BP at Basa de la Mora. The discrepancies are smaller at Estanya (Figure 6a-c), at lower elevation (Figure S6.1).

4.3 Potential issues with the application

The method assumes that the abundance of each taxon follows a Gaussian (unimodal) curve with respect to each climate variable. A few pollen taxa do not have unimodal distributions in climate space. For example, *Artemisia* occurs in both warm and cold steppe environments, because its distribution is strongly controlled by plant-available moisture and largely insensitive to temperature [48]. It would be possible to screen the training data set for taxa that do not display unimodal Gaussian relationships with a specific climate variable. Previous research has used generalized additive models (GAMs) to view the climate space of the taxa [48], which can help check for non-unimodality. Inspection of GAMs cannot unambiguously detect multimodality, given that the true abundances of taxa are not observed. When a taxon has a large tolerance and a low true abundance, its proportion to the total abundance can show multimodality even when the true abundance is unimodal (Figure S7.2). Observed multimodality may make the estimated optimum and tolerance of such a taxon less accurate. However, the training data set includes 195 taxa in total, all of which contribute to the final reconstruction; and removing individual taxa has very little impact on the results (removing *Artemisia*, for example: Figure S7.3, S7.4).

Underpinning the assumption that each taxon follows a unimodal curve with respect to each climate variable is that the reconstructed variables influence the distribution and abundance of plant species independently. This assumption would need to be tested when fxTWA-PLS is applied in other regions, or using different indicators. In addition, different bin widths to capture the trend of *fx* might result in slightly different results, because using too large a bin would lose many details while using too small a bin would lose the overall trend. Different bin widths can be tried to determine which to use, when using different indicators or different training data sets.

The theory underpinning the new approach makes use of maximum likelihood estimation, which gives maximum efficiency in large samples [62]. In general, fossil pollen assemblages contain fewer taxa than modern pollen assemblages, and the number of taxa represented can be small. Depauperate fossil pollen assemblages tend to reflect anomalous situations, for example, where the sediments have been partially oxidised and more fragile pollen types have been lost. Depauperate assemblages also tend to occur when sedimentation is discontinuous or dominated by erosion. There is no obvious solution for these problems, except by using high reconstruction uncertainties to identify unreliable samples.

A further potential limitation in the use of fxTWA-PLS is that taxa with narrow climate ranges tend to be represented by fewer samples in the training data set (Table 3), which can make estimation of their optima and tolerances less reliable. Upweighting taxa with narrow climate ranges can make the reconstructions less stable and increase uncertainties. This is reflected in Table 2: tolerance weighting sometimes induces larger maximum bias, although with lower RMSEP and higher R^2 overall. The training data set used here contains > 6000 samples and covers a wide range of climates, but there still are gaps in the coverage of climate space. A further expansion of the data set [e.g. ref 63], targeting sites that might help to fill in the climate space of taxa with narrow climate ranges, would be beneficial for future applications.

Data accessibility

We have developed a package `fxTWAPLS` (https://special-uor.github.io/fxTWAPLS/) to apply the new approach. The codes for this package can be found at https://github.com/special-uor/fxTWAPLS. This package is now available on CRAN. See electronic supplementary material, S2 for a brief description of this package. Version 0.0.2 is used in this paper. We have uploaded the data and other codes used in this paper as electronic supplementary material, S8. We have also uploaded the data and codes for BUMPER as electronic supplementary material, S9.

Acknowledgements

We thank Dr. Penélope González Sampériz and Dr. Graciela Gil Romera for providing the fossil pollen data, and Dongyang Wei for her kind assistance with the assessment of the individual and joint effects of the bioclimatic variables for the SMPDS data set. We also thank Huanyuan Zhang for his help in running BUMPER on Imperial College's High Performance Computer system and Roberto Villegas-Diaz for his help in developing our codes into a package.

Funding

ML acknowledges support from Imperial College through the Lee Family Scholarship. SPH acknowledges support from the European Research Council (ERC) for "GC2.0: Unlocking the past for a clearer future" (ERC-2015-AdG 694481) and from the JPI-Belmont project "PAlaeo-Constraints on Monsoon Evolution and Dynamics (PACMEDY)" through the UK Natural Environmental Research Council (NERC: NE/P006752/1). ICP acknowledges funding from the ERC, under the European Union's Horizon 2020 research and innovation programme (grant agreement No: 787203 REALM).

References

1. Braconnot P, Harrison SP, Kageyama M, Bartlein PJ, Masson-Delmotte V, Abe-Ouchi A, Otto-Bliesner B, Zhao Y. 2012 Evaluation of climate models using palaeoclimatic data. *Nat. Clim. Chang.* **2**, 417–424.

2. Schmidt GA *et al.* 2014 Using palaeo-climate comparisons to constrain future projections in CMIP5. *Clim. Past* **10**, 221–250. (doi:10.5194/cp-10-221-2014)

3. Harrison SP, Bartlein PJ, Izumi K, Li G, Annan J, Hargreaves J, Braconnot P, Kageyama M. 2015 Evaluation of CMIP5 palaeo-simulations to improve climate projections. *Nat. Clim. Chang.* **5**, 735–743.

4. Brierley CM *et al.* 2020 Large-scale features and evaluation of the PMIP4-CMIP6 midHolocene simulations. *Clim. Past Discuss.* **2020**, 1–35. (doi:10.5194/cp-2019-168)

5. Freeman E *et al.* 2019 The International Comprehensive Ocean-Atmosphere Data Set – Meeting users needs and future priorities. *Front. Mar. Sci.* **6**, 435.

6. Morice CP, Kennedy JJ, Rayner NA, Jones PD. 2012 Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: The HadCRUT4 data set. *J. Geophys. Res. Atmos.* **117**. (doi:10.1029/2011JD017187)

7. Birks HJB. 2003 Quantitative palaeoenvironmental reconstructions from Holocene biological data. In *Global Change in the Holocene* (eds A Mackay, RW Battarbee, HJB Birks, F Oldfield), pp. 107–123. London: Arnold.

8. ter Braak CJF, Juggins S. 1993 Weighted averaging partial least squares regression (WA-PLS): An improved method for reconstructing environmental variables from species assemblages. *Hydrobiologia* **269**, 485–502. (doi:10.1007/BF00028046)

9. Shen C, Liu K, Tang L, Overpeck JT. 2006 Quantitative relationships between modern pollen rain and climate in the Tibetan Plateau. *Rev. Palaeobot. Palynol.* **140**, 61–77. (doi:https://doi.org/10.1016/j.revpalbo.2006.03.001)

10. Salonen JS, Ilvonen L, Seppä H, Holmström L, Telford RJ, Gaidamavičius A, Stančikaitė M, Subetto D. 2011 Comparing different calibration methods (WA/WA-PLS regression and Bayesian modelling) and different-sized calibration sets in pollen-based quantitative climate reconstruction. *The Holocene* **22**, 413–424. (doi:10.1177/0959683611425548)

11. Heiri O, Lotter AF, Hausmann S, Kienast F. 2003 A chironomid-based Holocene summer air temperature reconstruction from the Swiss Alps. *The Holocene* **13**, 477–484. (doi:10.1191/0959683603hl640ft)

12. Tarrats P, Heiri O, Valero-Garcés B, Cañedo-Argüelles M, Prat N, Rieradevall M, González-Sampériz P. 2018 Chironomid-inferred Holocene temperature reconstruction in Basa de la Mora Lake (Central Pyrenees). *The Holocene* **28**, 1685–1696. (doi:10.1177/0959683618788662)

13. Bigler C, Hall RI, Renberg I. 2000 A diatom-training set for palaeoclimatic inferences from lakes in northern Sweden. *SIL Proceedings, 1922-2010* **27**, 1174–1182. (doi:10.1080/03680770.1998.11901421)

14. Fritz SC, Juggins S, Battarbee RW, Engstrom DR. 1991 Reconstruction of past changes in salinity and climate using a diatom-based transfer function. *Nature* **352**,

706–708. (doi:10.1038/352706a0)

15. Brooks SJ, Birks HJB. 2001 Chironomid-inferred air temperatures from Lateglacial and Holocene sites in north-west Europe: Progress and problems. *Quat. Sci. Rev.* **20**, 1723–1741. (doi:https://doi.org/10.1016/S0277-3791(01)00038-5)

16. Seppä H, Birks HJB, Odland A, Poska A, Veski S. 2004 A modern pollen–climate calibration set from northern Europe: developing and testing a tool for palaeoclimatological reconstructions. *J. Biogeogr.* **31**, 251–267. (doi:10.1111/j.1365-2699.2004.00923.x)

17. ter Braak CJF, Barendregt LG. 1986 Weighted averaging of species indicator values: Its efficiency in environmental calibration. *Math. Biosci.* **78**, 57–72. (doi:https://doi.org/10.1016/0025-5564(86)90031-3)

18. ter Braak CJF, van Dam H. 1989 Inferring pH from diatoms: a comparison of old and new calibration methods. *Hydrobiologia* **178**, 209–223. (doi:10.1007/BF00006028)

19. Juggins S, Birks HJB. 2012 Quantitative Environmental Reconstructions from Biological Data. In *Tracking Environmental Change Using Lake Sediments: Data Handling and Numerical Techniques.* (eds HJB Birks, AF Lotter, S Juggins, JP Smol), pp. 431–494. Dordrecht: Springer Netherlands.

20. Birks HJB, Simpson GL. 2013 'Diatoms and pH reconstruction' (1990) revisited. *J. Paleolimnol.* **49**, 363–371. (doi:10.1007/s10933-013-9697-7)

21. Mauri A, Davis BAS, Collins PM, Kaplan JO. 2015 The climate of Europe during the Holocene: A gridded pollen-based reconstruction and its multi-proxy evaluation. *Quat. Sci. Rev.* **112**, 109–127. (doi:10.1016/j.quascirev.2015.01.013)

22. ter Braak CJF, Prentice IC. 1988 A theory of gradient analysis. *Adv. Ecol. Res.* **18**, 271–317. (doi:10.1016/S0065-2504(08)60183-X)

23. Forbes C, Evans M, Hastings N, Peacock B. 2010 Multinomial Distribution. In *Statistical distributions*, pp. 135–136. Oxford: Wiley-Blackwell.

24. Millar RB. 2011 *Maximum likelihood estimation and inference with examples in R, SAS, and ADMB*. Chichester, Sussex, U.K.: Wiley.

25. Zacks S. 1971 *The theory of statistical inference*. New York (N.Y.) : Wiley. See http://lib.ugent.be/catalog/rug01:000474697.

26. ter Braak CJF. 1988 Partial canonical correspondence analysis. In *Classification and related methods of data analysis* (ed HH Bock), pp. 551–558. Amsterdam: Elsevier Science Publishers B.V. (North-Holland).

27. Juggins S. 2017 rioja: Analysis of Quaternary Science Data.R package version (0.9-21).

28. ter Braak CJF, Verdonschot PFM. 1995 Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquat. Sci.* **57**, 255–289. (doi:10.1007/BF00877430)

29. ter Braak CJF. 1987 Ordination. In *Data analysis in community and landscape ecology* (eds RHG. Jongman, CJF ter Braak, OFR. Van Tongeren), pp. 91–173. Pudoc, Wageningen.

30. Venables WN, Ripley BD. 2002 *Modern Applied Statistics with S*. Fourth. New York: Springer. See http://www.stats.ox.ac.uk/pub/MASS4/.

31. Birks HJB, ter Braak CJF, Line JM, Juggins S, Stevenson AC. 1990 Diatoms and pH reconstruction. *Philos. Trans. R. Soc. London. B, Biol. Sci.* **327**, 263–278. (doi:10.1098/rstb.1990.0062)

32. Harrison SP. 2020 Climate reconstructions for the SMPDS v1 modern pollen data set. (doi:10.5281/zenodo.3605003)

33. Davis BAS *et al.* 2013 The European Modern Pollen Database (EMPD) project. *Veg. Hist. Archaeobot.* **22**, 521–530. (doi:10.1007/s00334-012-0388-5)

34. Marinova E *et al.* 2017 Pollen-derived biomes in the Eastern Mediterranean–Black Sea–Caspian-Corridor. *J. Biogeogr.* **45**, 484–499. (doi:10.1111/jbi.13128)

35. Saadi F, Bernard J. 1991 Rapport entre la pluie pollinique actuelle, le climat et la 733 vegetation dans les steppes à Artemisia et les milieu limitrophes au Maroc. *Palaeoecol. Africa* **22**, 67–86.

36. de Klerk P, Haberl A, Kaffke A, Krebs M, Matchutadze I, Minke M, Schulz J, Joosten H. 2009 Vegetation history and environmental development since ca 6000 cal yr BP in and around Ispani 2 (Kolkheti lowlands, Georgia). *Quat. Sci. Rev.* **28**, 890–910. (doi:10.1016/j.quascirev.2008.12.005)

37. Grüger E, Jerz H. 2010 Untersuchung einer Doline auf dem Zugspitzplatt. *Quat. Sci. J.* **59**, 66–75.

38. Werner K, Tarasov PE, Andreev AA, Müller S, Kienast F, Zech M, Zech W, Diekmann B. 2010 A 12.5-kyr history of vegetation dynamics and mire development with evidence of Younger Dryas larch presence in the Verkhoyansk Mountains, East Siberia, Russia. *Boreas* **39**, 56–68. (doi:10.1111/j.1502-3885.2009.00116.x)

39. Müller S, Tarasov PE, Andreev AA, Tütken T, Gartz S, Diekmann B. 2010 Late Quaternary vegetation and environments in the Verkhoyansk Mountains region (NE Asia) reconstructed from a 50-kyr fossil pollen record from Lake Billyakh. *Quat. Sci. Rev.* **29**, 2071–2086. (doi:https://doi.org/10.1016/j.quascirev.2010.04.024)

40. Tarasov PE *et al.* 2011 Progress in the reconstruction of Quaternary climate dynamics in the Northwest Pacific: A new modern analogue reference dataset and its application to the 430-kyr pollen record from Lake Biwa. *Earth-Science Rev.* **108**, 64–79. (doi:https://doi.org/10.1016/j.earscirev.2011.06.002)

41. Matthias I, Semmler MSS, Giesecke T. 2015 Pollen diversity captures landscape structure and diversity. *J. Ecol.* **103**, 880–890. (doi:10.1111/1365-2745.12404)

42. Niemeyer B, Klemm J, Pestryakova LA, Herzschuh U. 2015 Relative pollen productivity estimates for common taxa of the northern Siberian Arctic. *Rev. Palaeobot. Palynol.* **221**, 71–82. (doi:https://doi.org/10.1016/j.revpalbo.2015.06.008)

43. Bell BA, Fletcher WJ. 2016 Modern surface pollen assemblages from the Middle and High Atlas, Morocco: Insights into pollen representation and transport. *Grana* **55**, 286–301. (doi:10.1080/00173134.2015.1108996)

44. Novenko E, Mazei N, Kusilman M. 2017 Tree pollen representation in surface pollen assemblages from different vegetation zones of European Russia. *Ecol. Quest. Vol 26*

45. Woodward FI. 1987 *Climate and Plant Distribution*. Cambridge, UK: Cambridge University Press.

46. Boucher-Lalonde V, Morin A, Currie DJ. 2012 How are tree species distributed in climatic space? A simple and general pattern. *Glob. Ecol. Biogeogr.* **21**, 1157–1166. (doi:10.1111/j.1466-8238.2012.00764.x)

47. Wang H, Prentice IC, Ni J. 2013 Data-based modelling and environmental sensitivity of vegetation in China. *Biogeosciences* **10**, 5817–5830. (doi:10.5194/bg-10-5817-2013)

48. Wei D, Prentice IC, Harrison SP. 2020 The climatic space of European pollen taxa. *Ecology* **101**, e03055. (doi:10.1002/ecy.3055)

49. Turner MG, Wei D, Prentice IC, Harrison SP. 2020 The impact of methodological decisions on climate reconstructions using WA-PLS. *Quat. Res.* , 1–16. (doi:DOI: 10.1017/qua.2020.44)

50. New M, Lister D, Hulme M. 2002 A high-resolution data set of surface climate over global land areas . *Clim. Res.* **21**, 1–25.

51. Rymes MD, Myers DR. 2001 Mean preserving algorithm for smoothly interpolating averaged data. *Sol. Energy* **71**, 225–231. (doi:https://doi.org/10.1016/S0038-092X(01)00052-4)

52. Davis TW *et al.* 2017 Simple process-led algorithms for simulating habitats (SPLASH v.1.0): robust indices of radiation, evapotranspiration and plant-available moisture. *Geosci. Model Dev.* **10**, 689–708. (doi:10.5194/gmd-10-689-2017)

53. Prentice IC, Cleator SF, Huang YH, Harrison SP, Roulstone I. 2017 Reconstructing ice-age palaeoclimates: Quantifying low-CO_2 effects on plants. *Glob. Planet. Change* **149**, 166–176. (doi:https://doi.org/10.1016/j.gloplacha.2016.12.012)

54. Sposito G. 2017 Understanding the Budyko equation. *Water* **9**, 236. (doi:10.3390/w9040236)

55. Pérez-Sanz A *et al.* 2013 Holocene climate variability, vegetation dynamics and fire regime in the central Pyrenees: the Basa de la Mora sequence (NE Spain). *Quat. Sci. Rev.* **73**, 149–169. (doi:https://doi.org/10.1016/j.quascirev.2013.05.010)

56. González-Sampériz P *et al.* 2017 Environmental and climate change in the southern Central Pyrenees since the Last Glacial Maximum: A view from the lake records. *Catena* **149**, 668–688. (doi:https://doi.org/10.1016/j.catena.2016.07.041)

57. Morellón M *et al.* 2011 Climate changes and human activities recorded in the sediments of Lake Estanya (NE Spain) during the Medieval Warm Period and Little Ice Age. *J. Paleolimnol.* **46**, 423–452. (doi:10.1007/s10933-009-9346-3)

58. Holden PB, Mackay AW, Simpson GL. 2008 A Bayesian palaeoenvironmental transfer function model for acidified lakes. *J. Paleolimnol.* **39**, 551–566. (doi:10.1007/s10933-007-9129-7)

59. Holden PB, Birks HJB, Brooks SJ, Bush MB, Hwang GM, Matthews-Bird F, Valencia BG, van Woesik R. 2017 BUMPER v1.0: a Bayesian user-friendly model for palaeo-environmental reconstruction. *Geosci. Model Dev.* **10**, 483–498. (doi:10.5194/gmd-10-483-2017)

60. Tsukada M. 1958 Untersuchungen über das Verhältniss zwischen dem Pollengehalt der Oberflächenproben und der Vegetation des Hochlandes Shiga. *J. Inst. Polytech.* **9**, 235–249.

61. Takahara H, Sugita S, Harrison S, Miyoshi N, Morita Y, Uchiyama T. 2000 Pollen-based reconstructions of Japanese biomes at 0,6000 and 18,000 14C yr BP. *J. Biogeogr.* **27**, 665–683. (doi:10.1046/j.1365-2699.2000.00432.x)

62. Wilks SS. 1938 The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *Ann. Math. Stat.* **9**, 60–62. (doi:10.1214/aoms/1177732360)

63. Davis BAS *et al.* 2020 The Eurasian Modern Pollen Database (EMPD), Version 2. *Earth Syst. Sci. Data Discuss.* **2020**, 1–41. (doi:10.5194/essd-2020-14)

Figure 1. The modern pollen and climate data sets. (a) Distribution of modern pollen data from the SMPDS data set; inferred (b) mean temperature of the coldest month (MTCO), (c) growing degree days above a baseline of 0 °C (GDD_0) and (d) plant-available moisture (α) at each pollen site as estimated using geographically-weighted regression. The MTCO and GDD_0 estimates are from the SMPDS data set; α was calculated from the moisture index (MI) provided in the SMPDS data set.

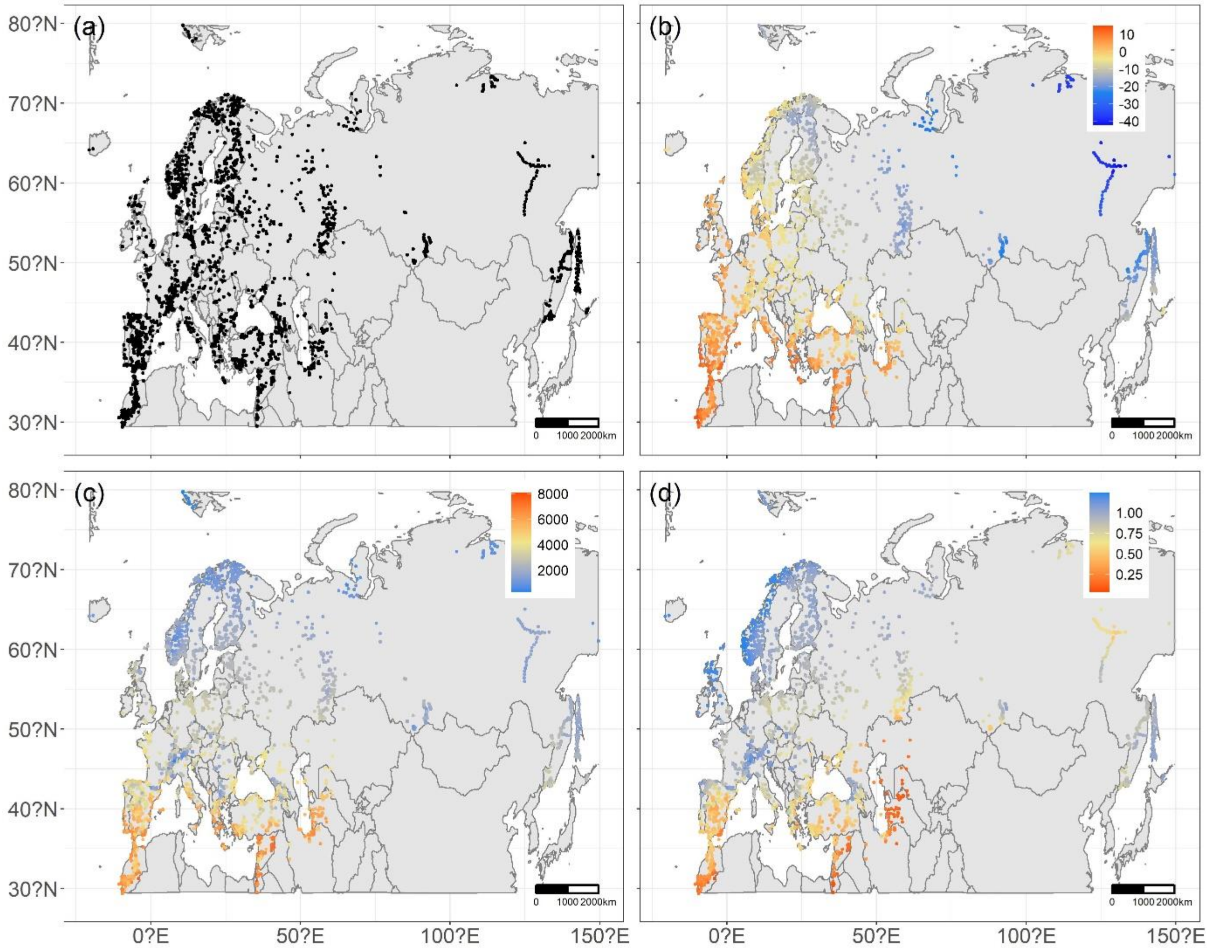

Figure 2. Reconstructed modern climates using the last significant number of components. The x axis is the observed modern climate value, the y axis is the modern climate value reconstructed from modern pollen data using WA-PLS, TWA-PLS, WA-PLS with *fx* correction, TWA-PLS with *fx* correction, respectively from top to bottom. The 1:1 line is shown in black, the linear regression line is shown in red, to show the degree of overall compression.

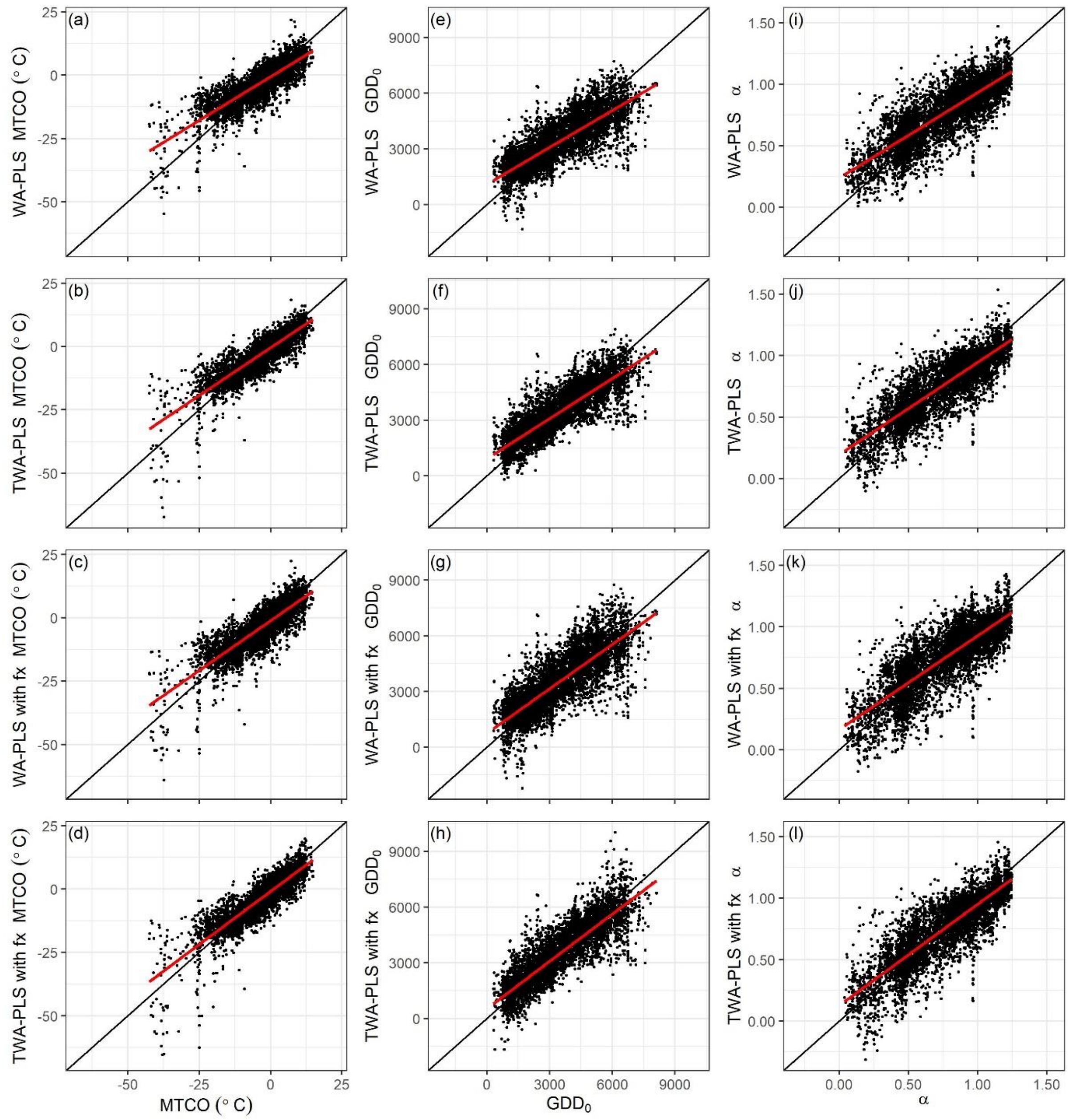

Figure 3. Residuals of reconstructed modern climates using the last significant number of components. The x axis is the observed modern climate value, the y axis is the residual of modern climate reconstruction using WA-PLS, TWA-PLS, WA-PLS with *fx* correction, TWA-PLS with *fx* correction, respectively, from top to bottom. The zero line is shown in black, the locally estimated scatterplot smoothing is shown in red, to show the degree of local compression.

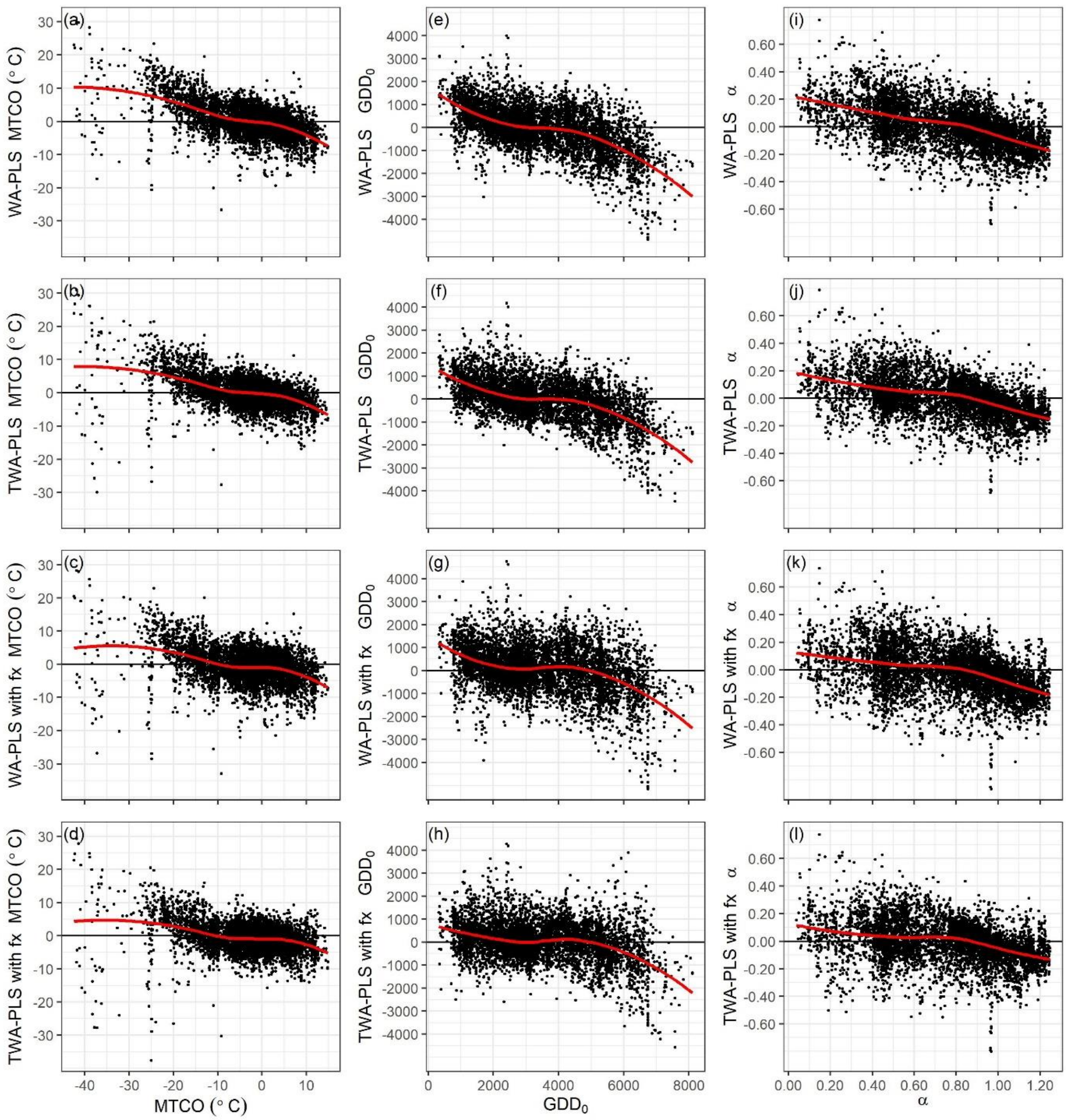

Figure 4. The principle of compression in WA-PLS. (a), (b), (c) and (d) are the four circumstances of optimum (*u*) and tolerance (*t*), $\hat{x}_{iWA}$ is the reconstructed value without tolerance, $\hat{x}_i$ is the reconstructed value with tolerance, the curves are the unimodal Gaussian curves (abundance to *x*) of the taxa. For (e), (f) and (g), the y axis is the tolerance of MTCO, GDD_0 and α, respectively; the x axis is the optimum of MTCO, GDD_0 and α, respectively. (h), (i) and (j) show the histograms of MTCO, GDD_0 and α, using bins of 0.02, 20, 0.002, respectively.

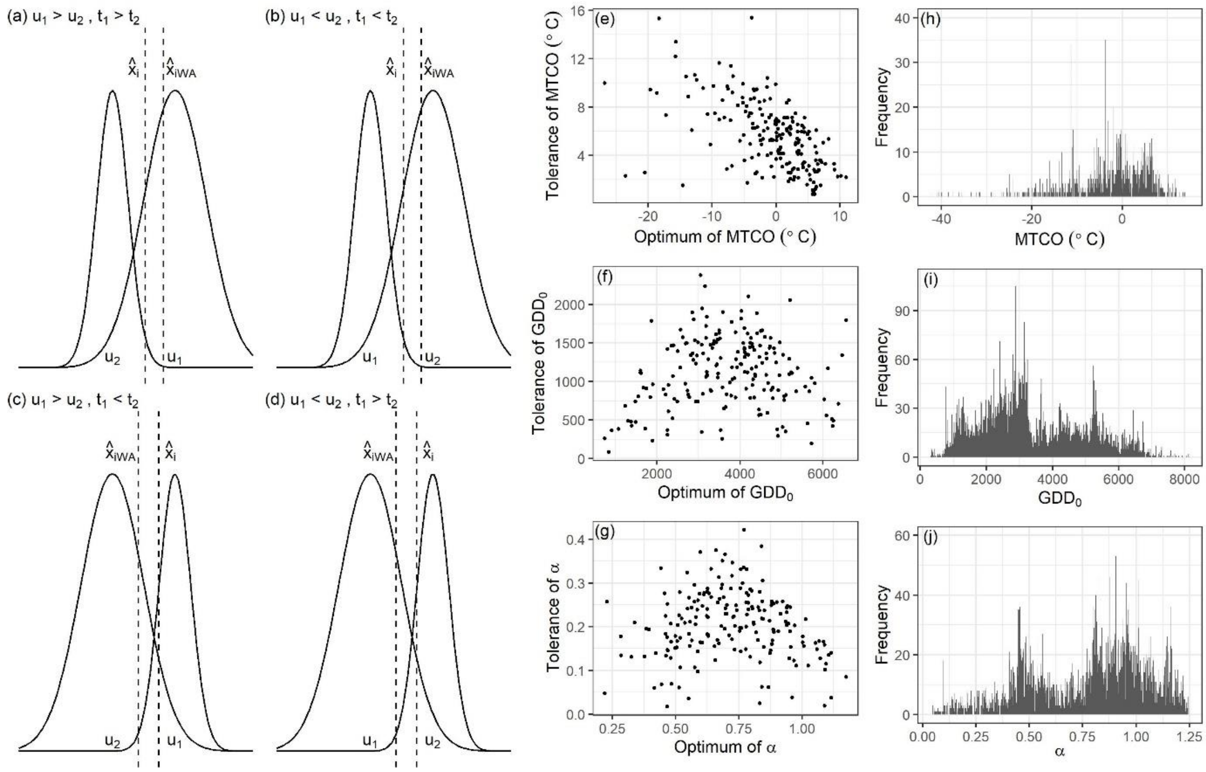

Figure 5. Comparison of downcore reconstructions of (a) MTCO, (b) GDD_0 and (c) α, at Basa de la Mora made using WA-PLS (black line) and TWA-PLS with *fx* correction (red line). The shades are 95% confidence intervals (reconstructions plus or minus 1.96 times their bootstrap estimates of sample-specific errors) of reconstructions using WA-PLS (black shade) and TWA-PLS with *fx* correction (red shade). The bar graphs show the range (maximum minus minimum) of reconstructed values over the Holocene for (d) MTCO, (e) GDD_0, (f) α using the two methods. Lines at 0 cal yr BP show the observed modern climate values at the site. Dashed horizontal lines show the estimate of the central range of the climate in the training data set.

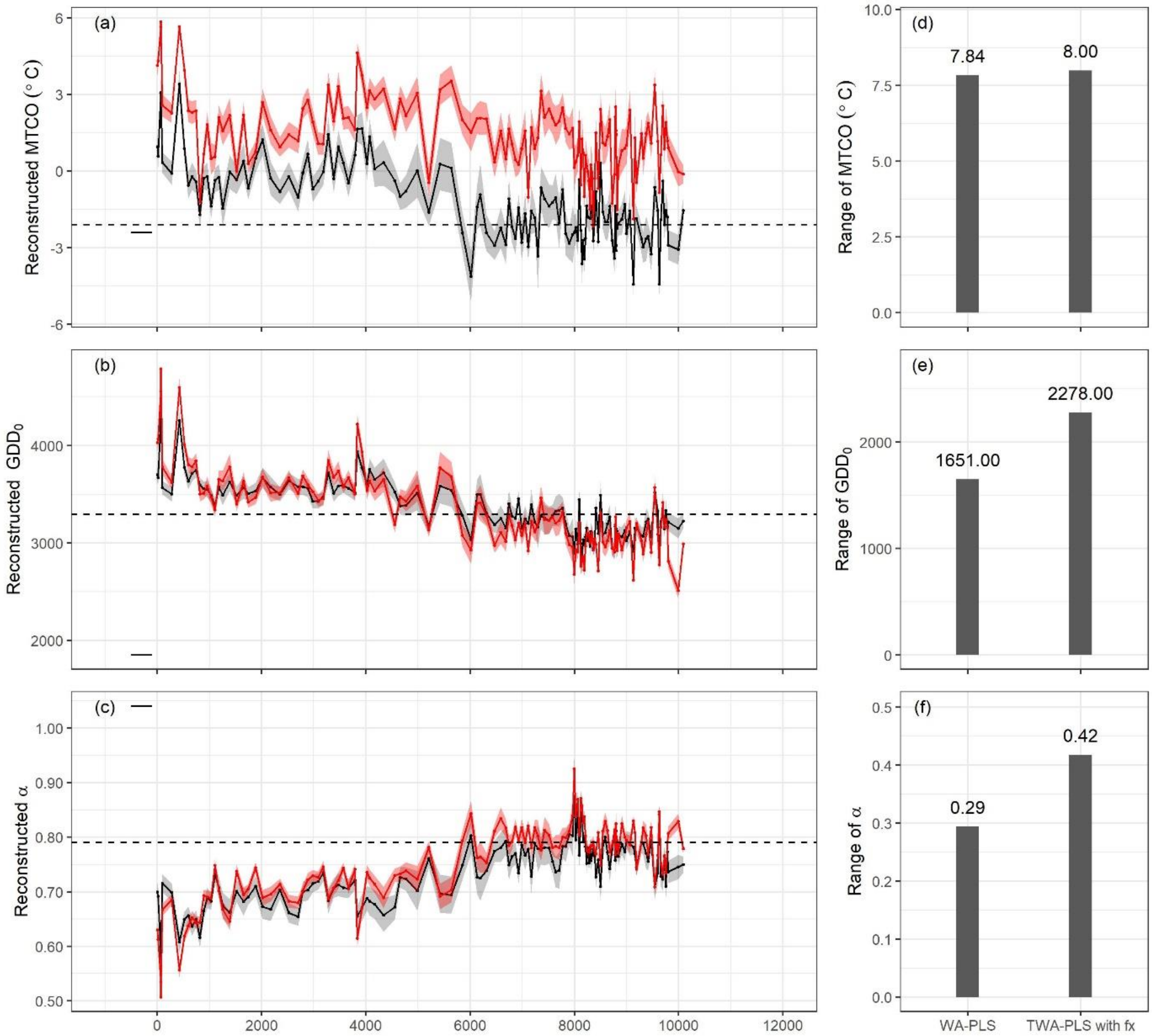

Figure 6. Comparison of downcore reconstructions of (a) MTCO, (b) GDD_0 and (c) α, at Estanya made using WA-PLS (black line) and TWA-PLS with *fx* correction (red line). The shades are 95% confidence intervals (reconstructions plus or minus 1.96 times their bootstrap estimates of sample-specific errors) of reconstructions using WA-PLS (black shade) and TWA-PLS with *fx* correction (red shade). The bar graphs show the range (maximum minus minimum) of reconstructed values over the Holocene for (d) MTCO, (e) GDD_0, (f) α using the two methods. Lines at 0 cal yr BP show the observed modern climate values at the site. Dashed horizontal lines show the estimate of the central range of the climate in the training data set.

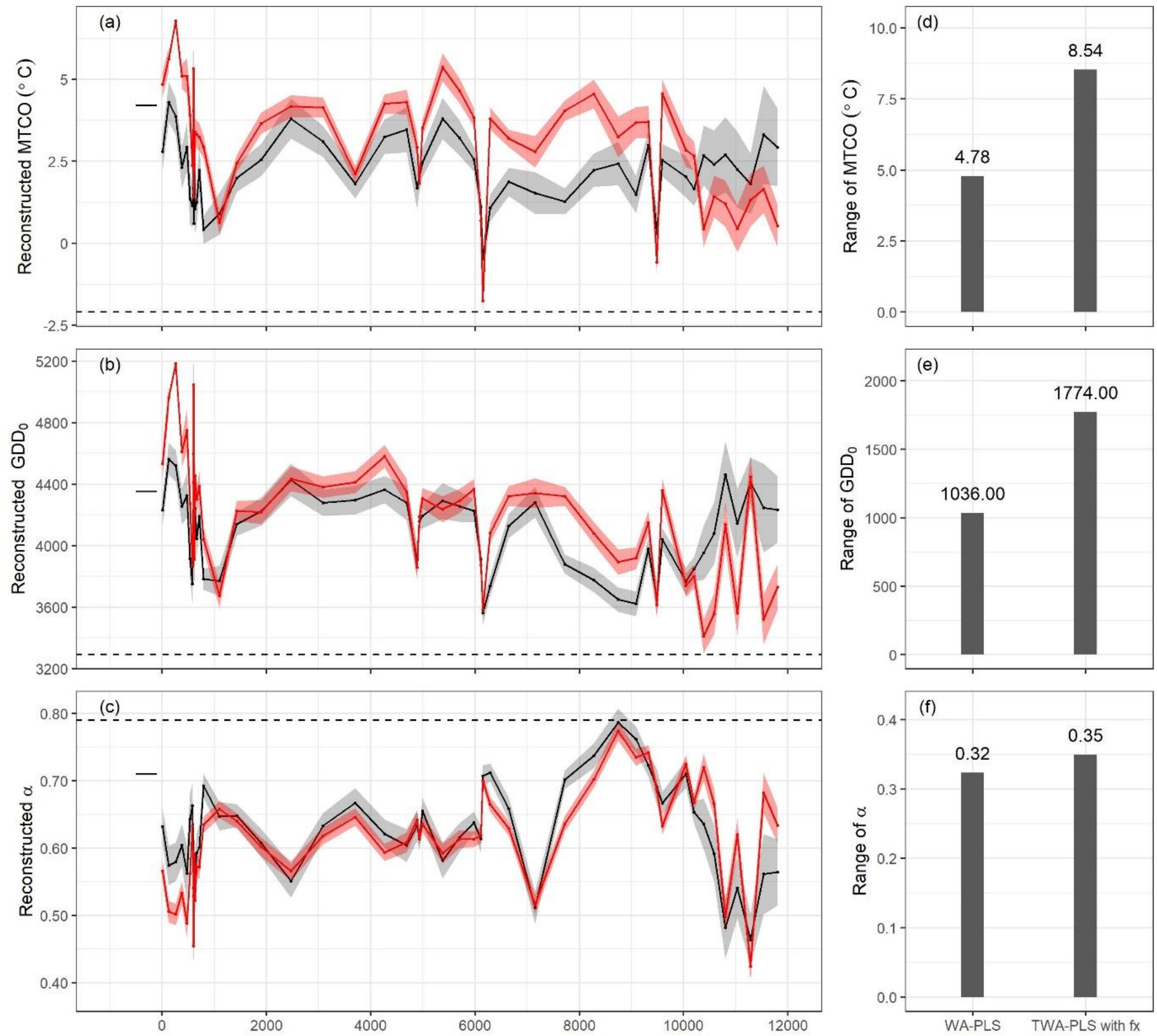

Table 1. Algorithms for WA-PLS and TWA-PLS, with and without *fx* correction. Here "←" is used instead of "=" to show the assigning of values. rlm means robust fitting of linear models [ref 30].

	WA-PLS	TWA-PLS
Step 0. Centre the environmental variable	$x_i \leftarrow x_i - \frac{\sum_{i=1}^{n}(\sum_{k=1}^{m} y_{ik})\, x_i}{\sum_{i=1}^{n}(\sum_{k=1}^{m} y_{ik})}$	$x_i \leftarrow x_i - \frac{\sum_{i=1}^{n}(\sum_{k=1}^{m} y_{ik}) x_i}{\sum_{i=1}^{n}(\sum_{k=1}^{m} y_{ik})}$
Step 1. Take the centred environmental variable as initial site scores	$r_i \leftarrow x_i$	$r_i \leftarrow x_i$
Step 2. Calculate new species scores	$u_k \leftarrow \frac{\sum_{i=1}^{n} y_{ik} r_i}{\sum_{i=1}^{n} y_{ik}}$	$u_k \leftarrow \frac{\sum_{i=1}^{n} y_{ik} r_i}{\sum_{i=1}^{n} y_{ik}}$ $t_k \leftarrow \sqrt{\frac{\sum_{i=1}^{n} y_{ik}(r_i - u_k)^2}{(1 - 1/N_{2k})\ \sum_{i=1}^{n} y_{ik}}}$ where $N_{2k} \leftarrow \frac{1}{\sum_{i=1}^{n}\left(\frac{y_{ik}}{\sum_{i'=1}^{n} y_{i'k}}\right)^2}$
Step 3. Calculate new site scores	$r_i \leftarrow \frac{\sum_{k=1}^{m} y_{ik} u_k}{\sum_{k=1}^{m} y_{ik}}$	$r_i \leftarrow \frac{\sum_{k=1}^{m} \frac{y_{ik} u_k}{t_k^2}}{\sum_{k=1}^{m} \frac{y_{ik}}{t_k^2}}$
Step 4. For the first axis go to Step 5. For second and higher components, make the new site scores uncorrelated with the previous components by orthogonalization	Using the orthogonalization procedure in Table 5.2.b in ref 29.	Using the orthogonalization procedure in Table 5.2.b in ref 29.
Step 5. Standardize the new site scores	Using the standardization procedure in Table 5.2.c in ref 29.	Using the standardization procedure in Table 5.2.c in ref 29.
Step 6. Take the standardized score as the new component	$comp_{pls} \leftarrow r_i$	$comp_{pls} \leftarrow r_i$
Step 7. Regress the environmental variable on the components obtained so far using weights and take the fitted values as current estimates. Go to Step 2 with the residuals of the regression as the new site scores.	If without *fx* correction, $rlm(x_i \sim comp_1 + \cdots + comp_{pls}, weights = \frac{\sum_{k=1}^{m} y_{ik}}{\sum_{i=1}^{n}(\sum_{k=1}^{m} y_{ik})})$ If with *fx* correction, $rlm(x_i \sim comp_1 + \cdots + comp_{pls}, weights = \frac{1}{f_{x_i}^2})$	If without *fx* correction, $rlm(x_i \sim comp_1 + \cdots + comp_{pls}, weights = \frac{\sum_{k=1}^{m} y_{ik}}{\sum_{i=1}^{n}(\sum_{k=1}^{m} y_{ik})})$ If with *fx* correction, $rlm(x_i \sim comp_1 + \cdots + comp_{pls}, weights = \frac{1}{f_{x_i}^2})$

Table 2. Leave-out cross-validation (with geographically and climatically close sites removed) fitness of WA-PLS and TWA-PLS methods, with and without *fx* correction, for mean temperature of the coldest month (MTCO), growing degree days above a baseline of 0 °C (GDD_0) and plant-available moisture (α), showing results for all the components. For WA-PLS with *fx* correction, only 4 components can be extracted for GDD_0 and α. RMSEP is the root-mean-square error of prediction. ΔRMSEP is the percent change of RMSEP using the current number of components than using one component less. *p* assesses whether using the current number of components is significantly different from using one component less, which is used to choose the last significant number of components (indicated in bold) to avoid over-fitting. The degree of overall compression is assessed by doing linear regression to the cross-validation result and the climate variable, b_0, b_1, b_0.se, b_1.se are the intercept, slope, standard error of the intercept, standard error of the slope, respectively. The closer the slope (b_1) is to 1, the less the overall compression is.

	Method	ncomp	R^2	Avg. Bias	Max. Bias	Min. Bias	RMSEP	ΔRMSEP	*p*	b_0	b_1	b_0.se	b_1.se
MTCO	WA-PLS	1	0.61	0.24	33.31	0.00	5.43	− 37.28	0.001	− 0.82	0.59	0.05	0.01
		2	0.65	0.12	31.93	0.00	5.12	− 5.72	0.001	− 0.76	0.66	0.05	0.01
		3	**0.66**	**0.17**	**30.52**	**0.00**	**5.05**	**− 1.49**	**0.002**	**− 0.66**	**0.68**	**0.05**	**0.01**
		4	0.66	0.18	42.56	0.00	5.06	0.22	0.669	− 0.64	0.68	0.05	0.01
		5	0.65	0.17	59.92	0.00	5.12	1.25	0.956	− 0.64	0.69	0.06	0.01
	TWA-PLS	1	0.66	0.30	33.64	0.00	5.07	− 41.48	0.001	− 0.66	0.62	0.05	0.01
		2	0.71	0.19	32.50	0.00	4.65	− 8.27	0.001	− 0.57	0.70	0.05	0.01
		3	**0.72**	**0.16**	**31.43**	**0.00**	**4.58**	**− 1.44**	**0.001**	**− 0.51**	**0.74**	**0.05**	**0.01**
		4	0.72	0.15	37.48	0.00	4.57	− 0.26	0.308	− 0.50	0.75	0.05	0.01
		5	0.72	0.16	58.07	0.00	4.61	0.86	0.754	− 0.49	0.75	0.05	0.01
	WA-PLS with *fx* correction	1	0.61	− 1.04	30.54	0.00	5.67	− 34.49	0.001	− 1.73	0.73	0.07	0.01
		2	0.65	− 0.83	35.67	0.00	5.32	− 6.31	0.001	− 1.43	0.76	0.06	0.01
		3	**0.66**	**− 0.65**	**33.70**	**0.00**	**5.20**	**− 2.24**	**0.001**	**− 1.23**	**0.77**	**0.06**	**0.01**
		4	0.66	− 0.74	44.52	0.00	5.20	0.09	0.537	− 1.33	0.77	0.06	0.01
		5	0.66	− 0.78	58.51	0.00	5.28	1.47	0.998	− 1.36	0.77	0.06	0.01
	TWA-PLS with *fx* correction	1	0.66	− 0.86	31.17	0.00	5.21	− 39.82	0.001	− 1.48	0.76	0.06	0.01
		2	0.72	− 0.52	36.61	0.00	4.70	− 9.80	0.001	− 1.03	0.80	0.06	0.01
		3	0.73	− 0.47	41.14	0.00	4.63	− 1.62	0.001	− 0.93	0.82	0.06	0.01
		4	**0.73**	**− 0.51**	**44.79**	**0.00**	**4.58**	**− 1.01**	**0.002**	**− 0.97**	**0.82**	**0.06**	**0.01**
		5	0.73	− 0.41	58.36	0.00	4.62	0.86	0.732	− 0.85	0.83	0.06	0.01
GDD_0	WA-PLS	1	0.59	− 21.47	4507.27	0.17	1000.20	− 35.91	0.001	1355.83	0.59	22.90	0.01
		2	**0.63**	**− 38.43**	**5077.66**	**0.12**	**950.25**	**− 4.99**	**0.001**	**1151.68**	**0.65**	**22.98**	**0.01**
		3	0.64	− 35.20	6518.62	0.03	941.76	− 0.89	0.040	1084.15	0.67	23.34	0.01
		4	0.63	− 35.09	9593.39	0.14	947.84	0.64	0.771	1066.94	0.67	23.71	0.01
		5	0.62	− 34.13	13849.39	0.03	964.51	1.76	0.976	1054.64	0.68	24.41	0.01
	TWA-PLS	1	0.66	− 19.13	4542.77	0.12	912.16	− 41.55	0.001	1144.48	0.66	21.87	0.01
		2	**0.69**	**− 17.48**	**4446.54**	**0.09**	**862.91**	**− 5.40**	**0.001**	**980.65**	**0.70**	**21.64**	**0.01**
		3	0.70	− 24.65	7094.83	0.18	857.52	− 0.62	0.132	919.11	0.72	21.91	0.01
		4	0.69	− 16.52	11556.79	0.16	865.31	0.91	0.740	892.03	0.73	22.46	0.01
		5	0.68	− 16.32	16283.18	0.05	885.25	2.30	0.907	881.10	0.73	23.21	0.01

	WA-PLS with *fx* correction	1	0.59	81.17	4540.86	0.07	1055.63	− 32.35	0.001	920.42	0.75	29.04	0.01
		2	**0.63**	**72.23**	**5401.87**	**0.10**	**998.53**	**− 5.41**	**0.001**	**814.61**	**0.78**	**27.75**	**0.01**
		3	0.63	42.61	9133.98	0.29	990.12	− 0.84	0.179	763.45	0.79	27.65	0.01
		4	0.63	39.35	11557.30	0.32	997.10	0.71	0.845	743.37	0.79	27.95	0.01
	TWA-PLS with *fx* correction	1	0.66	68.45	4534.20	0.08	951.90	− 39.00	0.001	753.05	0.80	26.57	0.01
		2	0.70	41.87	4700.48	0.27	882.35	− 7.31	0.001	649.66	0.82	24.80	0.01
		3	**0.71**	**21.37**	**7943.48**	**0.23**	**868.85**	**− 1.53**	**0.006**	**594.64**	**0.83**	**24.55**	**0.01**
		4	0.71	34.18	9748.44	0.19	869.25	0.05	0.564	597.03	0.83	24.60	0.01
		5	0.71	38.35	10978.72	0.11	872.40	0.36	0.770	605.52	0.83	24.67	0.01
α	WA-PLS	1	0.59	0.001	0.724	0.000	0.174	− 36.18	0.001	0.30	0.61	0.01	0.01
		2	0.63	0.001	0.798	0.000	0.166	− 4.54	0.001	0.27	0.66	0.01	0.01
		3	**0.64**	**0.001**	**0.780**	**0.000**	**0.165**	**− 0.79**	**0.005**	**0.26**	**0.67**	**0.01**	**0.01**
		4	0.64	0.001	0.792	0.000	0.165	− 0.14	0.207	0.25	0.67	0.01	0.01
		5	0.64	0.001	0.796	0.000	0.165	0.23	0.963	0.25	0.67	0.01	0.01
	TWA-PLS	1	0.63	0.002	0.746	0.000	0.166	− 39.12	0.001	0.28	0.64	0.00	0.01
		2	0.68	− 0.001	0.841	0.000	0.155	− 6.54	0.001	0.23	0.70	0.00	0.01
		3	0.68	0.001	0.772	0.000	0.154	− 1.17	0.001	0.22	0.71	0.00	0.01
		4	**0.69**	**0.000**	**0.789**	**0.000**	**0.153**	**− 0.50**	**0.007**	**0.21**	**0.72**	**0.00**	**0.01**
		5	0.69	0.001	0.793	0.000	0.153	0.01	0.524	0.21	0.72	0.00	0.01
	WA-PLS with *fx* correction	1	0.59	− 0.021	0.855	0.000	0.183	− 33.09	0.001	0.18	0.74	0.01	0.01
		2	**0.63**	**− 0.019**	**0.889**	**0.000**	**0.172**	**− 6.11**	**0.001**	**0.19**	**0.73**	**0.01**	**0.01**
		3	0.63	− 0.022	0.803	0.000	0.171	− 0.31	0.214	0.17	0.75	0.01	0.01
		4	0.63	− 0.020	0.867	0.000	0.172	0.43	0.990	0.16	0.76	0.01	0.01
	TWA-PLS with *fx* correction	1	0.63	− 0.020	0.773	0.000	0.175	− 36.03	0.001	0.15	0.77	0.01	0.01
		2	0.68	− 0.012	0.902	0.000	0.158	− 9.73	0.001	0.15	0.79	0.01	0.01
		3	**0.69**	**− 0.011**	**0.820**	**0.000**	**0.156**	**− 1.29**	**0.001**	**0.15**	**0.79**	**0.01**	**0.01**
		4	0.69	− 0.010	0.787	0.000	0.156	0.26	0.881	0.14	0.81	0.01	0.01
		5	0.69	− 0.010	0.787	0.000	0.156	0.10	1.000	0.14	0.81	0.01	0.01

Table 3. Representation of taxa in the modern training data set. The table shows the mean number of sites at which the taxa are represented (present) or not represented (absent) for all the taxa in the data set and for those taxa with a relatively narrow climate niche (width of the niche defined for MTCO, GDD_0 and α separately). There are 6458 sites in total.

	Present	Absent
All taxa	832	5626
Taxa with tolerance of MTCO < 3	94	6364
Taxa with tolerance of GDD_0 < 500	111	6347
Taxa with tolerance of α < 0.1	71	6387