https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions

Can machine learning algorithms improve upon classical palaeoenvironmental reconstruction models?

Peng Sun[1], Philip. B. Holden[2], and H. John B. Birks[3,4]

[1]Institute of Environmental Sciences (CML), Leiden University, 2333 CC Leiden, the Netherlands
[2]Environment, Earth and Ecosystem Sciences, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[3]Department of Biological Sciences and Bjerknes Centre for Climate Research, University of Bergen, PO Box 7803, Bergen N-5020, Norway [4]Environmental Change Research Centre, University College London, London WC1 6BT, UK

Correspondence to: Philip B. Holden (philip.holden@open.ac.uk)

Abstract. Classical palaeoenvironmental reconstruction models often incorporate biological ideas and commonly assume that the taxa comprising a fossil assemblage exhibit unimodal response functions of the environmental variable of interest. In contrast, machine learning approaches do not rely upon any biological assumptions, but instead need training with large data-sets to extract some understanding of the relationships between biological assemblages and their environment. We have developed a two-layered machine learning reconstruction model MEMLM (Multi Ensemble Machine Learning Model). The first layer applies three different ensemble machine learning models of random forests, extra random trees and lightGBM, trained on the modern taxon assemblage and associated environmental data to make reconstructions based on the three different models, while the second layer uses multiple linear regression to integrate these three reconstructions into a consensus reconstruction. We consider three versions of the model: 1) A standard version of MEMLM, which uses only taxon abundance data, 2) MEMLMe, which uses embedded assemblage information, using a natural language processing model (GLOVE) to detect associations between taxa across the training data-set and 3) MEMLMc which incorporates both taxon abundance and assemblage data. We train these MEMLM model variants with three high quality diatom and pollen training sets and compare their reconstruction performance with three weighted averaging (WA) approaches of WA-Cla (classical deshrinking), WA-Inv (inverse deshrinking) and WA-PLS (partial least squares). In general, the MEMLM approaches, even when trained on only embedded assemblage data, perform substantially better than the WA approaches under cross-validation in the larger data-sets. However, when applied to fossil data, MEMLM and WA approaches sometimes generate qualitatively different palaeoenvironmental reconstructions. We applied a statistical significance test to all the reconstructions. This successfully identified each incidence where the reconstruction is not robust with respect to the model choice. We find that machine learning approaches can outperform classical approaches, but can sometimes catastrophically fail, despite showing high performance under cross-validation, likely indicating problems when extrapolation occurs. We find that the classical approaches are generally more robust, although they can also generate reconstructions which have modest statistical significance, and therefore may be unreliable. We conclude that cross-validation is not a sufficient measure of transfer-function performance, and we recommend that the results of statistical significance tests are provided alongside the down-core reconstructions based on fossil assemblages.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions

1 INTRODUCTION

The distribution and abundance of taxa are interrelated with the environment (Ovaskainen et al., 2017). By using the concept of *space instead time*, the palaeoenvironment can be reconstructed by applying modern taxon-environment relationships to the fossil record (e.g. Battarbee et al., 2005; Cleator et al., 2020; Turner et al., 2020).

With the development of palaeoecological research, large training data-sets for environmental reconstruction have emerged in recent years. (e.g. Harrison 2019; Bush et al., 2021). Data assimilation has long been a focus of Earth science and ecology, and the integration of larger data-sets will provide more comprehensive training information (e.g. Christin et al., 2019; Houssaye et al., 2019; Bush et al., 2020). For large data-sets, machine learning methods have strong advantages and may be appropriate to extract the non-linear relationships between taxon compositional information and the environment, and to integrate a variety of sources of data (e,g, Helama et al., 2009; Aguirre-Gutierrez et al., 2021; Wei et al., 2021b).

In recent years, machine learning has been applied to a wide range of applications in palaeoecology (Hais et al., 2015; Jordan et al., 2016). Wei et al., (2021b) reconstructed palaeoclimate using five different machine learning methods based on digital leaf physiognomic data and integrated the predictions by averaging. Hais et al., (2015) predicted the Pleistocene biota distributions in palaeoclimate using machine learning. Huang et al., (2020) used one series of palaeoclimate sequences to predict the climate in another period. These studies show that machine learning has strong versatility and effectiveness, and suggest it should be more widely applied.

However, machine learning approaches do not make any biological assumptions, which may weaken their performance relative to mathematically simpler classical approaches that do. For instance, weighted averaging (WA) approaches apply the simple but informative assumption that taxa have a unimodal response to the environmental variable of interest (ter Braak and Barendregt, 1986). The absence of any such prior understanding is likely to place additional demands on the minimum adequate size of a modern training set. Moreover, it may weaken the ability of machine learning to operate under extrapolation, critically important when applying any reconstruction approach to past taxon assemblages that lack modern analogues. To address these questions, we have developed the Multi Ensemble Machine Learning Model (MEMLM) to apply in a systematic comparison with classical WA reconstruction approaches.

The benefit of machine learning is that it has strong data mining and information extraction ability. An associated problem, however, is that when a sample size is limited, machine learning is more likely to learn the noise component and generate prediction errors due to over-fitting (Yeom et al., 2018; Syam and Kaul 2021). This suggests that an ensemble learning method, which integrates models with potentially different biases, may reduce over-fitting errors (Legendre et al., 1997) and improve the prediction performance (Wei et al., 2021b). This is the motivation for the ensemble learning approach we present, namely the Multi Ensemble Machine Learning Model (MEMLM).

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023


Climate of the Past Discussions

Classical studies have integrated different reconstruction approaches by calculating the mean of their predictions (Norberg et al., 2019). An arithmetic mean does not attribute weights to each model, even though the models may have different advantages in different applications (Schulte and Hinckley 1985; Zhou 2012). Similarly, most classical models do not consider different weights for different taxa, which may reduce their prediction potential and smooth the reconstruction (e.g. Brooks and Birks 2001; Heiri et al., 2003; Battarbee et al., 2005; Wei et al., 2021a). Tolerance downweighted WA-PLS (TWA-PLS) makes it possible to assign different weights to each taxon in reconstructing the environment (Liu et al., 2020), while Bayesian approaches such as BUMPER (Holden et al., 2017) are built on classical assumptions and are highly constrained by taxa with low environmental tolerances, especially when characterised with high confidence.

In MEMLM, we apply both taxon weights and model weights. The first calculation layer applies three different machine learning ensemble models of random forests, extra random trees and lightGBM, trained on modern taxon assemblage and environmental data. In these ensemble models, each taxon has a different predicted contribution which is used to weight its contribution to the ensemble. The three reconstructions are then integrated into a consensus reconstruction using a weak learning algorithm which weights each model according to its predictive power under cross-validation.

We develop three versions of MEMLM; the standard version which only considers raw taxon abundance data; MEMLMe includes encoded assemblage information; and MEMLMc includes both. The motivation for the more complex versions is to explore whether considering known associations between taxa can improve the palaeoenvironmental reconstructions. For this, we use the natural language processing (NLP) model GLOVE (Pennington et al., 2014), which calculates the relationships between co-occurring words in the same sentence. GLOVE is a form of dimension reduction which assigns vectors (also called embedding) to each word according to the word connection relationships, so that each sentence can be represented as a superposition of the word embeddings within that sentence. In environmental assemblages, there are analogous co-occurrence relationships between taxa which we hypothesise convey information on their ecological functioning. We therefore use GLOVE to generate the embedding vectors of different taxa in different samples based on assemblage information and then to integrate the embeddings within each sample to represent the assemblage.

We apply MEMLM to high quality pollen and diatom training sets to generate down-core reconstructions. We calculate training set cross-validation metrics and we quantify the statistical significance and robustness of the core reconstructions. We compare these performance metrics with those of classical WA approaches to evaluate whether, and under what circumstances, machine learning approaches might be able to outperform classical reconstruction approaches.

2 MATERIALS AND METHODS

We apply MEMLM to high quality pollen and diatom training sets to generate down-core reconstructions. We calculate training set cross-validation metrics and we quantify the statistical significance and robustness of the core reconstructions. We compare these performance metrics with those of classical WA approaches to evaluate whether, and under what circumstances, machine learning approaches might be able to outperform classical reconstruction approaches.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions
Open Access
EGU

2.1. **MEMLM**

MEMLM combines a series of modules **(Figure 1)**. In this section, we introduce the functions of each module and the data processing routes.

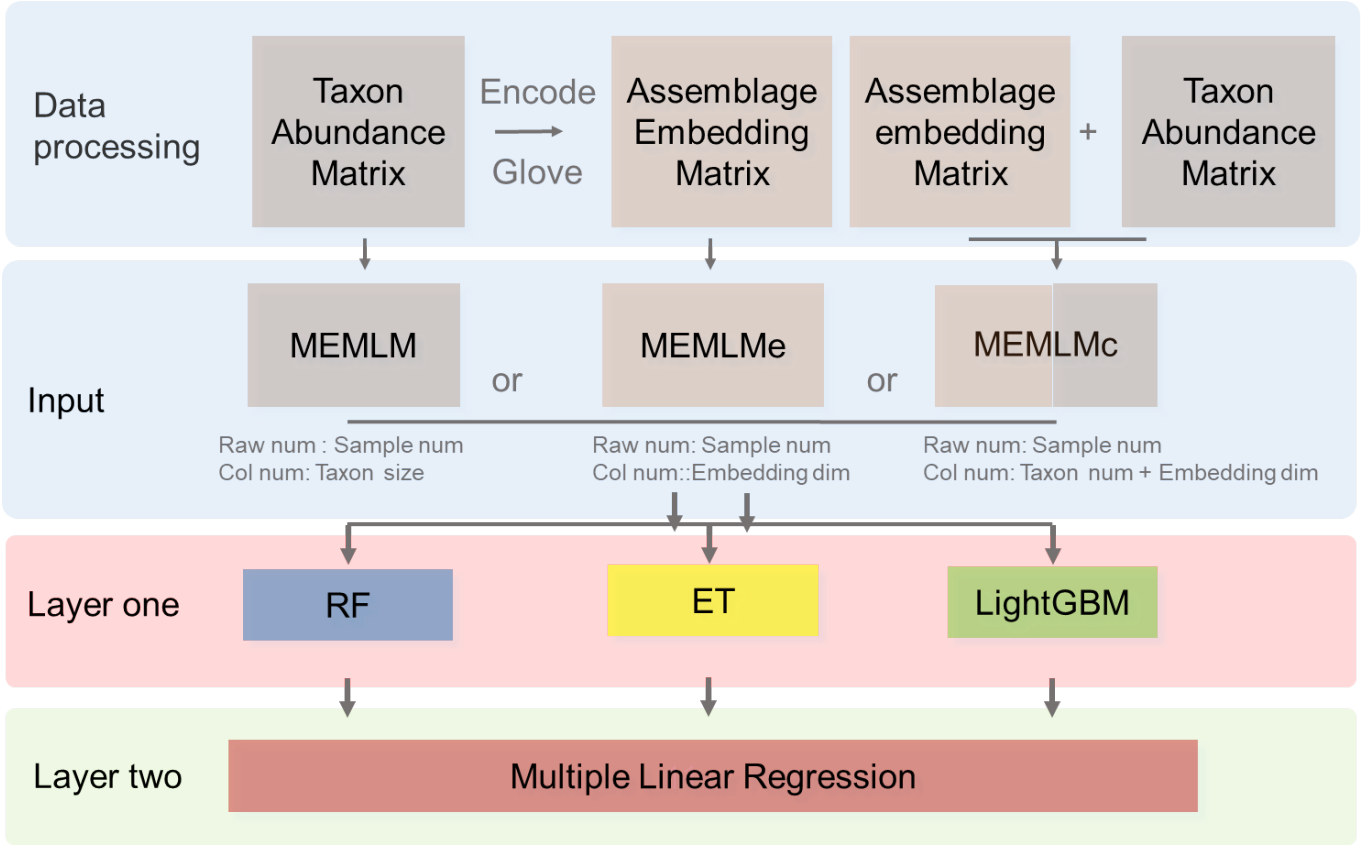


Figure 1: Multi Ensemble Machine Learning Model (MEMLM) model framework. MEMLM has a modular building block architecture so that components can be easily changed. Raw num and Col num are the number of rows and columns in input matrix; dim is the number of dimensions.

2.1.1. First layer

The input data comprise environmental data together with either the taxon abundance matrix, the assemblage embedding matrix, or both matrices (see section 2.2.3 for a description of the embedding algorithm to develop the assemblage matrix). We apply three ensemble machine learning models to derive the mapping between taxon composition information and environmental factors:

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023


Climate of the Past Discussions

(1) Random forests (RF) is an ensemble machine learning model composed of multiple decision trees. The overall model framework is determined based on the predictive power of each decision tree applied to the training data-set under bootstrapping. Individual decision trees with better predictive performance are allocated higher weights, and the 'forest' integrates the weighted result from each tree (Liaw and Wiener 2002).

(2) Extra Random Tree (ET) is similar to RF, except that it uses the entire data-set rather than a bootstrapped subset (Geurts et al., 2006).

(3) LightGBM is based on the Gradient Boosting Decision Tree. This also integrates decision trees, but LightGBM differs by applying 'gradient boosting' to add new trees, building each new model on the residuals of the previous model to improve the prediction. It has the ability to merge sparse data sets to increase computational efficiency (Friedman 2001; Ke et al., 2017).

2.1.2 Second layer (consensus reconstruction)

It is possible to improve prediction performance by integrating the prediction of multiple models into a consensus reconstruction (Yeom et al., 2017; Syam and Kaul 2021). Averaging is widely used to integrate the output prediction of multiple models. However, the integration weight of each model is the same under averaging. MEMLM applies multiple linear regression to allocate an integration weight to each model rather than attaching each model with the same weight. The consensus reconstruction is derived as follows. First, the three upstream models are applied to reconstruct the training data-set under five-fold cross-validation. We then build a multiple linear regression model to fit the reconstructed values to the actual value in the training set. This approach is designed to avoid the risk of over-fitting while reducing the impact of low-performance models on the consensus reconstruction. Exploratory analysis applied to the NIMBIOS data-set, building models for each of 18 environment attributes, demonstrated that the multiple linear regression approach reduced the root mean squared error of prediction (RMSEP) relative to the individual reconstructions by an average of 8% (Table A1). A consensus reconstruction based on the mean of the three ensemble approaches also improved predictive power but reduced the RMSEP errors relative to the individual reconstructions by an average of 5%.

2.1.3 Embedding

The GLOVE algorithm (Pennington et al., 2014) is a very widely used linguistic dimensional reduction approach. It uses co-occurrences of words in phrases to characterise numerically their meaning. Words are represented as vectors in high dimensional space, where each dimension captures an aspect of meaning so that in this space words that have similar meanings are located near to each other. To illustrate, in word vector space, we would expect the difference vectors $\boldsymbol{queen - king}$ and $\boldsymbol{girl - boy}$ to be similar, as they both reflect only a change of gender, with other dimensions of meaning (species, age, social status etc) constant. Embedding reduces the dimensionality of a vocabulary from tens of thousands of words to hundreds of meaning dimensions, known as features.

In ecology, co-existence among taxa can reflect characteristics of the environment (Legendre et al., 1997; Ovaskainen et al., 2017). We hypothesise that taxa within an assemblage have relationships that are analogous to words within a phrase, so that

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

in the feature space of ecological 'meaning' the vectorial representation of a taxon describes its ecological function. We apply GLOVE to ecological assemblages. Instead of analysing co-occurrences of words within phrases, we analyse co-occurrences of taxa within assemblages. The objective is to extract ecological information by associating taxa with their ecosystem functioning.

The GLOVE algorithm is fully detailed in Pennington et al (2014), and here we introduce the underlying philosophy and illustrate it in the context of ecological functioning. Consider P_{ij} the probability that taxon j appears in the same assemblage as taxon i:

$$P_{ij} = P(j|i) = X_{ij}/X_i \quad (1)$$

where X_{ij} is the number of assemblages which contain both taxa i and j, and X_i is the number of assemblages containing taxon i. This probability does not necessarily indicate the strength of the relationship. Consider, for instance, that a high value may simply reflect that taxon j is common and therefore provides little information about the environment.

To determine associative relationships, GLOVE considers the ratio P_{ik}/P_{jk} where taxon k is some probe taxon used to differentiate the ecological functioning of i and j. If taxon k has a strong association with taxon i but not with taxon j then $P_{ik}/P_{jk} \gg 1$. However, if all three taxa are either commonly found together or have no relationship (i.e. low but random co-occurrence) between each other, $P_{ik}/P_{jk} \sim 1$, indicating that taxon k provides very little information to help distinguish the ecological functions of i and j. The value of P_{ik}/P_{jk} can therefore inform us about the direction of difference vector $i - j$. GLOVE is trained on assemblages to map taxa onto vectors in feature space, so that the assemblages can be described as linear combinations of the features. For application to MEMLMc, the feature matrices are provided together with the raw taxon count data to provide richer training data for the ensemble learning algorithms.

2.2. Assemblage data

For model training purposes we apply two large pollen data-sets, SMPDSV1 (Harrison, 2019) and NIMBIOS (Bush et al., 2020), and the diatom SWAP data-set (Stevenson et al., 1991). In order to demonstrate the palaeoenvironment reconstructions of each model, we apply i) SWAP to reconstruct lake-water pH from diatoms in a core from The Round Loch of Glenhead (RLGH) (Allott et al., 1992, Jones et al., 1989), ii) SMPDSV1 to reconstruct mean temperature of the coldest month (MTCO) from pollen in the Villarquemado core (Harrison, 2019), and iii) NIMBIOS to reconstruct the mean average temperature (MAT) from pollen in the Consuelo (Urrego et al., 2010) and Llaviucu (Kannan et al., 1983, Colinvaux et al., 1988) cores.

2.2.1. Training data-sets

SWAP: The SWAP training set (Stevenson et al., 1991) was developed as part of an international scientific effort directed at establishing and understanding the impacts of acid rain on freshwaters. It includes relative abundance data for 277 diatom

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023


Climate of the Past Discussions

taxa from 167 modern samples with clear identification criteria standards (Birks et al., 1990). We apply these data to reconstruct lake-water pH.

The NIMBIOS data-set (Bush et al 2020), includes samples from 636 neotropical locations with various habitat types. There are 533 pollen types (some taxa can only be identified to family level), ranging from soil samples to mud-water interface samples from lakes. We use it to reconstruct mean annual temperature (MAT).

The SMPDSv1 data set was developed as an environmental calibration data-set to provide training data for palaeoclimate reconstructions (Harrison, 2019). SMPDSv1 contains the relative abundancies of the 247 most important pollen taxa in 6458 terrestrial samples from Europe, northern Africa, the Middle East, and Eurasia, compiled from multiple different published sources. We use it to reconstruct mean temperature of the coldest month (MTCO).

2.2.2. Core data-sets

We apply the SWAP training set to the RLGH and RLGH3 core data-sets. RLGH is a fossil diatom data-set from The Round Loch of Glenhead, Scotland, taken to explore anthropogenic acidification (Allott et al., 1992). The data-set includes the relative abundances of 41 diatom taxa in 20 samples which span the industrial era. RLGH3 was sampled to explore natural acidification driven by weathering and soil development during the Holocene (Jones et al., 1989). This data-set includes abundances for 225 diatom taxa in 101 samples.

We apply the NIMBIOS training set to the Consuelo and Llaviucu core data-sets. The core from Lake Consuelo, Bolivia, is an 8.8 m sediment sequence, which records the long-term evolution of cloud forest in response to environmental changes over the last 46,300 years (Urrego et al., 2010).

Lake Llaviucu is a temperature-sensitive lake in the Ecuadorian Andes (Kannan et al., 1983; Colinvaux et al., 1988). It lies behind a moraine in the system dated by Clapperton (1987) within the last glaciation (35,000 yr B.P.). At nearly 37 degrees S latitude, the lake is perched on the eastern face of the Cordillera Occidental and has been lifted 2,200 m since the glacial age. It shows the possibility of significant cooling of tropical latitude rain-forest near San Juan Bosco (Colinvaux et al., 1997).

We apply the SMPDSv1 training set to the Villarquemado core data-set (Wei et al 2021a), a pollen record from the western Mediterranean Basin spanning the interval from the last part of MIS-6 to the late Holocene. The fossil pollen data were assigned to the subset of pollen taxa recognized in the modern SMPDSv1 data-set. There are 104 taxa represented in the final taxon list based on the 361 core samples.

2.3. Performance and validation metrics

2.3.1. Model parameters

We build the GLOVE model under the PyTorch deep learning frame for efficient matrix computation and error gradient feedback (Paszke et al., 2019). In embedding training, we set the number of epochs (training loops) to 1000 and the number

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023


Climate of the Past Discussions

of embedding dimensions to 256. For first layer, we build an ensemble of 1000 decision trees number with parallel computing. MEMLM has an external interface so that these parameters can be easily changed for third party application.

We note that we originally developed the GLOVE analysis using the pre-packaged software 'glove-python' [https://github.com/maciejkula/glove-python] but we subsequently re-wrote the GLOVE algorithm from first principles. Cross-validation and down-core reconstructions from the two algorithms were not materially different and so the statistical significance testing, which is highly expensive computationally, requiring one month of parallel computing, was not repeated.

2.3.2. The prediction importance indicator

The MEMLM models are ensembles based on the results of multiple decision trees. Each time a decision tree forks, the algorithm will explore how to integrate each taxon's abundance values to have more predictive power. The algorithm works through an internal cross-validation analysis to determine whether each predictor reduces the prediction errors in each decision tree, and then summarizes the results across all decision trees. The approach ascribes an importance index to each taxon which is normalised to total 1 across all taxa and provides a measure of that taxon's predictive power. The ten most important taxa for each upstream model are detailed in Tables A2. These are used in the inference of taxon importance for climate reconstruction.

2.3.3. Cross-validation

The predictive powers of the MEMLM variants are compared with classical WA models (ter Braak and Barendregt 1986) and WA-PLS (ter Braak and Juggins 1993). We take RMSEP and R^2 score as evaluation indicators, using the scikit-learn package (Pedregosa et al., 2012). We use five-fold cross validation in this study.

For evaluation of the classical models we use the rioja package in R (Juggins 2017) with default settings. As WA-PLS performance is sensitive to the number of components; we accept a higher PLS component only if it exhibits a 5% improvement on the previous component (Birks 1998) and we present results for that component.

2.3.4. Statistical significance of reconstructions

While cross-validation is a useful measure of predictive power, which implicitly tests a model for over-fitting (Yates at al. 2023), it is likely to over-estimate predictive power in practice as fossil assemblages may lie outside the high dimensional space of the modern training assemblages, for instance by lacking close modern analogues. Telford and Birks (2011) developed an easily applied method for testing the robustness of a reconstruction of a specific site. The approach is to create an ensemble of transfer functions using the same biological assemblage as the training set, but with randomised values of the environmental variable. If the reconstructed variable is found to explain more of the variance than 95% of the random reconstructions, then the reconstruction is deemed to be statistically significant. We apply this approach with the palaeoSig package in R (Telford and Birks, 2015) to all core reconstructions as an indicator of their robustness.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023


Climate of the Past Discussions

2.4. Computing hardware

In this study, the computing CPU is Intel Core i7-4710MQ; the model is supported by the scikit-learn package (Pedregosa et al., 2012). MEMLM supports parallel computing: with more CPU cores, the computing time will decrease significantly.

3 RESULTS

3.1. Cross-validation

Table 1 compares the cross-validated RMSEP for the three training sets and the five reconstruction approaches (See Figure A1 for regression visualization of predicted values against observed values). WA-PLS was found to be the best performing classical approach in all three training sets as evaluated by RMSEP, but in each case it was outperformed by MEMLM, which reduced RMSEP by 6% (SWAP, 167 training samples, 277 taxa), 22% (NIMBIOS, 636 training samples, 533 taxa) and 50% (SMPDSv1, 6548 samples, 257 taxa). The additional learning power with increasing training-set size is evident.

MEMLMe is trained only on embedded assemblage data from GLOVE. The approach does not work well for the SWAP training set, but it significantly improves upon WA approaches when using the larger NIMBIOS and SMPDVs1 training sets, suggesting that when the training set is large enough, embedding is able to extract most of the predictive power of the assemblages. However, MEMLMe consistently under- performs relative to MEMLM and MEMLMc, and so we do not use it in the reconstructions.

We performed additional cross-validation tests on MEMLMe to confirm that the embedding approach does indeed encode useful information, noting that with an embedding dimension of 256 (comparable to the number of taxa in the training sets) we are not applying the approach under significant dimensional reduction. We applied a progressively increasing embedding dimension applied to an MEMLMe model of MAT using the 533-taxon NIMBIOS data-set (Figure A1b). This sensitivity demonstrates that only about 30-dimensions are required for MEMLMe to outperform WA-PLS (RMSEP 2.914°C), so that that dimension reduction by more than an order of magnitude retains sufficient information to build a useful model. Increasing the embedding dimension towards 256 unsurprisingly progressively improves RMSEP further by encoding additional assemblage information. Figure A2b illustrates the learning power of increased training, with RMSEP decreasing by around 0.4°C as the number of training epochs is increased from 40 to 1000.

MEMLMc uses both the taxon abundance and the embedding matrices. These additional data do not significantly affect the predictive performance relative to MEMLM under-cross validation, suggesting that conventional ensemble machine learning approaches are sufficient to encode adequately the assemblage information in training sets comprising a few hundred taxa. However, we retain this model for down-core reconstructions to explore whether the addition of embedding information can affect reconstructions in a way that is not captured by RMSEP.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions

		MEMLM	MEMLMe	MEMLMc	WA-Cla	WA-Inv	WA-PLS (best)
RMSEP							
SWAP	pH	**0.289**	0.376	0.294	0.307	0.313	0.299
NIMBIOS	MAT/ °C	2.203	2.193	**2.092**	3.194	3.577	2.914
SMPDSv1	MTCO/ °C	**2.360**	2.827	2.449	5.315	6.674	4.964
R^2 score							
SWAP	pH	**0.837**	0.670	0.831	0.795	0.837	0.822
NIMBIOS	MAT/ °C	0.841	0.846	**0.861**	0.585	0.711	0.683
SMPDSv1	MTCO/ °C	**0.920**	0.881	0.913	0.397	0.624	0.518

Table 1. Cross-validated RMSEP and R^2 score for the three training sets. MEMLM uses the abundance matrix. MEMLMe uses the assemblage embedding matrix. MEMLMc uses the spliced abundance and embedding matrices. WA-Cla is weighted averaging with a classical deshrinking regression, WA-Inv is weighted averaging with an inverse deshrinking regession (Birks et al. 1990). WA-PLS is the 'best' model (see 2.2.3), see Table A3 for other components. Bold highlights the model with the lowest RMSEP or highest R^2 score.

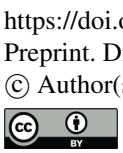
3.2. Environmental reconstructions and comparisons

For each core we compare the reconstructions from the models with lowest RMSEP, being the MEMLM and MEMLMc machine learning approaches and WA-PLS, the best classical approach (section 2.3.3), which is PLS component 1 for SWAP and PLS component 2 for NIMBIOS and SMPDSV1. In each reconstruction we additionally provide the statistical significance test results (Telford and Birks, 2011). A reconstruction is considered significant when that reconstruction explains more of the variance than 95% of 1000 randomised reconstructions, which apply the same training assemblage but with randomized environmental characteristics.

3.2.1. pH reconstructions from RLGH using the SWAP training set

MEMLM and WA-PLS1 show similar trends of acidification, with pH declining from around 5.2 at about 1870 to around 4.8 at about 1980. MEMLMc shows a similar trend but understates the degree of acidification relative to the other approaches. All three reconstructions are statistically significant, and with high explained variance, though WA-PLS1 explains more variance (58%) than MEMLM (46%) or MEMLMc (52%.). The variance explained by the first principal component of the fossil core

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023



Climate of the Past Discussions
Open Access
EGU

assemblages is 62%, indicating that the reconstructed pH explains most of the dominant part of the variance in the fossil diatom assemblages (Figure 2).

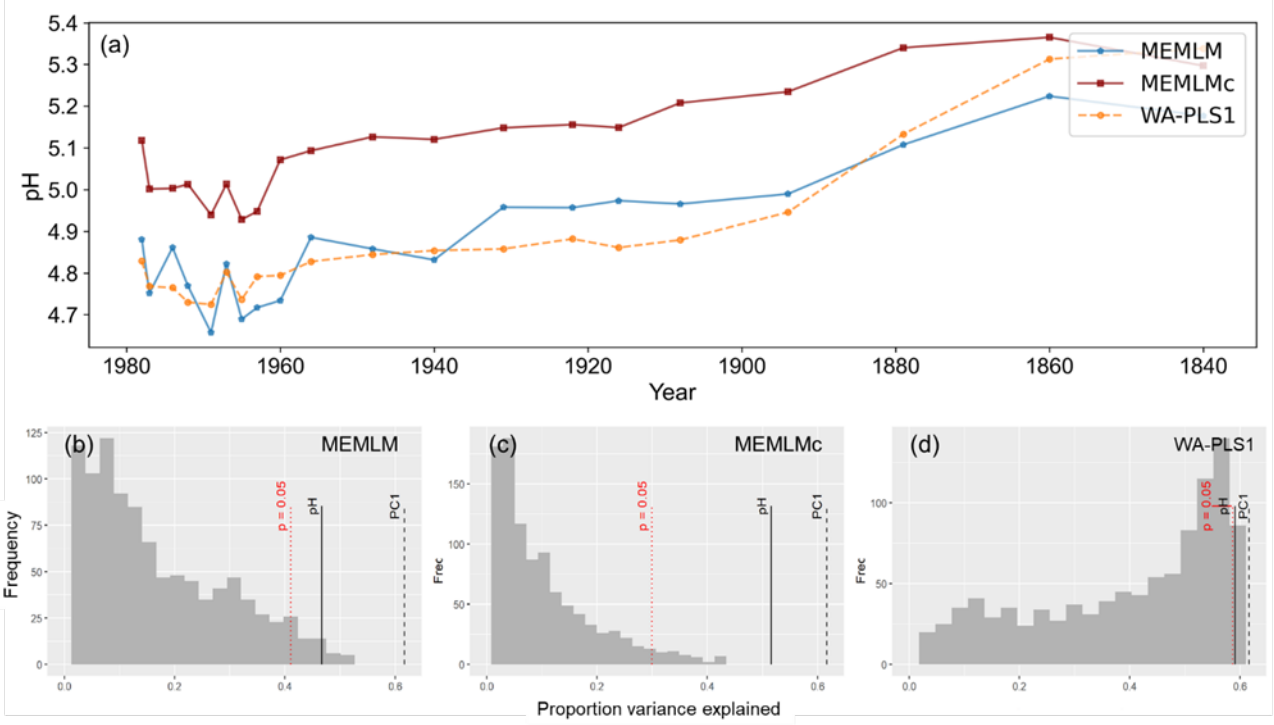


Figure 2: a) pH reconstruction for the RLGH core. b, c & d) statistical significance of MEMLM, MEMLMc and WA-PLS1 reconstructions, respectively.

3.2.2. pH reconstruction from RLGH3 using SWAP

All three methods provide reconstructions that show similar trends of lake-water pH, with gradual acidification in the early record from around 5.6 to 5.2 pH, attributed to the development of organic soils (Jones et al., 1989) and then a rapid post-industrial acidification from around 5.2. to 4.8 pH. The three reconstructions also exhibit similar variability, previously attributed to loss of tree cover and peat erosion (Jones et al., 1989), further suggesting reconstruction robustness. Moreover, all three reconstructions are statistically significant, explaining between 23% and 27% of the core variance, which compares to 32% variance explained by the first principal component of the fossil assemblages (Figure 3).

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

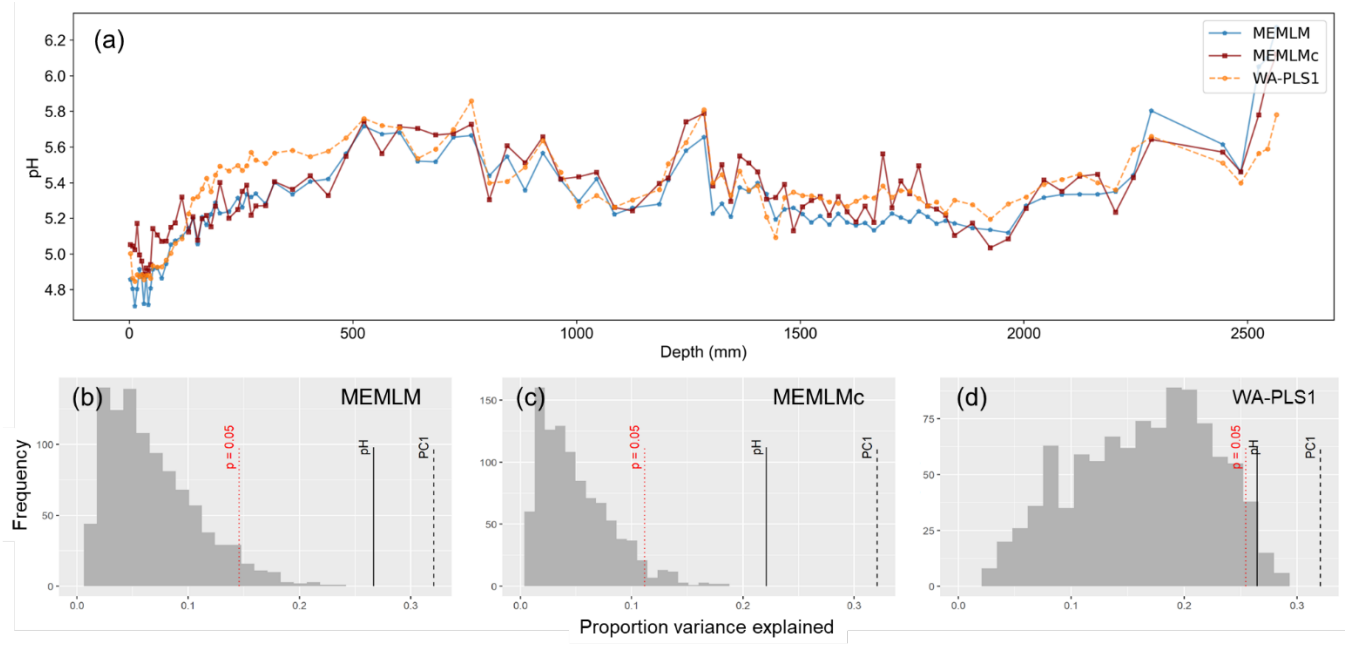


Figure 3: a) pH reconstruction for the RLGH3 core. b, c & d) statistical significance of MEMLM, MEMLMc and WA-PLS1 reconstructions, respectively.

3.2.3. MAT reconstruction from Consuelo using the NIMBIOS training set

All three methods display similar trends, most notably reconstructing about a 4°C warming from the Last Glacial Maximum at 21,000 BP to the start of the Holocene at 11,000 BP. The MEMLM approaches are more variable in general, although variability is largely synchronous between the three reconstruction approaches and may be associated with Dansgaard-Oeschger (D/O) events (Bond et al. 1993; Blunier &and Brook 2001). At 8000 BP, WA-PLS2 displays a 10°C cooling excursion which is not apparent in the MEMLM reconstructions. Although a cooling event at 8.2ka is well known, the cooling reconstructed by WA-PLS2 cooling is excessive. All three methods are statistically significant and explain core assemblage variance of between 27% and 29%, compared to 32% explained by the first principal component (Figure 4).

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

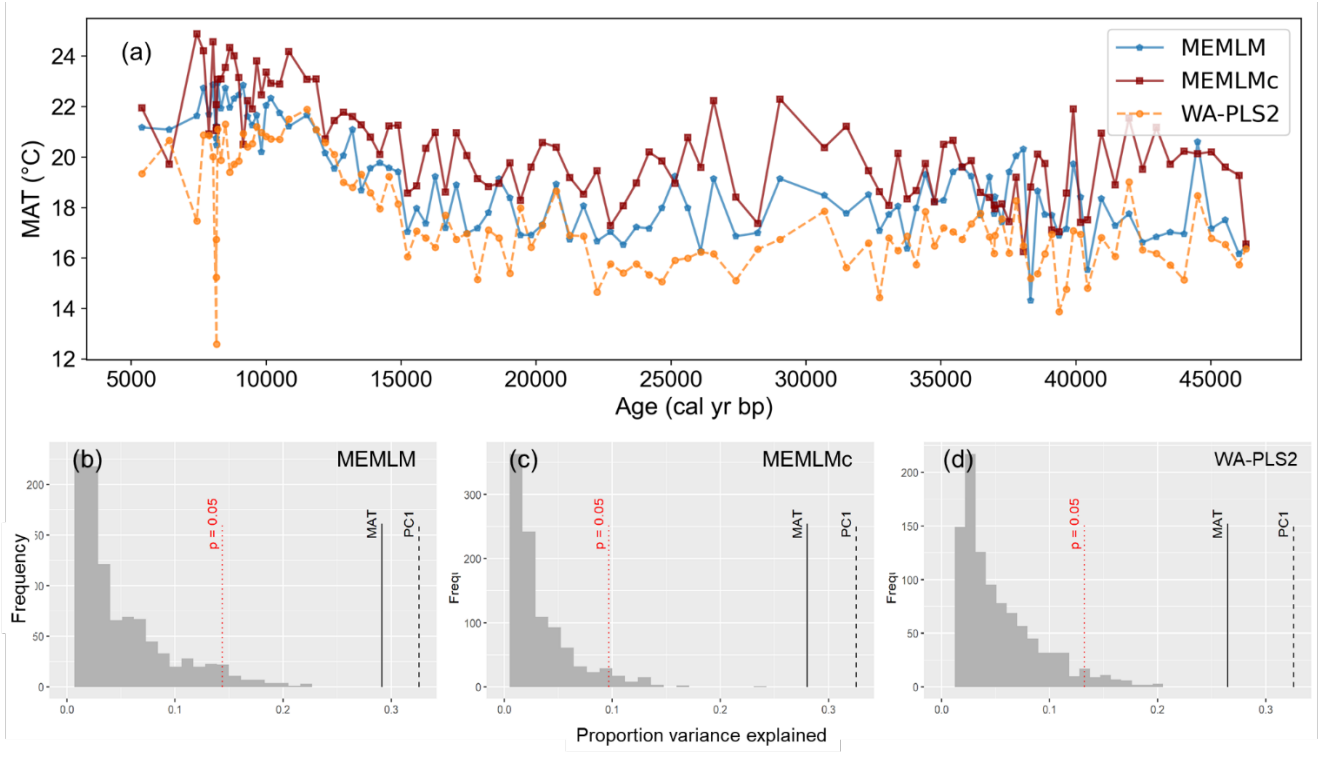


Figure 4: a) MAT reconstruction for the Consuelo core. b, c & d) statistical significance of MEMLM, MEMLMc and WA.PLS2 reconstructions, respectively.

3.2.4. MAT reconstruction from Llaviucu by the NIMBIOS training set

All three methods display similar overall trends with mid-Holocene warming, but each display different centennial variability, which for the MEMLMc reconstruction is clearly unrealistic for the Holocene, with temperature excursions as large as 8°C. Neither of the MEMLM approaches are statistically significant at the 95% confidence level, so neither can be accepted as robust. The WA-PLS2 reconstruction is statistically significant, although it only explains 13% of the core assemblage variance compared to the 28% explained by the first principal component of the core data (Figure 5).

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

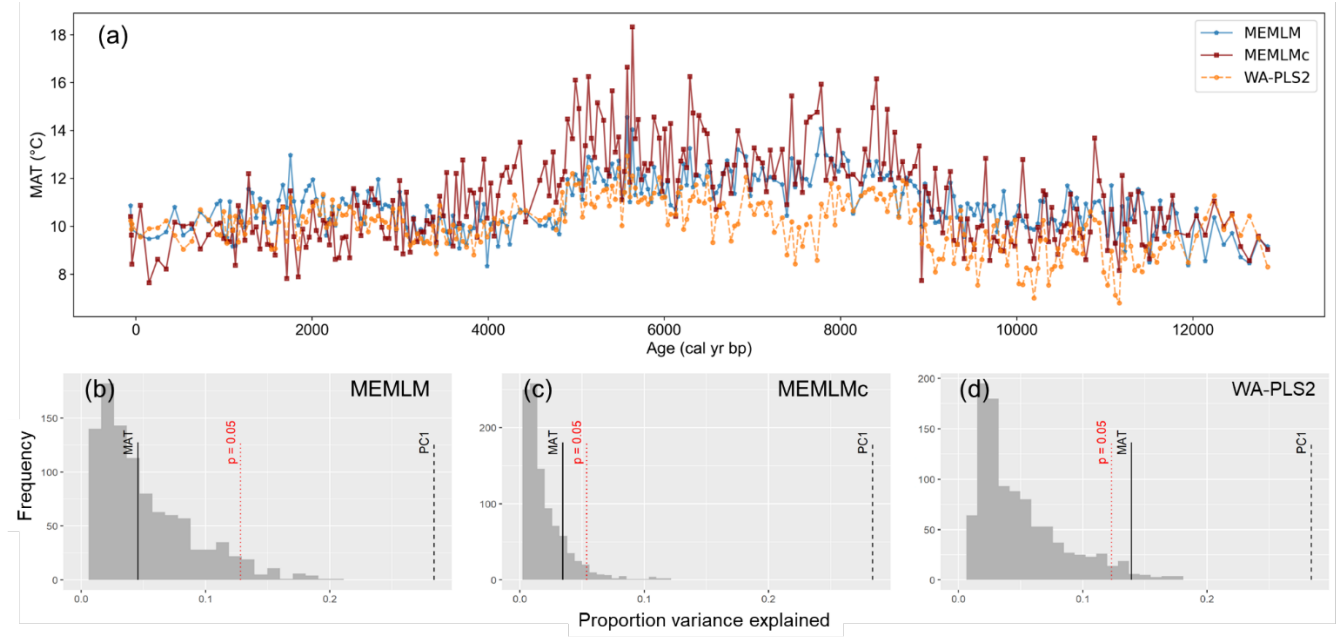


Figure 5: a) MAT reconstruction for the Llaviucu core. b, c & d) statistical significance of MEMLM, MEMLMc and WA-PLS2 reconstructions, respectively.

3.2.6. Reconstruction for core Villarquemado using the SMPDSV1 training set

All three approaches generate noisy reconstructions with high variability that is inconsistent. It is difficult to discern any meaningful trends. None of the reconstructions, including WA-PLS2, are statistically significant. The low (17%) variance associated with the first principal component suggests that the fossil assemblages are responding to multiple environmental factors with responses that are too complex to be captured by a single explanatory environmental variable (Figure 6).

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

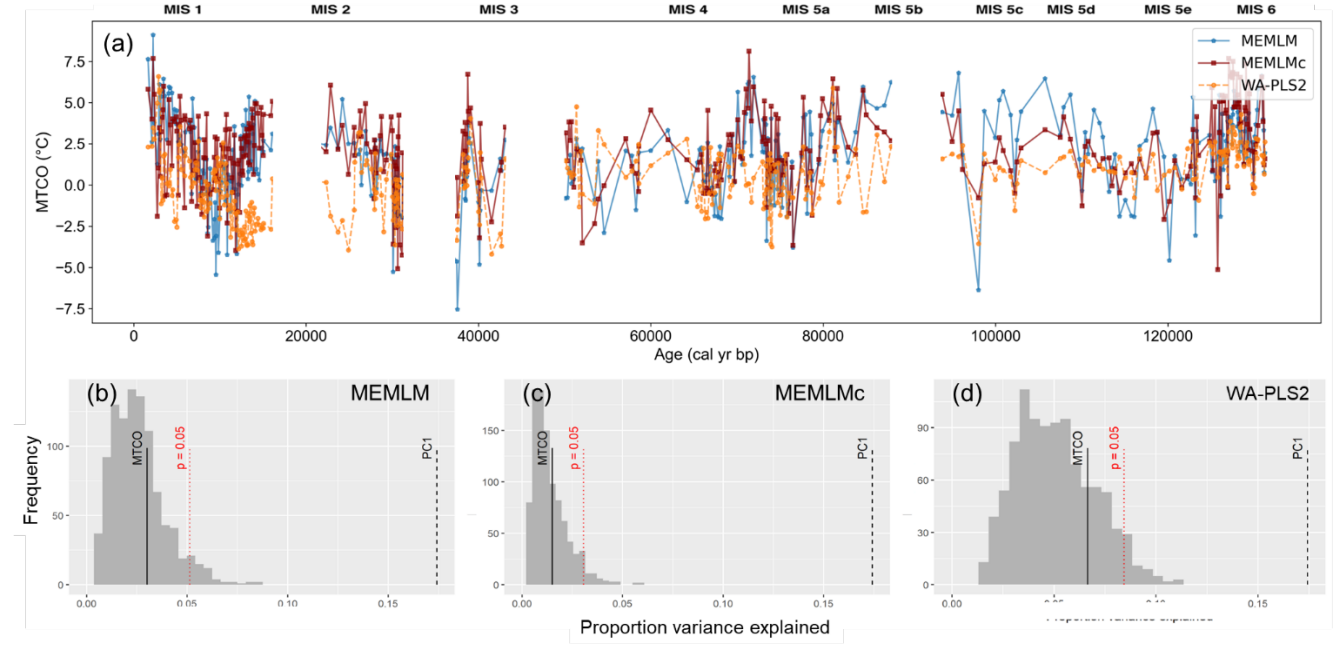


Figure 6: a) MTCO reconstruction for the Villarquemado core. b, c & d) statistical significance of MEMLM, MEMLMc and WA-PLS2 reconstructions, respectively.

4 Discussion and conclusions

We have developed three variants of a multi-model ensemble machine learning algorithm, MEMLM. These each train three separate ensemble machine learning algorithms (random forests, extremely random trees and lightGBM) and combine them into a consensus reconstruction using a weak learner approach based on multiple regression. The three approaches only differ in their input data. The simpler MEMLM takes only taxon abundance data. MEMLMe, built only upon the GLOVE embedding matrix, does not perform as well as MEMLM. However, MEMLMe was found to be a useful reconstruction model, at least when applied to the larger NIMBIOS and SMPDSV1 training sets, and the embedding was able to usefully summarise taxon assemblages with fewer than 50 dimensions. The additional complexity of MEMLMc, which uses both taxon count and embedding, did not significantly affect the predictive performance relative to MEMLM under cross-validation, suggesting that conventional ensemble machine learning approaches are sufficient to encode adequately ecological information in the relatively small data-sets used in these palaeoclimate reconstructions. We note that the real power of embedding (dimension reduction) approaches in ecology is likely to be in their applications to much larger data-sets, when ecological relationships between 10,000's of taxa and their environment are being considered.

The MEMLM approaches were both found to perform better than classical weighted averaging approaches under cross validation. In the case of the smallest SWAP data-set the advantages were modest, but in the largest SMPDSV1 data-set RMSEP errors were reduced by a factor of two relative to the best performing classical WA approach. These improvements

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

in performance clearly validate the potential benefits of strong data mining abilities of machine learning to create a more complete description of a data-set, suggesting these techniques have the potential to improve upon classical reconstruction approaches.

When applied to core reconstructions, MEMLM approaches were found to generate considerably more variability than the WA-PLS reconstructions. While some elements of this additional variability might be realistic, especially considering that WA-PLS approaches are known to bias reconstructions towards the centre of their training data (Liu et al., 2020), the variability was not always found to be coherent between different reconstruction approaches and the magnitude of MEMLM variability was in some cases implausibly high, for example by suggesting Holocene variability of up to 8°C in the Ecuadorian Llaviucu core.

We performed significance testing on all core reconstructions and found that five of the fifteen reconstructions were not statistically significant and therefore should not be considered robust. Both MEMLM and MEMLMc approaches failed on the Llavuicu core, confirming our suspicion that the unrealistic variability was an artefact even though the overall trends of the reconstruction were consistent with the robust WA-PLS2 reconstruction. All three approaches failed the statistical robustness test at Villarquemado, which is sensitive to multiple environmental factors and has responses which appear too complex to be captured by a single explanatory variable.

In summary, while MEMLM can generate useful reconstructions, it should always be used in conjunction with statistical significance testing to ensure the reconstructions are robust and potentially realistic and reliable. The additional complexities of providing assemblage information to MEMLMc did not reduce RMSEP or spurious variability and nor did it improve statistical significance. However, MEMLMe demonstrated that embedding is useful as it can summarise ecological assemblages using significantly fewer dimensions. Its benefits may be felt more clearly in applications to much larger data-sets and in applications beyond palaeoenvironmental reconstructions. The poor performance of MEMLM in some reconstructions may be due to extrapolation due to poor or no analogue fossil assemblages. Even though all models were applied under the same extrapolation, the WA-PLS2 reconstructions were found to be more reliable than MEMLM, although WA-PLS2 also failed to generate robust reconstructions at Villarquemado. We infer that that the use of simpler WA models, which include a major biological assumption (unimodal environmental response) can be more powerful than the use of brute-force learning, despite reductions in RMSEP. We reiterate our recommendation that all reconstructions using any approach, should be accompanied with statistical significance testing. Seemingly useful models may fail when applied under extrapolation or when the assemblage variance is only weakly dependent on the reconstructed environmental variable.

Acknowledgements

PS was funded by a PhD scholarship from China Scholarship Council (CSC, no. 202104910033). HJBB's participation has been possible thanks to the European Research Council under the European Union's Horizon 2020 Research and Innovation Programme grant agreement 74143 to the project 'HOPE: Humans on Planet Earth - long-term impacts on biosphere dynamics'

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

awarded to HJBB. We thank Prof Mark Bush & Dr Alex Correa-Metrio for generously providing the NIMBIOS modern pollen data and Llaviucu fossil pollen data and Prof Graciela Gil Romera for generously providing the SMPDS v1 modern pollen data and Villarquemado fossil pollen data to the public.

Data availability

All datasets can be found in the cited datasets and articles in references, except the RLGH3 and Llaviucu core, which are available from the authors on request.

Code availability

All codes are available in github. WA and WA-PLS use the *rioja* package (https://github.com/nsj3/rioja). Telford and Birks (2011) statistical significance uses *randomTF* in the *palaeoSig* package (https://github.com/richardjtelford/palaeoSig). MEMLM can be found at https://github.com/Schimasuperbra/MEMLM.

Competing interests

The authors declare that they have no conflict of interest.

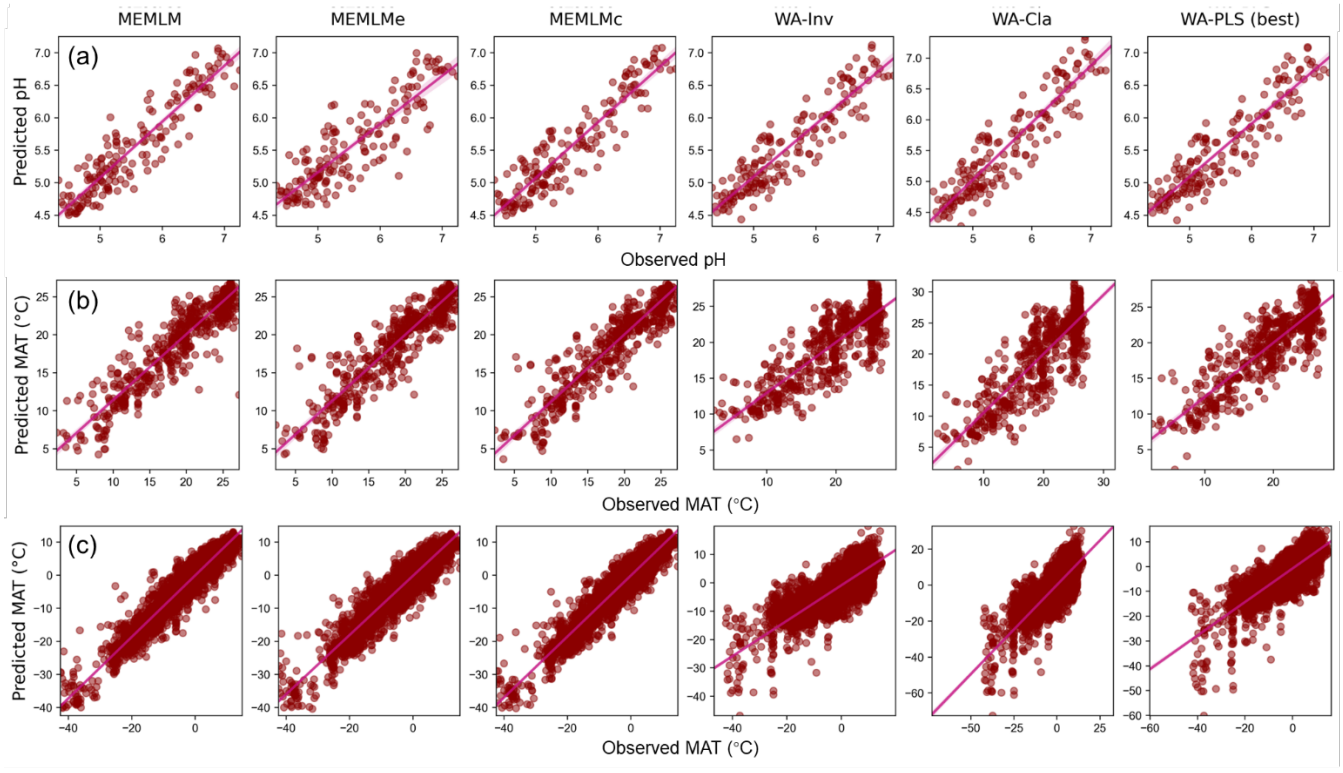
Author contribution

PS conceptualised the application of GLOVE to ecological assemblages. PS, PBH and HJBB conceptualised the experimental design. PS developed the MEMLM model and performed all analysis and graphical visualisation. HJBB contributed assemblage data. PS and PBH wrote the manuscript with reviewing and editing by HJBB.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions
Open Access
EGU

Appendices A



395 **Figure A1: Regression visualization of predicted values against observed values in three training sets. MEMLM uses the abundance matrix. MEMLMe uses the assemblage embedding matrix. Component number of WA-PLS was selected for each training set as the lowest component that showed a 5% improvement over the previous component (Table A3). WA-Cla is weighted averaging with a classical deshrinking regression, WA-Inv is weighted averaging with an inverse deshrinking regression (Birks et al. 1990). WA-PLS's components selected based on method described in 2.3.3, see Table S3 for full results.**

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions
Open Access
EGU

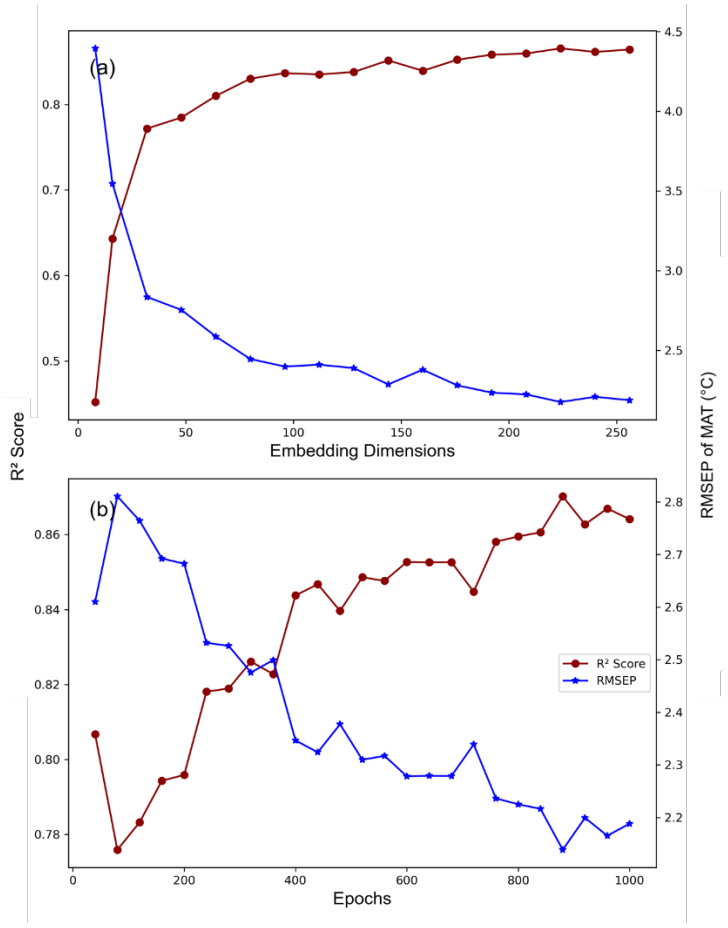


Figure A2: MEMLMe prediction performance under different GLOVE hyper-parameter settings. a) Fix epoch = 1000, set embedding dimensions from 8 to 256; b) Fix embedding dimensions = 256, set epoch from 40 to 1000. The model is developed from the NIMBIOS set and trained upon MAT.

Table A1: RMSEP and R^2 values (based on cross-validation) of the 18 environment elements prediction of MEMLMc in the NIMBIOS data-set. Mean is average of the prediction of the three downstream models. RF presents the random forest values; ET presents the extended random tree results. Bold highlights the model with the best prediction performance.

	Elements	RF	ET	lightGBM	MEMLMc	Mean
RMSEP	Precipitation of the warmest quarter	138.17	131.513	133.124	**125.531**	129.042
	Isothermality	3.065	2.793	3.09	**2.778**	2.838
	Annual precipitation	483.099	442.623	479.217	**430.291**	445.813
	Mean temperature coldest quarter	23.162	21.228	23.061	**21.023**	21.387
	Maximum temperature warmest month	25.104	22.343	24.01	**21.181**	22.421
	Minimum temperature coldest month	26.66	24.15	26.226	**23.734**	24.369
	Mean temperature warmest quarter	22.898	21.435	22.655	**20.727**	21.316
	Precipitation of the coldest quarter	157.458	135.741	151.69	**129.674**	139.075
	Precipitation of the driest month	28.907	25.898	27.892	**23.898**	25.723

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions

	Temperature seasonality	227.1	203.536	221.23	**203.281**	207.248
	Precipitation of the wettest month	64.759	60.669	64.387	**58.822**	60.572
	Temperature annual range	22.179	20.917	22.101	**20.524**	20.83
	Mean temperature wettest quarter	18.312	16.515	18.722	**16.094**	16.865
	Precipitation of the wettest quarter	171.418	161.823	173.802	**157.769**	162.204
	Precipitation seasonality	11.581	10.858	11.506	**10.635**	10.852
	Mean diurnal temperature range	13.139	11.684	12.968	**11.258**	11.855
	Mean temperature driest quarter	23.754	22.225	23.556	**21.989**	22.198
	Precipitation of the driest quarter	96.455	85.831	92.351	**79.657**	85.539
R^2 score	Precipitation of the warmest quarter	0.656	0.688	0.68	**0.716**	0.7
	Isothermality	0.862	0.886	0.86	**0.887**	0.882
	Annual precipitation	0.81	0.841	0.813	**0.85**	0.838
	Mean temperature coldest quarter	0.862	0.884	0.863	**0.886**	0.882
	Maximum temperature warmest month	0.771	0.819	0.79	**0.837**	0.817
	Minimum temperature coldest month	0.887	0.907	0.89	**0.91**	0.905
	Mean temperature warmest quarter	0.85	0.868	0.853	**0.877**	0.87
	Precipitation of the coldest quarter	0.845	0.885	0.856	**0.895**	0.879
	Precipitation of the driest month	0.821	0.857	0.834	**0.878**	0.859
	Temperature seasonality	0.835	**0.868**	0.844	**0.868**	0.863
	Precipitation of the wettest month	0.752	0.782	0.755	**0.795**	0.783
	Temperature annual range	0.848	0.865	0.85	**0.87**	0.866
	Mean temperature wettest quarter	0.822	0.855	0.814	**0.862**	0.849
	Precipitation of the wettest quarter	0.761	0.787	0.755	**0.798**	0.786
	Precipitation seasonality	0.773	0.8	0.776	**0.809**	0.801
	Mean diurnal temperature range	0.803	0.845	0.809	**0.856**	0.84
	Mean temperature driest quarter	0.87	0.887	0.873	**0.889**	0.887
	Precipitation of the driest quarter	0.806	0.846	0.822	**0.868**	0.847

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions

410 **Table A2: The top 10 important taxa for the environmental reconstructions in the SWAP, NIMBIOS and SMPDSv1 training sets sorted by the random forests results.**

	Taxon	RF	ET	LightGBM
SWAP	EU047A	0.505	0.139	0.033
	AC013A	0.072	0.182	0.028
	EU048A	0.061	0.064	0.02
	TA003A	0.048	0.043	0.017
	PE002A	0.031	0.013	0.027
	CM048A	0.023	0.006	0.029
	BR001A	0.022	0.012	0.032
	TA004A	0.018	0.02	0.017
	NA140A	0.012	0.007	0.01
	CM017A	0.011	0.01	0.019
NIMBIOS	Alnus	0.263	0.096	0.045
	Poaceae	0.146	0.161	0.124
	Plantago	0.118	0.039	0.006
	MoracUrtic	0.105	0.02	0.068
	Bursera	0.049	0.016	0.008
	Myrtaceae	0.024	0.007	0.016
	Ericaceae	0.022	0.042	0.021
	Hedyosmum	0.015	0.03	0.035
	Asteraceae	0.013	0.083	0.056
	Cyperaceae	0.013	0.02	0.068
SMPDSv1	Picea	0.339	0.038	0.029
	Fagus	0.169	0.016	0.012
	Betula Chamaebetula.	0.103	0.22	0.008
	Betula	0.042	0.077	0.041
	Alnus Alnobetula	0.039	0.017	0.007
	Larix	0.03	0.03	0.009
	Quercus deciduous	0.028	0.017	0.03
	Olea	0.027	0.072	0.013
	Oxyria Rumex	0.017	0.009	0.019
	Poaceae	0.014	0.0144	0.028

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Climate of the Past Discussions

Table A3: RMSEP (based on cross-validation) of the first five components in WA-PLS of the three training sets. Bold highlights the 'best' component, noting that we accept a higher PLS component only if it exhibits a 5% improvement on the previous component (Birks 1998).

Dataset	Feature	WA-PLS Comp01	Comp02	Comp03	Comp04	Comp05
NIMBIOS	MAT	31.971	**29.136**	30.224	31.712	33.557
SMPDSv1	MTCO	5.304	**4.964**	4.854	4.842	4.876
SWAP	pH	**0.307**	0.299	0.313	0.325	0.344

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

References

Aguirre-Gutiérrez, J., Rifai, S., Shenkin, A., Oliveras, I., Bentley, L. P., Svátek, M., Girardin, C. A. J., Both, S., Riutta, T., Berenguer, E., Kissling, W. D., Bauman, D., Raab, N., Moore, S., Farfan-Rios, W., Figueiredo, A. E. S., Reis, S. M., Ndong, J. E., Ondo, F. E., N'ssi Bengone, N., Mihindou, V., Moraes de Seixas, M. M., Adu-Bredu, S., Abernethy, K., Asner, G. P., Barlow, J., Burslem, D. F. R. P., Coomes, D. A., Cernusak, L. A., Dargie, G. C., Enquist, B. J., Ewers, R. M., Ferreira, J., Jeffery, K. J., Joly, C. A., Lewis, S. L., Marimon-Junior, B. H., Martin, R. E., Morandi, P. S., Phillips, O. L., Quesada, C. A., Salinas, N., Schwantes Marimon, B., Silman, M., Teh, Y. A., White, L. J. T., and Malhi, Y.: Pantropical modelling of canopy functional traits using Sentinel-2 remote sensing data, Remote Sens Environ, 252, 112122, doi:10.1016/j.rse.2020.112122, 2021.

Allott, T. E. H., Harriman, R., and Battarbee, R. W.: Reversibility of lake acidification at the Round Loch of Glenhead, Galloway, Scotland, Environmental Pollution, 77, 219–225, doi:10.1016/0269-7491(92)90080-T 1992, 1992.

Bannar-Martin, K. H., Kremer, C. T., Ernest, S. K. M., Leibold, M. A., Auge, H., Chase, J., Declerck, S. A. J., Eisenhauer, N., Harpole, S., Hillebrand, H., Isbell, F., Koffel, T., Larsen, S., Narwani, A., Petermann, J. S., Roscher, C., Cabral, J. S., and Supp, S. R.: Integrating community assembly and biodiversity to better understand ecosystem function: the Community Assembly and the Functioning of Ecosystems (CAFE) approach, Ecol Lett, 21, 167–180, doi:10.1111/ele.12895, 2018.

Battarbee, R. W., Stevenson, A. C., Rippey, B., Fletcher, C., Natkanski, J., Wik, M., and Flower, R. J.: Causes of Lake Acidification in Galloway, South-West Scotland: A Palaeoecological Evaluation of the Relative Roles of Atmospheric Contamination and Catchment Change for Two Acidified Sites with Non-Afforested Catchments, J Ecol, 77, 651–672, doi:10.2307/2260976, 1989.

Battarbee, R. W., Monteith, D. T., Juggins, S., Evans, C. D., Jenkins, A., and Simpson, G. L.: Reconstructing pre-acidification pH for an acidified Scottish loch: A comparison of palaeolimnological and modelling approaches, Environ Pollut, 137, 135–149, doi:10.1016/j.envpol.2004.12.021, 2005.

Birks, H. J. B, ter Braak C.J.F, Line J.M., Juggins S. and Stevenson A.C. Diatoms and pH reconstruction Phil. Trans. R. Soc. Lond. B, 327, 263–278, doi:10.1098/rstb.1990.0062, 1990.

Birks, H., Birks, H.: D.G. Frey and E.S. Deevey: Review 1: Numerical tools in palaeolimnology – Progress, potentialities, and problems, Journal of Paleolimnology, 20, 307–332. doi:10.1023/A:1008038808690, 1998.

Birks, H. j. b., Braak, C. j. f. Ter, Line, J. M., Juggins, S., Stevenson, A. C., Battarbee, R. W., Mason, B. J., Renberg, I., and Talling, J. F.: Diatoms and pH reconstruction, Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 327, 263–278, doi:10.1098/rstb.1990.0062, 1990.

Blunier, T. and Brook, E. J.: Timing of Millennial-Scale Climate Change in Antarctica and Greenland During the Last Glacial Period, Science, 291, 109–112, doi:10.1126/science.291.5501.109, 2001.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Bond, G., Broecker, W., Johnsen, S., McManus, J., Labeyrie, L., Jouzel, J., and Bonani, G.: Correlations between climate records from North Atlantic sediments and Greenland ice, Nature, 365, 143–147, doi:10.1038/365143a0, 1993.

ter Braak. C. J. F. and Juggins, S.: 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, 1993.

ter Braak, C. J. F. and Barendregt, L. G.: Weighted averaging of species indicator values: Its efficiency in environmental calibration, Math Biosci, 78, 57–72, doi:10.1016/0025-5564(86)90031-3, 1986.

Brook, B. W., Sodhi, N. S., and Bradshaw, C. J. A.: Synergies among extinction drivers under global change, Trends Ecol Evol, 23, 453-460, doi:10.1016/j.tree.2008.03.011, 2008.

Brooks, S. J. and Birks, H. J. B.: Chironomid-inferred air temperatures from Lateglacial and Holocene sites in north-west Europe: progress and problems, Quat Sci Rev, 20, 1723–1741, doi:10.1016/S0277-3791(01)00038-5, 2001.

Bush, M. B., Correa-Metrio, A., van Woesik, R., Collins, A., Hanselman, J., Martinez, P., and McMichael, C. N. H.: Modern pollen assemblages of the Neotropics, J Biogeogr, 48, 231–241, doi:10.1111/jbi.13960, 2021.

Christin, S, Hervet, É, Lecomte, N: Applications for deep learning in ecology, Methods Ecol Evol, 10, 1632–1644, doi:10.1111/2041-210X.13256, 2019.

Clapperton. C. M: Maximum extent of the late Wisconsin glaciation in the Ecuadorian Andes Quaternary of South America and Antarctic Peninsula, Balkema, Rotterdam, 165–180, doi: ISBN 9781003079323, 1987

Cleator, S. F., Harrison, S. P., Nichols, N. K., Colin Prentice, I., and Roulstone, I.: A new multivariable benchmark for Last Glacial Maximum climate simulations, Clim Past, 16, 699–712, doi:10.5194/cp-16-699-2020, 2020.

Colinvaux, P. A., Olson, K., and Liu, K. B.: Late-glacial and holocene pollen diagrams from two endorheic lakes of the inte-andean plateau of ecuador, Rev Palaeobot Palynol, 55, 83–99, doi:10.1016/0034-6667(88)90055-3, 1988.

Colinvaux, P. A., Bush, M. B., Steinitz-Kannan, M., and Miller, M. C.: Glacial and Postglacial Pollen Records from the Ecuadorian Andes and Amazon, Quat Res, 48, 69–78, doi:10.1006/qres.1997.1908, 1997.

Dormann, C. F., Calabrese, J. M., Guillera-Arroita, G., Matechou, E., Bahn, V., Bartoń, K., Beale, C. M., Ciuti, S., Elith, J., Gerstner, K., Guelat, J., Keil, P., Lahoz-Monfort, J. J., Pollock, L. J., Reineking, B., Roberts, D. R., Schröder, B., Thuiller, W., Warton, D. I., Wintle, B. A., Wood, S. N., Wüest, R. O., and Hartig, F.: Model averaging in ecology: a review of Bayesian, information-theoretic, and tactical approaches for predictive inference, Ecol Monogr, 88, 485–504, doi:10.1002/ecm.1309, 2018.

Féret, J. B., Berger, K., de Boissieu, F., and Malenovský, Z.: PROSPECT-PRO for estimating content of nitrogen-containing leaf proteins and other carbon-based constituents, Remote Sens Environ, 252, doi:10.1016/j.rse.2020.112173, 2021.

Geurts, P., Ernst, D., and Wehenkel, L.: Extremely randomized trees, Mach Learn, 63, 3–42, doi:10.1007/s10994-006-6226-1, 2006.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Friedman J.H.: Greedy function approximation: A gradient boosting machine, Ann. Statist, 29(5), 1189–1232, doi:10.1214/aos/1013203451, 2011.

Harrison, S. P: Modern pollen data for climate reconstructions, version 1 (SMPDS), University of Reading, doi: 10.17864/1947.194, 2019.

Harrison, S.P., González-Sampériz, P., Gil-Romera, G.: Fossil pollen data for climate reconstructions from El Cañizar de Villarquemado, University of Reading, doi: 10.17864/1947.219, 2019.

Harrison, S.P.: Climate reconstructions for the SMPDSv1 modern pollen data set. doi: 10.5281/zenodo.3605003, 2020.

Hais, M., Komprdová, K., Ermakov, N., and Chytrý, M.: Modelling the Last Glacial Maximum environments for a refugium of Pleistocene biota in the Russian Altai Mountains, Siberia, Palaeogeogr Palaeoclimatol Palaeoecol, 438, 135–145, doi:10.1016/j.palaeo.2015.07.037, 2015.

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

Helama, S., Makarenko, N. G., Karimova, L. M., Kruglun, O. A., Timonen, M., Holopainen, J., Meriläinen, J., and Eronen, M.: Dendroclimatic transfer functions revisited: Little Ice Age and Medieval Warm Period summer temperatures reconstructed using artificial neural networks and linear algorithms, Ann Geophys, 27, 1097–1111, doi:10.5194/angeo-27-1097-2009, 2009.

Holden, P. B., Birks, H. J. B., Brooks, S. J., Bush, M. B., Hwang, G. M., Matthews-Bird, F., Valencia, B. G., and Van Woesik, R.: 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, 2017.

Huang, Y., Yang, L. and Fu, Z.: Reconstructing coupled time series in climate systems using three kinds of machine-learning methods, Earth Syst Dynam, 11, 835–853, doi:10.5194/esd-11-835-2020, 2020.

V.J. Jones, A.C. Stevenson, R.W: Battarbee, Acidification of lakes in Galloway, south-west Scotland: a diatom and pollen study of the post-glacial history of the Round Loch of Glenhead, J Ecol., 77, 1-22, doi:10.2307/2260912, 1989.

Houssaye, B. D. La, Flaming, P. L., Nixon, Q., and Acton, G. D.: Machine Learning and Deep Learning Applications for International Ocean Discovery Program Geoscience Research, in: SMU Data Science Review, 2(3), 9, https://scholar.smu.edu/datasciencereview/vol2/iss3/9, 2019.

Jordan, G. J., Harrison, P. A., Worth, J. R. P., Williamson, G. J., and Kirkpatrick, J. B.: Palaeoendemic plants provide evidence for persistence of open, well-watered vegetation since the Cretaceous, Glob Ecol Biogeogr, 25, 127–140, doi:10.1111/geb.12389, 2016.

Juggins, S: Rioja: analysis of Quaternary science data, CRAN [code], R package version (0.9–15.1), https://github.com/nsj3/rioja, 2017.

Steinitz-Kannan M., Colinvaux P.A., Kannan R.: Limnological Studies in Ecuador 1. A survey of chemical and physical properties of Ecuadorian lakes, Arch Hydrobiol, Suppl., 65, 61-105, 1983.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., and Liu, T.-Y.: LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, 30, 2017.

Legendre, P., Galzin, R., and Harmelin-Vivien, M. L.: Relating behavior to habitat: solutions to the fourth-corner problem, Ecology, 78, 547–562, doi:10.1890/0012-9658(1997)078[0547:RBTHST]2.0.CO;2, 1997.

Liaw, A., Wiener, M.: Classification and regression by randomForest, R news, 2(3), 18-22, http://www.stat.berkeley.edu/users/breiman/, 2002.

Liu, M., Prentice, I. C., Ter Braak, C. J. F., and Harrison, S. P.: An improved statistical approach for reconstructing past climates from biotic assemblages, Proceedings of the Royal Society A, 476, doi:10.1098/rspa.2020.0346, 2020.

Norberg, A., Abrego, N., Blanchet, F. G., Adler, F. R., Anderson, B. J., Anttila, J., Araújo, M. B., Dallas, T., Dunson, D., Elith, J., Foster, S. D., Fox, R., Franklin, J., Godsoe, W., Guisan, A., O'Hara, B., Hill, N. A., Holt, R. D., Hui, F. K. C., Husby, M., Kålås, J. A., Lehikoinen, A., Luoto, M., Mod, H. K., Newell, G., Renner, I., Roslin, T., Soininen, J., Thuiller, W., Vanhatalo, J., Warton, D., White, M., Zimmermann, N. E., Gravel, D., and Ovaskainen, O.: A comprehensive evaluation of predictive performance of 33 species distribution models at species and community levels, Ecol Monogr, 89, doi:10.1002/ecm.1370, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É.: Scikit-Learn: Machine Learning in Python. J. Mach. Learn. Res., 12, 2825–2830, 2011.

Pennington, J., Socher, R., and Manning, C.: GloVe: Global Vectors for Word Representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543, doi:10.3115/v1/D14-1162, 2014.

Ovaskainen, O., Tikhonov, G., Norberg, A., Blanchet, F.G., Duan, L., Dunson, D., Roslin, T. & Abrego, N.: How to make more out of community data? A conceptual framework and its implementation as models and software, Ecol Lett, 20, 561-576, doi:10.1111/ele.12757. 2017, 2017.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A., Andreas K., Edward Z. Y., Zachary D., Martin R., Alykhan T., Sasank C., Benoit S., Lu F., Junjie B. and Soumith C.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32, 8024-8035, arXiv:1912.01703, 2019.

Rasmussen, S. O., Bigler, M., Blockley, S. P., Blunier, T., Buchardt, S. L., Clausen, H. B., Cvijanovic, I., Dahl-Jensen, D., Johnsen, S. J., Fischer, H., Gkinis, V., Guillevic, M., Hoek, W. Z., Lowe, J. J., Pedro, J. B., Popp, T., Seierstad, I. K., Steffensen, J. P., Svensson, A. M., Vallelonga, P., Vinther, B. M., Walker, M. J. C., Wheatley, J. J., and Winstrup, M.: A stratigraphic framework for abrupt climatic changes during the Last Glacial period based on three synchronized Greenland ice-core records: Refining and extending the INTIMATE event stratigraphy, Quat Sci Rev, 106, 14–28, doi:10.1016/j.quascirev.2014.09.007, 2014.

https://doi.org/10.5194/cp-2023-69
Preprint. Discussion started: 27 September 2023

Telford, R. J. and Birks, H. J. B.: A novel method for assessing the statistical significance of quantitative reconstructions inferred from biotic assemblages, Quat Sci Rev, 30, 1272–1278, doi:10.1016/j.quascirev.2011.03.002, 2011.

Telford, R. J.and Trachsel, M.: palaeoSig: Significance Tests for Palaeoenvironmental Reconstructions. R Package Version 1.1-3. Bergen: University of Bergen, 2015.

Schulte, P.J. and Hinckley, T.M.: A Comparison of Pressure-Volume Curve Data-Analysis Techniques, J Exp Bot, 36, 1590-1602, doi: 10.1093/jxb/36.10.1590, 1985.

Stevenson, A.C, Juggins, S., Birks, H.J.B., Anderson, D.S., Anderson, N.J., Battarbee, R.W., Berge, F., Davis, R.B, Flower, R.J. & Haworth, E.Y, The Surface Waters Acidification Project Palaeolimnology Programme: Modern Diatom/Lake-Water Chemistry Data-Set, UCL Environmental Change Research Centre, doi:10.1098/rstb.1990.0056, 1991.

Syam, N. and Kaul, R.: Overfitting and Regularization in Machine Learning Models in Machine Learning and Artificial Intelligence in Marketing and Sales, Emerald Publishing Limited, Bingley, 65-84, doi: 10.1108/978-1-80043-880-420211004, 2021

Turner, M. G., Wei, D., Prentice, I. C., and Harrison, S. P.: The impact of methodological decisions on climate reconstructions using WA-PLS, Quat Res, 99, 341–356, doi:10.1017/qua.2020.44, 2020.

Tylianakis, J. M., Didham, R. K., Bascompte, J., and Wardle, D. A.: Global change and species interactions in terrestrial ecosystems, Ecol Lett, 11, 1351-1363, doi:10.1111/j.1461-0248.2008.01250.x, 2008.

Urrego, D. H., Bush, M. B., and Silman, M. R.: A long history of cloud and forest migration from Lake Consuelo, Peru, Quat Res, 73, 364–373, doi:10.1016/j.yqres.2009.10.005, 2010.

Wei, D., González-Sampériz, P., Gil-Romera, G., Harrison, S. P., and Prentice, I. C.: Seasonal temperature and moisture changes in interior semi-arid Spain from the last interglacial to the Late Holocene, Quat Res, 101, 143–155, doi:10.1017/qua.2020.108, 2021a.

Wei, G., Peng, C., Zhu, Q., Zhou, X., and Yang, B.: Application of machine learning methods for paleoclimatic reconstructions from leaf traits, International Journal of Climatology, 41, E3249–E3262, doi:10.1002/joc.6921, 2021b.

Yates, L. A., Aandahl, Z., Richards, S. A., and Brook, B. W.: Cross validation for model selection: A review with examples from ecology, Ecol Monogr, 93, doi:10.1002/ecm.1557, 2023.

Yeom, S., Giacomelli, I., Fredrikson, M. & Jha, S.: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting, 2018 IEEE 31st Computer Security Foundations Symposium (CSF), 268-282, doi:10.1109/CSF.2018.00027, 2018.

Zhou, Z.H.: Ensemble Methods: Foundations and Algorithms, 236, CRC Press, New York, ISBN 9780429151095, 2012.