Clim. Past Discuss.,
https://doi.org/10.5194/cp-2018-87-AC1, 2018

Climate of the Past Discussions

Interactive comment on "Combining a pollen synthesis and climate simulations for spatial reconstructions of European climate using Bayesian modelling" *by* Nils Weitzel et al.

Nils Weitzel et al.

nils.weitzel@uni-bonn.de

Received and published: 10 December 2018

We thank the referee for taking the time to review our discussion paper, and for helpful and interesting comments which will improve the quality of our manuscript. As a response to the referee's suggestions, we plan substantial changes of the manuscript. In the following, referee's comments (RC) are given in blue italicised text and followed by our respective responses (AR).

"What is the meaning of the covariance Σ? if it really is the inter-model variability,

why mix over $\omega_k N(\mu_k, \Sigma)$ instead of just using the distribution $N(\bar{\mu}, \Sigma)$? What is the meaning of this distribution that your are mixing over? Σ is the covariance of the model ensemble, not a particular realization μ_k, thus the distribution you propose doesn't make much sense to me as a meaningful statsitical object. I'm open to being convinced that this idea makes sense, but right now I don't see the purpose.
Also, using this mixing distribution is likely to eliminate the spatial autocorrelation in the μ_k as neighboring locations are now no longer draws from the same climate model. Instead, it might be simpler and make more sense to mix over $\omega_k \mu_k$. This mixing distribution is over the climate model ensemble and will preserve the spatial autocorrelations in the climate models. Ultimately, I'm not convinced that this covariance is what you want to model."

The general motivation for using the kernel mixture distribution is that each ensemble member is seen as a sample from an unknown distribution of all possible climate states. Because of limitations of the climate models and the small amount of available simulations this is of course a very small subset of possible states. We agree with the referee that ideally one would like the covariance matrix Σ_k of each kernel μ_k to correspond to the climate model k such that the spatial autocorrelation of model k is preserved by Σ_k and such that a draw from kernel k is a draw from climate model k. Unfortunately, for each model there is only one run available for the mid-Holocene and the internal variability in those runs is much smaller than the inter-model differences. Therefore, using the internal variability of those runs would lead to very distinct kernels and therefore the range of possible states would be very small. Another possibility would be to use long PI control runs, from which more samples could be extracted but still the problem persists that in these runs the internal variability is much smaller than the inter-model differences. Therefore, we estimate Σ from the inter-model differences as a compromise that allows to sample from a much broader range of states even though autocorrelation from the individual models is lost. To our knowledge, using the empirical covariance of the samples is a very common choice in kernel based prob-

ability density approximations (Silverman, 1986, Chapter 3 and 4) if there is no good estimate of the covariance corresponding to each sample available. Two advantages of mixing over $\omega_k \mathcal{N}(\mu_k, \Sigma)$ compared to using simply $\mathcal{N}(\bar{\mu}, \Sigma)$ are that we do not have to assume that the unknown prior distribution is Gaussian and that we do not rely on an iid assumption for the first moment properties of the kernels. In other words, for the first moment properties of the spatial prior distribution, we do not assume that the model runs are iid samples from a Gaussian distribution but just that there exists an abstract probability density of possible states and that the kernel corresponding to each μ_k is Gaussian. On the other hand, we do rely on an iid assumption for the second moment properties, which is the compromise we make as described above, due to not seeing a better alternative unless one prescribes second moment properties that are not calculated from climate models. In a meteorology context, the advantages of kernel filters compared to standard Gaussian filtering are described in Anderson and Anderson (1999). A more recent application of kernel filtering in data assimilation is Liu et al. (2016). In both examples, the sample covariance is used as covariance of the samples.

The referee suggests two other models for the spatial prior distribution which would make the inference procedure easier, namely using a Gaussian distribution with the ensemble mean as mean and the inter-model variability to estimate the covariance, i.e. $\mathcal{N}(\bar{\mu}, \Sigma)$, or mixing just over $\omega_k \mu_k$, i.e. $\mathcal{N}(\sum_{k=1}^{K} \omega_k \mu_k, \Sigma)$. Both models have advantages and disadvantages that we want to mention.

Using $\mathcal{N}(\bar{\mu}, \Sigma)$ is the most common approach in data assimilation applications in climatology. It is based on the assumption that the ensemble members are iid samples from an unknown Gaussian distribution which contains all possible climate states. The main advantage of this model is that inference becomes much simpler because the prior distribution is unimodal and Gaussian and not multi-modal as in the kernel approach that we applied. The disadvantage is that it relies on the very strong as-

sumption that the samples μ_k are iid samples from an unknown Gaussian distribution. This assumption tends to be more realistic for samples from just one climate model, whereas statistics of multi-model ensembles are often not well described by purely Gaussian distributions (Knutti et al., 2010). A second disadvantage of this model is that it reduces the degrees of freedom in the model compared to the kernel model, and therefore limits the possibilities of the model to adjust the posterior distribution to the data. This could be particularly important due to the small ensemble size in our application.

The prior distribution $\mathcal{N}(\sum_{k=1}^{K} \omega_k \mu_k, \Sigma)$ has the advantage that similar to the kernel approach, it is not relying on an iid assumption of the samples for the first moment properties of the distribution. In addition, it introduces more degrees of freedom, as it allows weighted averages of the ensemble members beyond choosing individual members with a certain probability. In addition, the inference becomes easier compared to the kernel approach as this model allows smooth transitions between the μ_k such that the MC^3 (parallel tempering) method is not required to achieve efficient sampling. Similar models are popular in postprocessing of climate model ensembles as in many applications weighted averages outperformed each individual ensemble member (Krishnamurti et al., 1999). A disadvantage of this type of mixing is that even more spatial structure of the μ_k could be lost compared to the kernel approach because in addition to using the inter-model covariances, the prior mean is a linear combination of the different climate models and does not represent just one single model.

Summarizing, there are good arguments to use each of the three models, the one that we proposed as well as the two suggested be the referee. Similarly, each of the models has disadvantages. Therefore, we decided to test all three models. We additionally compared the performance of using the empirical covariance matrix regularized by the

glasso algorithm with a shrinkage approach, where the empirical correlation matrix is combined with a Matérn type correlation matrix. This is an approach to account for potential model inadequacies to explain the data (see the response to Referee 2). By estimating the shrinkage weight α from the proxy data, this allows deviations from the spatial structures prescribed by the climate simulations if the spatial modes of the Matérn type correlation matrix fit the proxy data better than the inter-model differences.

To compare these six different models, we use cross-validation experiments as described in Sect. 3.5 and 4.2. In addition, we perform identical twin experiments (also named pseudo-proxy experiments or observation system simulation experiments), where one climate model is left out as reference climatology, pseudo-proxies are simulated from this reference climatology and the skill of the corresponding posterior distribution to predict the reference climatology is analysed (see also the response to Referee 2). Initial results suggest that the three statistical models where the glasso regularized covariance matrix Σ is used, perform on a similar level, with almost equal Brier scores in the cross-validation experiments and similar skill in the identical twin experiments. On the other hand, the statistical models with the shrinkage covariance matrix outperfom the glasso regularized covariance models in cross-validation as well as identical twin experiments. The main reason for this better performance seems to be a strong reduction of the underdispersion of the posterior distribution due to the increase of spatial modes in the covariance matrix.

In the revised manuscript, we will enhance the motivation for our statistical models. In addition, we will discuss the three different statistical models, the one which we proposed in the manuscript as well as the two that the referee suggested. Moreover, we will describe the two types of covariance matrices, which we compared as a response to the suggestions of the referees. A section on the comparison of the statistical models based on cross-validation and identical twin experiments will be

added. We hope that these changes will not just improve the results of this study but can also provide guidelines for future applications of Bayesian filtering methods in paleoclimate applications.

"I would also suggest the title should be "... using Bayesian Filtering" rather than "Bayesian Modelling." You aren't modeling the climate, just assimilating the proxy data to filter the climate model ensemble. For instance, if the proxies suggest a temperature higher than all of the climate models, then the best your framework can do is the highest temperature from the climate model ensemble (plus a little error from Σ). Hence, your models are only as reasonable as the climate models (which are likely not great estimates for a given time and location...)."

We agree with the referee that the current manuscript title is slightly misleading and that replacing "Bayesian modelling" by "Bayesian filtering" is a more accurate description of our study. Therefore, we will change the title according to the suggestion of the referee in the revised manuscript.

"The fit of the pollen data by a normal distribution is really poor. The fitted distributions in Figure 3 look nothing like the data distributions. Perhaps a better model is needed."

While we agree with the referee that there is room for improvement of the response surfaces parametrized by Eq. (7) in the manuscript, we do not think that the fit is "really poor". But maybe the visualisation of the response surfaces in Fig. 3 can be improved. To underline that the chosen parametrization is a reasonable approximation of the probability of presence of the taxa, we plot an alternative to Fig. 3 at the end

of this response. In that plot the coloured raster represents the ratio of taxa presence in bins of size 1K $\times$ 1K, and the contour lines visualize the fitted response surface. Considering that the imperfect sampling of the climate space by the calibration dataset leads to weaker signals in some parts of the climate space, particularly at the edges of the sampled area of the climate space, we think that our current parametrization is a reasonable choice. Therefore, we do not plan to change the parametrization of the response surfaces given by Eq. (7) in the revised manuscript but will look at options to improve the visualisation of the data.

"In addition, if you are only using presences in the fossil pollen, your calibration is biased. It would be better to treat absences as a zero-inflated model where an absence could be a true absence or a missing presence due to non-climatic reasons. This is easily done by introducing a latent variable (like you did for the z). In the ecological literature on occupancy modeling, this is known as detection modeling (MacKenzie et. a. 2002)"

The idea of the indicator taxa method is to use taxa which are sensitive to certain climate variables to constrain past climate based on the presence of a taxa in a fossil sample. The probabilistic indicator taxa method (PITM) is an extension of this method where probability distributions have been used to characterize the climate space at which a taxa occurs instead of using binary limits (e.g. a taxa occurs above a certain temperature but not below it) to acknowledge that most taxa have a preferred climate space but the transitions between climates where they usually occur and those where they do not grow is soft. This extension was named pdf method in the literature (Kühl et al., 2002). To estimate these distributions, vegetation data is used instead of modern pollen data, because it contains more accurate information on the presence or absence of a taxa on the spatial scales that we are interested in. As

our Bayesian model is more straightforward formulated by modelling vegetation given climate (forward model) instead of climate given vegetation (inverse model), we have rewritten the pdf method as a forward model by fitting the quadratic logistic regression model, but the aim of this reformulation is to imitate this well-tested method as closely as possible because an extension or improvement of this method is beyond the scope of this study. Because of the reliability of the modern vegetation and climate data, we use presence and absence data to fit the logistic regression. The disadvantage of using vegetation data for the calibration is that the probability of presence of a taxa is only valid in vegetation space on the spatial scale taken for the training data but not in the pollen or macrofossil space, where an absence of a taxa in a pollen or macrofossil sample can have multiple non-climatic reasons like local plant competition or pollen transport effects, as well as local climate effects below the resolution of our study such that the taxa did not grow in the immediate surrounding of the sample. Therefore, the only reliable information on the presence or absence of a taxa in the respective spatial domain (grid box) in the past is the occurrence of the taxa in a pollen or macrofossil sample. Hence, we only use presence information in the reconstruction step.

We agree with the referee that using only present taxa in the fossil pollen is inconsistent with the modern calibration. Despite this inconsistency, our reconstructions are in agreement with previous versions of the probabilistic indicator taxa method, where this inconsistency with the calibration did not appear as previously only presence information were used to fit the probability density functions.

However, we do not see a simple solution for the problem that our calibration is in vegetation space whereas the absence of taxa is an information in the pollen or macrofossil space. The referee suggests to model the absence due to non-climatic reasons as a zero-inflated model by adding a latent variable to estimate the detection probability of a taxa. We think that this is a very promising idea but while the

formulation of this model is simple, the estimation of the detection probability is a very challenging task because it depends on many factors like pollen influx area of the fossil sample, local topography, soil properties, and plant competition which might change over time. It is a priori unclear which of these factor can be marginalized and whether a single detection probability for each taxa is a reasonable approximation. In addition, our fossil dataset combines macrofossils with pollen. The processes that influence the detection probability of macrofossils are very different than for pollen. Therefore, a different detection probability has to be estimated for pollen than for macrofossils.

To our knowledge, the detection probability has never been estimated explicitly as a probability of a presence of a taxa in a pollen or macrofossil sample given the occurrence of the taxa in the respective grid box, but only as a combination of climate as well as non-climate related zero-inflation (e.g Salter-Townshend and Haslett, 2012). We acknowledge that extending the indicator taxa method to include presence and absence information in fossil pollen and macrofossil samples by modelling detection probabilities of taxa should be a focus of future research. But resolving all the described issues requires extensive cooperation of (paleo)climatologists, (paleo)botanists, and statisticians, and is beyond the scope of this study.

To acknowledge the comment of the referee, we will change the following in the revised manuscript: We will describe the underlying assumptions of the PITM model more detailed and elaborate on the inconsistency between the calibration and reconstruction procedure. In addition, we will mention the modelling of detection probabilities as a topic for future research and name the involved issues that need to be solved to accurately model detection probabilities. Finally, we will point out more explicitly that the proxy dataset contains pollen and macrofossil data and the differences of those two data types.

"Equation 13: If mixing over the z's is the problem, why not integrate/marginalize them out? You can always recover them using composition sampling later. It seems like an awfully complex computational framework (MCËĘ3) for such a simple model framework that could be fixed simply by marginalization."

The reason for using the MC^3 (parallel tempering) framework is not the mixing of z but the multi-modality of the prior distribution Eq. (6). The poor mixing of z in our case if we do not use MC^3 is just the manifestation of that problem. It is widely acknowledged in the literature that the design of efficient MCMC methods for multi-modal models is a challenging task, in particular in multivariate settings. The main issue is the construction of efficient proposal samples, which explore the distribution of the individual modes and jump from one mode to another. MC^3 (parallel tempering) is a common technique to solve this issue (Tawn and Roberts, 2018).

Marginalization of z just shifts the general issue to another part of the inference algorithm, as it makes the creation of efficient proposal samples for C a lot more complicated. The introduction of z leads to a conditional Gaussian prior distribution of C which facilitates sampling from the full conditional distribution of C for the grid boxes without proxy data and sequential updates of the grid boxes with proxy data. We do not see a simpler strategy which produces efficient proposal samples, when z is marginalized (for example, a recent study shows that gradient based MCMC methods like Hamiltonian Monte Carlo are not faster than random walk Metropolis-Hastings algorithms for multi-modal problems (Mangoubi et al., 2018), which is intractable in our case due to the degrees of freedom of the posterior distribution).

Summarizing, we agree that MC^3 is a complex framework, for a model which looks

simple at first sight. However, we do not think that marginalization of z is a simple solution because it only shifts the problem. Therefore, we do not see a simple modification of our MCMC algorithm to fit the kernel model, which would make the algorithm less complex. When the two other models suggested by the referee (see above) are fitted, the multi-modality of the posterior distribution vanishes such that a much simpler MCMC algorithm without parallel tempering can be used. This would indicate an additional advantage of those two models in the case of proper reconstruction performance of the two added models.

"Maybe I missed it but what is the size of the model ensemble K and the number of calibration sites?"

The model ensemble has $K = 7$ members (implicitly stated in Sect. 2.2 and Table 1 of the manuscript). The regions that have been used for the transfer function calibration were determined separately for each taxa by pollen experts (Kühl et al., 2007). The number of calibration sites varies between 14.543 and 28.844, depending on the taxa. We will report these numbers in the revised manuscript.

"How to you evaluate the Brier score when you don't include the absences? This seems to introduce a bias and could make the Brier score improper (which limitis its usefulness in comparing models)."

We agree with the referee that using only occurring taxa to evaluate the Brier score is problematic and could make the Brier score improper when it is used to compare general models for predicting taxa presence and absence. However, our goal is an

indirect evaluation of predictions of past climate via transfer functions. In that context, it would lead to inconsistencies between the local reconstructions and the Brier score evaluations when we would include absences as there is currently no model available to accurately estimate detection probabilities for the reasons described above. In that context, inconsistencies mean that the local reconstructions could prefer systematically different climates than the Brier scores, when in one case absences would be included but not in the other case.

Therefore, we think that the evaluation would be much improved from future research to accurately estimate detection probabilities as this would allow the inclusion of present and absent taxa. But for the reasons described above, such an estimation is beyond the scope of this study. We think that the way how we use the Brier scores is still a useful technique to indirectly evaluate climate reconstructions. It should be noted that for each taxa the Brier scores are minimal for a unique climate state, but that minimum is bounded away from zero because the occurrence probability is bounded below one by the response surfaces. In addition, they are a convex function of the climate state for each taxa. We think that these two properties make our methodology useful for the comparison of climate reconstructions but it has to be noted that the comparison is conditioned on the correctness of the response surfaces.

Alternatively, predictions of past climate could be directly compared with probabilistic local reconstructions from the inversion of forward models, but this would mean that the local reconstruction have to be treated as (noisy) observations and not as an inferred product. Hence, we prefer to apply the forward model to the climate reconstructions and then compare the resulting predictions with observations in taxa space. This indirect strategy is also a recent way to infer skill of weather predictions.

In the revised manuscript, we will discuss the limitations of our evaluation methodology

more extensively and describe more detailed in which way it should be seen and interpreted. However, we do not plan to change the methodology because we do not see an easy way to fix its disadvantages for the reasons described above.

"Scientific quality: Are the scientific approach and applied methods valid? Are the results discussed in an appropriate and balanced way (consideration of related work, including appropriate references)?
There are some questions about the implementation of the statistical model that are not completely resolved. (particularly the mixing distribution for climate doesn't make sense and the lack of absence data introduces bias in the estimates). The comments above can provide some guidance in resolving these issues."

We hope that our changes according to the responses to the referee's comments above will improve the scientific quality of the revised manuscript.

"Presentation quality: Are the scientific results and conclusions presented in a clear, concise, and well-structured way (number and quality of figures/tables, appropriate use of English language)?
The paper is reasonably well written from a technical perspective, although more motivation of why particular methods/equations are chosen would be useful. In other words, there is a lot written about what the methods are by not much about why the methods are chosen and what the ideas are trying to solve."

We agree with the referee that more motivation for the statistical models and the respective inference algorithm would improve the quality of the manuscript substantially.

Therefore, we will explain our modelling choices in Sect. 3 (Methods) more extensively in the revised manuscript.

"Are the scientific methods and assumptions valid and clearly outlined?
Not always (at least for the statistical methods)."

We hope that our changes according to the responses to the referee's comments above will improve the validity of the scientific methods and assumptions as well as its presentation.

"Does the title clearly reflect the contents of the paper?
Yes, with a small change of emphasis"

The title of the revised manuscript will be changed according to the referee's suggestion.

"Is the description of experiments and calculations sufficiently complete and precise to allow their reproduction by fellow scientists (traceability of results)?
For the most part. A gitHub repository/code base would go a long way."
"Is the amount and quality of supplementary material appropriate?
I would like to see more done for reproducibility. The computational methods seem overly complex and making code availalbe for replication would be useful."

The revised manuscript will include paragraphs on data and code availability. We will create a repository to share our code. This repository will be referenced in the revised manuscript.

References

Anderson, J. L. and Anderson, S. L.: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts, Monthly Weather Review, 127, 2741–2758, 1999.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, Journal of Climate, 23, 2739–2758, 2010.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science, 285, 1548–1550, 1999.

Kühl, N., Gebhardt, C., Litt, T., and Hense, A.: Probability Density Functions as Botanical-Climatological Transfer Functions for Climate Reconstruction, Quarternary Research, 58, 381–392, 2002.

Kühl, N., Litt, T., Schölzel, C., and Hense, A.: Eemian and Early Weichselian temperature and precipitation variability in northern Germany, Quarternary Science Reviews, 26, 3311–3317, 2007.

Liu, B., Ait-El-Fquih, B., and Hoteit, I.: Efficient Kernel-Based Ensemble Gaussian Mixture Filtering, Monthly Weather Review, 144, 781–800, 2016.

Mangoubi, O., Pillai, N. S., and Smith, A.: Does Hamiltonian Monte Carlo mix faster than a random walk on multimodal densities, arXiv:1808.03230v2, pp. 1–45, 2018.

Salter-Townshend, M. and Haslett, J.: Fast inversion of a flexible regression model for multivariate pollen counts data, Environmetrics, 23, 595–605, 2012.

Silverman, B.: Density Estimation for Statistics and Data Analysis, vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman & Hall / CRC, Boca Raton, 1986.

Tawn, N. G. and Roberts, G. O.: Accelerating Parallel Tempering: Quantile Tempering Algorithm (QuanTA), arXiv:1808.10415v1, pp. 1–39, 2018.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2018-87, 2018.

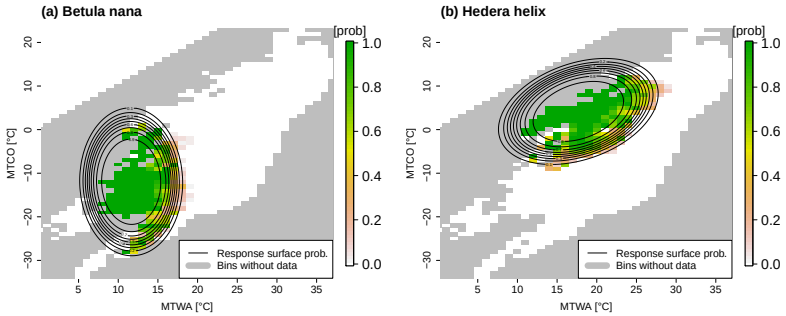

Figure 1: Response surfaces for Betula nana (a) and Hedera helix (b). The ratio of presence versus absence in the modern calibration data in each bin with at least one data point is shown in colours. Gray bins are bins without data. The response surfaces (probability of presence according to Eq. (7) in the original manuscript) are depicted by contour lines. In the climate space, combinations of MTWA and MTCO above a line at MTWA = MTCO cannot occur by definition. The white triangle in the upper left are artificial absence information added to account for this constraint, as described in the original manuscript.

Fig. 1.

Clim. Past Discuss.,
https://doi.org/10.5194/cp-2018-87-AC2, 2018

Climate of the Past Discussions

Interactive comment on "Combining a pollen synthesis and climate simulations for spatial reconstructions of European climate using Bayesian modelling" *by* Nils Weitzel et al.

Nils Weitzel et al.

nils.weitzel@uni-bonn.de

Received and published: 10 December 2018

We thank the referee for taking the time to review our discussion paper, and for helpful and interesting comments which will improve the quality of our manuscript. As a response to the referee's suggestions, we plan substantial changes of the manuscript. In the following, referee's comments (RC) are given in blue italicised text and followed by our respective responses (AR).

"I have some minor criticism of the presentation which often appears poorly-organised

to me. Specifically, many parameters and concepts are used prior to their definition and introduction which made the manuscript tough going for me."

We will go through the manuscript and try to rework Sect. 2 and 3 to improve the structure of the manuscript.

"A more fundamental concern I have is demonstrated in the results, which (unless I have misunderstood something) show a strong degree of instability and overconfidence. This is a common occurrence in situations where all members of a model ensemble are inadequate and a Bayesian framework is used which does not allow for this - for instance, by the explicit use of model inadequacy or model error terms. The z weight of 0.98 when all data are used is itself likely to indicate a problem, and this feeling is only strengthened by the fact that the weight on this model was essentially 0 when half the data were used, in which event two different models are preferred. What seems to be happening here is that the method is homing in on the model(s) for which the data are most likely, while in fact these data are incompatible with any model.
I've seen this in a variety of contexts and am sure this sort of phenomenon will be familiar to the authors. A trivial demonstration of the phenomenon can be given by considering a situation in which we have two models which generate data according to N(0,1) and N(10,1) respectively, whereas the data are in fact generated by N(5,1). A naive filtering which ignores the possibility of model inadequacy will assign essentially all weight to whichever model happens to be closer to the sample mean of the data set (which is of course close to, but not precisely, 5). As more data are sampled, the weight will jump randomly from one model to the other with extremely high confidence, rather than converging to the answer that the models are equiprobable but poor, which might be found if a reasonable model error term was considered. This issue seems to me to undermine the results and details of the application in a

fairly fundamental manner but I would hope that the authors could revise their method to adequately account for it, ie by accounting for model inadequacy in a formal manner."

We agree with the referee that the ensemble member weights in our model tend to degenerate towards the least wrong model and this degeneracy increases with increasing signal in the proxy data. This is a general problem of particle filter type models (or more general, Bayesian model selection problems). For that reason, we combine the particle part of our model with a part that is more similar to Kalman type filters by adjusting the ensemble members according to a prior covariance and the signal in the proxy data.

On the other hand, the strong changes of model weights between the joint reconstruction and the MTCO only reconstruction are a result of a different proxy data calibration and not of leaving out half the data. The proxy data in the MTCO only reconstruction is still the same but instead of calibrating it against a bivariate climate, it is only calibrated against MTCO. It is not surprising that different models perform better for winter than for summer climate. The reason that the joint reconstruction is dominated by the model performance for summer climate is a result of the higher signal to noise ratio for local MTWA reconstructions because most taxa are more sensitive to temperature during the growing season than in winter. We briefly report in Sect. 5.1 results from experiments where half of the proxy data is left out. In each of these results the same model than in the reconstruction with the full dataset is preferred, but because of the smaller signal in the proxy data the weights can be less degenerate (see Fig. at the end of this response). Moreover, the experiments show that the reconstructions are reasonably stable with respect to leaving out substantial parts of the data. We think that the issue of very different model weights between the joint reconstruction and the MTCO only reconstruction rather indicates that the joint reconstruction could be improved from different weights for MTWA and MTCO, even though this would reduce

the physical consistency of the estimates.

Nevertheless, we agree with the referee that our model produces overconfident estimates and an under-dispersed posterior distribution similar to many ensemble filtering models, and that this could be reduced by techniques to formally account for model inadequacy. The main reason for this behaviour is the small number of ensemble members leading to a small number of mixture kernels and a small number of spatial modes in the covariance matrix Σ. Accounting for model inadequacy in a physically reasonable way is a challenging issue and an active research area particularly in postprocessing applications. To reduce the underdispersion of our model and reduce reconstruction biases, we introduce two extensions of the model compared to the original manuscript.

First, we compare the proposed kernel model with a statistical model that facilitates a more flexible prior mean structure by mixing directly over $\omega_k \mu_k$, which was suggested by the other referee (see also response to Referee 1). This means that weighted averages of the ensemble member climatologies are used and the weights are adjusted according to the proxy data. Thereby, we increase the degrees of freedom but ensure that the spatial structures are still physically motivated. In particular, the model weights are less degenerate with this adjustment and the model can better cope with situations like the one described by the referee where the truth is located between two kernels.

Second, we compare the covariance matrix regularized by the glasso algorithm with a shrinkage matrix (Hannart and Naveau, 2014), where the empirical correlation matrix of the ensemble is combined with a Matérn type correlation matrix. This increases the spatial modes of the prior covariance matrix and estimating the shrinkage weight from the proxy data ensures that the amount of deviations

from the ensemble correlation matrix is in agreement with the proxy data. Initial results of identical twin experiments reveal that this increase of modes in the prior covariance matrix reduces the underdispersion of the posterior distribution significantly.

Combining ensemble filtering methods with additional techniques to account for model inadequacy should be an important part of future work on climate field reconstructions as a balance has to be found between underdispersion of the posterior distribution and overfitting to noisy proxy data by enhancing the degrees of freedom too much. Beyond our adjustments compared to the original manuscript, some directions that can be envisaged are the increase of permitted spatial structures in the prior mean by combining the climate simulation ensemble with patterns calculated from alternative physically motivated models, and the introduction of multiple shrinkage targets (Gray et al., 2018) in the prior covariance matrix which allows the proxy data to weight multiple spatial correlation modes. Identical twin experiments and cross-validation experiments with proxy compilations will be important techniques to understand the appropriateness of different modelling options as described below.

In the revised manuscript, we will extend the report on results from experiments with reduced data as a way to analyse the robustness of our reconstructions. In addition, we will include the two adjustments in the model described above in the methods section. Afterwards, we will add a section that compares these adjusted models with the original kernel method and the other alternative models described in the response to Referee 1. This section will include identical twin and cross-validation experiments as described below. Finally, we will put a focus in the discussion of future research strategies on additional methods to account for model inadequacy.

"Related to this, the Brier score analysis seems to indicate that the posterior is

generally closer to the data than the prior, but it does not indicate whether the posterior is valid in the sense of having reasonably calibrated uncertainties. Also, while the data are in taxa-space, the aim of the paper is actually to recreate a climate, so it is important that the validity of this estimate is assessed. Perhaps a cross-validation (in which one model is used as a target) could be tried."

We agree with the referee that ideally we want to evaluate the ability of our method to reconstruct climate. However, there are no direct observations of paleoclimate available such that evaluations against real observations have to be indirect. Therefore, evaluations with cross-validation experiments against indirect observations are a valuable tool to analyse the ability of reconstruction methods to reconstruct past climate. In particular, evaluating the reconstruction in taxa space by applying the forward model is a methodologically more stringent method than comparing reconstructions with other inferred quantities like local climate reconstructions. Therefore, for the evaluation of a reconstruction against observations, we do not see a way around evaluations against indirect data. This is also a recent way of model skill evaluation in weather forecasting.

The referee suggests to perform identical twin experiments, i.e. experiments in which one model is excluded from the ensemble as reference climatology and the goal is to reconstruct this reference climatology from the remaining models. This is a useful method to study the validity of the posterior distributions which is difficult to analyse in evaluations against observations. Therefore, we designed such experiments by simulating pseudo-proxies from the reference climatology using the proxy uncertainty structure from the local reconstructions, reconstructing the reference climatology from the pseudo-proxies with the model proposed in the original manuscript as well as alternative statistical models proposed by the referees, and analysing the skill of the corresponding posterior distributions to predict the reference climatology. In particular, this methodology is well-suited for studying potential underdispersion of the posterior

distributions and for comparing different statistical models.

Initial results from these experiments suggest that the models with the shrinkage matrix are substantially less under-dispersed than the models with the glasso optimized covariance matrices as a result of containing more spatial modes and therefore more degrees of freedom. Moreover, the kernel model that we proposed in the original manuscript performs similar to the purely Gaussian model and the mixture over $\omega_k \mu_k$ which were suggested by the first referee. Therefore, it seems like adding spatial modes to the covariance structure results in the biggest improvements of reconstruction skill.

In the revised manuscript, we will add a description of the design of identical twin experiments, and report on the results of these experiments. In addition, we will dedicate a section to the comparison of the different statistical models, the one that is proposed in the original manuscript as well as alternatives that were suggested by the referees, using identical twin experiments as well as cross-validations in taxa space.

"However, I am not convinced that the details of the application as presented here are adequate, nor, therefore, can I believe that their results are credible. I urge them to modify and extend their work in order to address this."

As described above, we will extend our work by addressing the comments and include the new results in the revised manuscript. We hope that this will increase the credibility of our results.

"As I mentioned, the ordering of some of the paper made it hard work for me.
The prior is introduced in Sect 2.2 and Fig 1 but only explained in Sect 3.2. The variables w and z are first mentioned in 3.1, but for their definitions we are referred forward to 3.2. It is only after substantial usage of the variables, right at the end of 3.2, that we are informed that z was only introduced as a help parameter, with its definition and explanation again pushed forward another two sections to 3.4 with the largely unrelated discussion of the transfer function (two words for this please, and note also that burn-in should be hyphenated) intervening in section 3.3. I would have found it far more straightforward to have had the variables explained as they were introduced.
The multinomial with n=1 would I think be better described as the categorical distribution.
In 3.4, "alternately" usually refers to two alternates, not a sequence. "Sequentially" might be better."

We thank the referee for pointing out several issues that hamper the readability of our manuscript. We will reorder Sect. 2 and 3 according to the referee's suggestions and incorporate the linguistic advises.

References

Gray, H., Leday, G. G., Vallejos, C. A., and Richardson, S.: Shrinkage estimation of large covariance matrices using multiple shrinkage targets, arXiv:1809.08024v1, pp. 1–32, 2018.
Hannart, A. and Naveau, P.: Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework, Journal of Multivariate Analysis, 131, 149–162, 2014.

Interactive comment on Clim. Past Discuss., https://doi.org/10.5194/cp-2018-87, 2018.

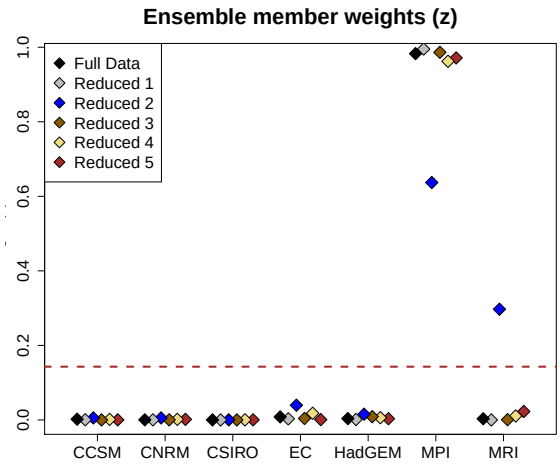

Figure 1: Posterior ensemble member weights (mean of z) of the the joint reconstruction and of five experiments where half of the proxy samples are left out. The black diamonds correspond to the reconstruction with the full dataset, whereas the coloured diamonds represent the experiments with reduced data. Prior weights (mean of z) are denoted by the dashed line.

Fig. 1.

Changes made to the manuscript

- We changed the title of the manuscript to emphasize that the proxy synthesis contains pollen and macrofossil data, and that the focus is on 'Bayesian filtering' instead of 'Bayesian modelling'.
- In the abstract, we added a sentence on the comparison study that we performed to address the referee's comments.
- We emphasize the joint use of pollen and macrofossils in the proxy synthesis more by referring to 'pollen and macrofossils' wherever applicable instead of just 'pollen'.
- We changed 'transferfunction' into transfer function' and 'burnin' into 'burn-in' throughout the manuscript.
- We added a description of the strengths of macrofossils in the introduction.
- We reordered Sect. 2 and 3 to improve readability and motivation of the statistical modelling choices.
- In Sect. 2.1, we added more information on the role of macrofossils in the proxy synthesis. In addition, we added a paragraph on the vegetation and climate data used to calibrate the transfer functions.
- In Sect. 2.2, we removed references to the prior distribution, as it is not defined before Sect. 3.
- We shortened Sect. 3.1 to a general introduction of the Bayesian framework, such that no variables are used that are not defined until later sections.
- We changed the order of the description of the transfer function (now Sect. 3.2) and the role of the climate simulation ensemble in the Bayesian framework (now Sect. 3.3).
- In Sect. 3.2 (Transfer function), we enhanced the motivation and description of assumptions of the probabilistic indicator taxa model. We discuss the inconsistency between using present and absent taxa in the calibration but only occurring taxa in the reconstruction, mention detection probabilities as a way to overcome this issue at least for pollen data, and highlight the challenges of estimating detection probabilities, which make the use of them not feasible in this study.
- In Sect. 3.3 (Process stage), we add two more ways to formulate the process stage (as Gaussian model or as regression model, in addition to the kernel model of the previous manuscript version) and another approach for the spatial covariance structure (the shrinkage approach, in addition to the glasso approach of the previous manuscript version). We enhance the motivation of all the models, and discuss advantages and disadvantages of the different modelling choices.
- In order to improve the readability of the manuscript, some technical parts of the inference strategy (Sect. 3.4) are moved to the appendices.

- Sect. 3.5 (Assessing the added value of reconstructions) has been removed and the different measures to evaluate the reconstruction skill are described when they are first used in Sect. 4.1 and Sect. 4.2.3.
- Sect. 4.1 (Comparison of different process stage frameworks) is added to the previous manuscript version. It describes results from comparing the different process stage models described in Sect. 3.3 (Process stage) with identical twin and cross-validation experiments.
- The results from the spatial reconstruction of European mid-Holocene climate (Sect. 4.2 in the revised manuscript) have been updated to reconstructions with the regression model using the shrinkage covariance matrix (instead of the kernel model with glasso covariance matrix), since this model has produced the best results in the comparison study of Sect. 4.1.
- Sect. 4.2.1 (Comparison of unconstrained PMIP3 ensemble and posterior distribution) has been added instead of the previous section on 'Posterior ensemble member weights' to emphasize stronger on the differences between the unconstrained PMIP3 ensemble and the posterior distribution instead of just the ensemble member weights.
- The section on 'Sensitivity with respect to the glasso penalty parameter' has been moved to the supplement.
- Sect. 5.1 (Robustness of the reconstruction) focuses more on the experiments with reduced proxy synthesis and the sensitivity with respect to the process stage formulation.
- Sect. 5.2 (Comparison with previous reconstructions) has been updated to account for the updated reconstruction results.
- In Sect. 5.3 (Climate model inadequacy and process stage structure), we discuss strategies to account for climate model inadequacy, as well as alternative ways to formulate the process stage.
- The section 'Towards global and spatio-temporal reconstructions' has been removed. Parts of it are moved to the section on 'Climate model inadequacy and process stage structure'.
- In the 'Conclusions', we added a paragraph on the results of the process stage model comparison study.
- We shortened some paragraphs that discussed possible directions of future research.
- We added a paragraph on 'Code and data availability'.
- We added appendices on the 'Determination of glasso penalty parameter', the ' Shrinkage target matrix', and the 'Metropolis coupled Markov chain Monte Carlo algorithm'. The description of the 'Full conditional distributions' has been updated to account for the added process stage models.
- We updated the Acknowledgments.
- Fig. 1 does not refer to the prior distribution anymore.
- Fig. 2 has been updated to account for the added process stage models.
- Fig. 3 has been updated to better visualize the calibration data and the estimated response surfaces.

- Fig. 5 and Fig. 6 have been added to visualize results from the process stage model comparison study.
- Fig. 7 to 10 are updated to present the updated results from Sect. 4.2.
- Table 1 has been shortened.
- Table 2 has been added to summarize results from the process stage model comparison study.
- Table 3 has been updated to the new results presented in Sect. 4.2.
- A Bitbucket repository has been created to improve reproducibility of our results. The link to the repository is given in 'Code and data availability'.
- We added a Supplement with pseudo-code for the MCMC and MC^3 algorithm, additional figures, results from spatial reconstructions with the other process stage models, and results from studying the sensitivity with respect to the glasso penalty parameter.
- Numerous small changes to improve clarity and readability of the manuscript are not listed here explicitly, but can be seen in the marked up version of the manuscript.

Combining a pollen ~~synthesis~~ and macrofossil synthesis with climate simulations for spatial reconstructions of European climate using Bayesian ~~modelling~~filtering

Nils Weitzel[1,2], Andreas Hense[1], and Christian Ohlwein[1]

[1]Institut für Geowissenschaften und Meteorologie, Rheinische Friedrich-Wilhelms-Universität Bonn, Auf dem Hügel 20, 53121 Bonn, Germany
[2]Institut für Umweltphysik, Ruprecht-Karls-Universität Heidelberg, Im Neuenheimer Feld 229, 69120 Heidelberg, Germany

Correspondence: Nils Weitzel (nils.weitzel@uni-bonn.de)

Abstract. Probabilistic spatial reconstructions of past climate states are valuable to quantitatively study the climate system under different forcing conditions because they combine the information contained in a proxy synthesis in a comprehensible product. Unfortunately, they are subject to a complex uncertainty structure due to complicated proxy-climate relations and sparse data, which makes interpolation between samples difficult. Bayesian hierarchical models feature promising properties to handle these issues like the possibility to include multiple sources of information and to quantify uncertainties in a statistically rigorous way.

We present a Bayesian framework that combines a network of pollen and macrofossil samples with a spatial prior distribution estimated from a multi-model ensemble of climate simulations. The use of climate simulation output aims at a physically reasonable spatial interpolation of proxy data on a regional scale. To transfer the pollen data into (local) climate information, we ~~apply~~ invert a forward version of the probabilistic indicator taxa model. The Bayesian inference is performed using Markov chain Monte Carlo methods following a Metropolis-within-Gibbs strategy.

~~We~~ Different ways to incorporate the climate simulations in the Bayesian framework are compared using identical twin and cross-validation experiments. Then, we reconstruct mean temperature of the warmest and mean temperature of the coldest month during the mid-Holocene in Europe using a published pollen and macrofossil synthesis in combination with the Paleoclimate Modelling Intercomparison Project Phase III mid-Holocene ensemble. The output of our Bayesian model is a spatially distributed probability distribution that facilitates quantitative analyses which account for uncertainties. ~~Our reconstruction performs well in cross-validation experiments and shows a reasonable degree of spatial smoothing.~~

1 Introduction

Spatial or climate field reconstructions of past near surface climate states combine information from proxy samples, which are mostly localized, with a model for interpolation between those samples. They are valuable for comparisons of the state of the climate system under different external forcing conditions, because they produce a comprehensible product containing the joint information in a proxy synthesis. Thereby, spatial reconstructions are more suitable for many quantitative analyses

of past climate than individual proxy records. Unfortunately, spatial reconstructions are subject to a complex uncertainty structure due to uncertainties in the proxy-climate relation and the sparseness of available proxy data which leads to additional interpolation uncertainties. Therefore, a meaningful reconstruction has to include these uncertainties (Tingley et al., 2012). A natural way to represent uncertainties in the proxy-climate relation are ~~so called probabilistic transferfunctions~~ so-called probabilistic transfer functions (Ohlwein and Wahl, 2012). To account for the uncertainties due to sparseness of proxy data, we suggest the use of stochastic interpolation techniques. Most standard geostatistical methods like kriging or Gaussian modelling with Matérn covariances are designed for interpolation in data rich situations, while in paleoclimatology we deal with sparse data. Therefore, their direct application to paleo situations is not suitable. Instead, we propose to use interpolation schemes that contain additional physical knowledge, such that the resulting product combines the information from a proxy network in a physically reasonable way (Gebhardt et al., 2008). ~~As will be shown, our~~ Our approach can in addition be used for structural extrapolation of the proxy data.

We use Bayesian ~~hierarchical modelling~~ statistics to combine the two modules mentioned above: The (local) proxy-climate relation and spatial interpolation. The Bayesian framework allows the combination of multiple data types. In our case, these are pollen and macrofossil records to constrain the local climate, and climate simulations, which produce physically consistent spatial fields for a given set of large scale external forcings. In addition, our framework accounts for several sources of uncertainty in a statistically rigorous way by estimating and inferring a multivariate probability distribution, the so-called posterior distribution (Gelman et al., 2013).

Pollen are the terrestrial proxy with the highest spatial coverage (Bradley, 2015), and there is a long tradition of using them for inferring past climate by applying statistical ~~transferfunctions~~ transfer functions (Birks et al., 2010). In recent years, several traditional ~~transferfunctions~~ transfer functions like indicator taxa, modern analogues, and weighted averaging have been translated to Bayesian frameworks (e.g., Kühl et al., 2002; Haslett et al., 2006; Holden et al., 2017). Pollen records contain information on the local climate during a time slice, where the spatial scale is constrained by the influx domain of horizontal pollen transport. Typically, macrofossils have a higher taxonomic resolution than pollen such that the climatic niche of the occurring taxa can be better constrained than with pollen alone (Bradley, 2015). Equilibrium simulations with earth system models (ESMs) produce a physically consistent estimate of the atmospheric and oceanic circulation and the regional energy balance given a set of forcings (boundary conditions). Important boundary conditions, for which information are available from proxy data and physical models, are insolation determined by the earth orbital parameters, greenhouse gas concentrations, ice sheet configurations, and land-sea masks. We use an ensemble of simulations from different ESMs to estimate a prior distribution, which contains a wide range of physically reasonable climate states. The combination of these two sources of information can be interpreted as a downscaling of forcing conditions via ESMs and an upscaling of local information contained in pollen records via spatial covariance matrices. The result is a spatially distributed and physically reasonable probabilistic climate reconstruction on continental domains.

We apply our framework to a mid-Holocene (MH, around 6ka) example for two reasons. First, compared with other time slices before the common era, the MH has a high proxy data coverage, particularly for Europe. Therefore, we can use raw pollen and macrofossil data with a sparse but relatively uniform spatial coverage over Europe as input for probabilistic

~~transferfunctions~~transfer functions, while still having other reconstructions available, that can be compared with our results. Second, a multi-model ensemble of climate simulations with boundary conditions adjusted to the MH was produced in the Paleoclimate Modelling Intercomparison Project Phase III (PMIP3) project (Braconnot et al., 2011). This ensemble is used to estimate the spatial prior distribution. The posterior distribution, which we estimate, is a multivariate probability distribution, with marginal distributions for each grid box, as well as spatial correlations and correlations between two climate variables, the mean temperature of the warmest month (MTWA) and the mean temperature of the coldest month (MTCO). For further analyses, we create samples from this distribution, such that each sample is an equally probable estimate of the bivariate spatial field. In the context of temporal reconstructions these samples were called "climate histories" by Parnell et al. (2016). From the samples, quantitative properties of the climate state during the MH, which account for uncertainties, are computed. In addition, our framework can be used to compare the model-data mismatch of multiple ESMs, to analyse the consistency of a given proxy network, and to help in the identification of potential outliers.

This work is related to several concepts that were developed for applications in paleoclimatology. In recent years, several authors constructed Bayesian hierarchical models (BHMs) for paleoclimate reconstructions: Tingley and Huybers (2010) introduced a spatio-temporal BHM for reconstructions of the last millennium with an underlying structure that is stationary, linear, and Gaussian. Other authors developed temporal ~~(Li et al., 2010; Parnell et al., 2015)~~ (Parnell et al., 2015) or small-scale spatio-temporal BHMs (Holmström et al., 2015). All of these approaches differ from our model in being purely proxy data driven. Additional information on orbital configurations were incorporated by Gebhardt et al. (2008) and Simonis et al. (2012) via an advection-diffusion model which is combined with proxy data using a variational inference approach. Li et al. (2010) included information on solar, greenhouse gas, and volcanic forcing for spatially averaged reconstructions of the last millennium via linear regression. Annan and Hargreaves (2013) combined Paleoclimate Modelling Intercomparison Project Phase II (PMIP2) simulations with the syntheses of Bartlein et al. (2011) and MARGO Project Members (2009) in a multi-linear regression model. We build on these approaches by incorporating fields that are simulated from a set of MH forcing conditions in a fully Bayesian framework. A different approach to combine proxy data and climate simulations for spatio-temporal reconstructions of the common era was developed by Steiger et al. (2014) and Dee et al. (2016) using ~~so called~~ so-called off-line data assimilation methods. They apply an ensemble Kalman filter, where the observation operators are forward models for proxy data, and the prior covariance is estimated from a database of transient climate simulations. Our purely spatial reconstructions can be interpreted as an off-line data assimilation with only one time step. This reduced dimensionality permits the exploration of the full posterior distribution despite incorporating non-linear and non-Gaussian elements in the observation operator and ~~a multi-modal spatial prior distribution estimated from a multi-model ensemble~~the spatial interpolation scheme.

The structure of the paper is as follows. In Sect. 2, we describe the pollen synthesis and climate simulations which we use. This is followed by a detailed description of our proposed Bayesian framework in Sect. 3. Results from ~~applying our methodology to the data~~ a comparison study of different ways to incorporate the climate simulations in the Bayesian framework and from our reconstruction of the European MH climate are presented in Sect. 4. Finally, we discuss and summarize our methodology and results in Sect. 5 and 6.

2 Data

2.1 Proxy and calibration data

The pollen and macrofossil synthesis, that we use in this study, stems from Simonis et al. (2012) as part of the European Science Foundation project DECVeg (Dynamic European Climate-Vegetation impacts and interactions). ~~In the following, we refer only to pollen instead of pollen and macrofossils for linguistic simplicity.~~ Out of the four time slices (6ka, 8ka, 12ka, 13ka), which were compiled, we only use the 6ka dataset because there is no ensemble of climate simulations available for the other three time slices. For 50 paleosites, ~~presence and absence information for~~ information on the occurrence of taxa is provided. 59 taxa ~~are provided. The data is already statistically preprocessed to remove co-occurring taxa that lead to an underestimation of uncertainty (Kühl et al., 2002; Gebhardt et al., 2008). The synthesis is optimised for the use in local climate reconstructions with different versions of the probabilistic indicator taxa model (PITM, originally called "pdf method"; Kühl et al., 2002), and is therefore well-suited for the implementation in a BHM that combines probabilistic transferfunctions with stochastic interpolation methods. The PITM model is described in detail in Sect. ??~~occur at least at one site. For some sites, information from very nearby sites are combined into a joint sample. 15 of the sites combine macrofossil and pollen information, three samples contain just macrofossil data, and for 32 sites only pollen data is available. In general, the macrofossil data provides more detailed taxonomic information than pollen. Because pollen is more prevalent than macrofossil data, pollen samples are included at sites with macrofossils as well as from additional sites to provide a broader spatial picture of the European vegetation at the MH.

The 50 paleosites are sparsely but relatively uniformly distributed over Europe. Their locations are delimited by 6.5° W, 26.5° E, 37.5° N and 69.5° N. Compared with other recent syntheses like Bartlein et al. (2011), less records are included due to high quality control criteria. The raw pollen or macrofossil data and radiocarbon measurements, from which at least one was supposed to be close to 6ka, had to be available to recalculate age-depth models and ensure the use of calibrated radiocarbon dates as common time scale. Each site is assigned to the corresponding cell of a 2° by 2° grid which we use for our reconstructions. The locations of the proxy samples are depicted by black dots in Fig. 1. The full list of sites included in the synthesis can be found in Simonis et al. (2012). The list of taxa, which ~~remain for each site after statistical preprocessing, is given~~ occur at the sites, is published in Simonis (2009).

Modern climate and vegetation data is used for the calibration of the transfer functions. The climate data is computed from the University of East Anglia Climatic Research Unit (CRU) 1961 to 1990 reference climatology (CRU TS v.4.01, Harris et al., 2014; Harris an). The vegetation data stems from digitized vegetation maps (Schölzel et al., 2002). The regions that are used for the transfer function calibration were determined individually for each taxa by pollen experts (Kühl et al., 2007). The number of calibration sites varies between 14.543 and 28.844, depending on the taxa.

2.2 Climate simulations

We use a multi-model ensemble of climate simulations which were run within PMIP3 with forcings adjusted to the MH. This includes changed orbital configurations and greenhouse gas concentrations (Braconnot et al., 2011). The ensemble contains

all available MH simulations in the PMIP3 database (downloaded from the German Climate Computing Center (DKRZ) long term archive, available under https://cera-www.dkrz.de), which have a grid spacing of at ~~most~~ least 2°. This constraint, which retains only the models with the smallest grid spacings, is chosen to better match the resolutions of ~~pollen~~ proxy samples and simulations. The condition results in using seven model runs performed with the CCSM4, CNRM-CM5, CSIRO-Mk2-6-0, EC-Earth-2-2, HadGEM2-CC, MPI-ESM-P, and MRI-CGCM3. Properties of the included simulations are given in Table 1. The ensemble is a multi-model ensemble with common boundary conditions. The models are run to an equilibrium state (spin-up), followed by typically around 100 simulated years to minimize noise due to high frequency internal variability. Therefore, the differences within the ensemble can be interpreted as modelling uncertainties (epistemic uncertainty).

The mean summer climate expressed as MTWA (Fig. 1a) from the MH ensemble is warmer than the ~~University of East Anglia Climatic Research Unit (CRU) 1961 to 1990~~ CRU reference climatology (CRU TS v.4.01 over land, Harris et al. (2014), Harris and Jones (2017), and HadCRUT absolute over sea, Jones et al. (1999)) in large parts of Europe, especially eastern Europe and the Norwegian Sea. These areas coincide predominantly with areas of large ensemble spreads, expressed as ~~the size of~~ point-wise ~~90% credible intervals (CIs) of the prior distribution~~ empirical standard deviations in Fig. 1c. The ~~construction of the prior distribution is described in Sect. ??. The spread increases up to 10K~~ standard deviations increase up to 4 K in some areas of southern and eastern Europe, which might originate from varying change patterns of the general circulation over Europe in the models. In contrast, the MH mean winter climate measured by MTCO in Fig. 1b shows a more dispersed structure with cooling in Fennoscandia, warming in the Mediterranean and Balkan peninsula, and mixed patterns in western and central Europe. The ensemble spread is predominantly small (Fig. 1d), but increases towards northern Europe with very large inter-model differences for the Norwegian Sea and eastern Fennoscandia. ~~Most of the grid box anomalies for MTWA as well as MTCO are not significant on a 5% level. Here, a positive anomaly is called significant if the probability of the prior temperature to exceed the reference climatology is at least 0.95. Significant negative anomalies are defined accordingly. The significance estimates are calculated point-wise.~~

2.3 Reconstruction variables

The spatial distribution of taxa is limited by three climatic factors: Temperature during growing season and in winter, and moisture availability (Huntley, 2012). Therefore, these three factors, translated into quantitative variables, are important for climate reconstructions from pollen. For large parts of Europe, it was shown in Simonis (2009) that the ~~PITM model (see Sect. ??)~~ pollen and macrofossil synthesis is well-suited for joint reconstructions of July and January temperature as measures for the warmth of growing season and cold of winter, because at least one of these two variables is a limiting factor for most taxa growing in the mid and high latitudes of Eurasia during the Holocene. Instead, testing various climate variables as indicators for moisture availability was less promising since ~~the~~ moisture availability is rarely a limiting factor for European taxa (Simonis, 2009). Hence, in this study, we choose MTWA and MTCO as the target variables for our climate reconstructions. This is a compromise between variables that are bioclimatically meaningful and variables for which accurate data is available to calibrate the ~~transferfunctions~~ transfer functions against modern vegetation and climate data. In the mid to high latitudes, MTWA and MTCO are highly correlated with July and January temperature, respectively, which is why they are also described

as "functionally equivalent" (Bartlein et al., 2011).

To calculate MTWA and MTCO from time series of monthly averages, the data is first interpolated to the desired spatial grid. Then, for each hydrological year (October to September), the warmest and coldest month are extracted. We choose the hydrological instead of the calendar year to ensure that the months are taken from connected seasons. Afterwards, the climatological mean is calculated by averaging over the values for each year.

3 Methods

3.1 Bayesian framework

We use Bayesian statistics to combine a network of pollen samples with an ensemble of PMIP3 simulations because in this approach each source of information has an associated uncertainty that is naturally included in the inference process. In this section, we specify the quantities that are combined in our reconstruction, and describe the inference algorithm that is used to create the results presented below~~as well as the statistical measures which we use to assess the reconstructions~~.

In the following, we denote fossil pollen and macrofossil data by P_p, past climate by C_p, modern vegetation and climate data for the calibration of ~~transferfunctions~~ transfer functions by P_m and C_m, respectively, and additional model parameters by ~~$\Theta := (\omega, z, \theta)$. Here, ω and z represent weights of the PMIP3 ensemble members as defined in Sect. ?? and ??, and θ are transferfunction parameters, which are specified in Sect. ??~~Θ. We are interested in the conditional distribution of C_p and Θ given fossil pollen and macrofossil, modern vegetation, and modern climate data, i.e. we want to estimate the posterior distribution $\mathbb{P}(C_p, \Theta \,|\, P_p, C_m, P_m)$.

Applying Bayes' theorem to $\mathbb{P}(C_p, \Theta \,|\, P_p, C_m, P_m)$ (in the following, we omit normalizing constants), we get:

$$\underbrace{\mathbb{P}(C_p, \Theta \,|\, P_p, C_m, P_m)}_{\text{Posterior}} \propto \underbrace{\mathbb{P}(P_p, P_m \,|\, C_p, C_m, \Theta)}_{\text{Data Stage / Likelihood}} \cdot \underbrace{\mathbb{P}(C_p, C_m \,|\, \Theta)}_{\text{Process Stage}} \cdot \underbrace{\mathbb{P}(\Theta)}_{\text{Prior Stage}} . \tag{1}$$

Following Tingley and Huybers (2010), we call $\mathbb{P}(P_p, P_m \,|\, C_p, C_m, \Theta)$ the data stage, $\mathbb{P}(C_p, C_m \,|\, \Theta)$ the process stage, and $\mathbb{P}(\Theta)$ the prior stage. ~~The structure of the Bayesian model can be expressed by a directed acyclic graph as shown in Fig. 2. In the graph, each node represents a variable and the arrows indicate dependences of variables.~~

~~We assume that the model weights ω and z are a priori independent of the transferfunction parameters θ, and that the data stage is conditionally independent of ω and z given C_p. Furthermore, by construction, P_m and C_m only contribute to the reconstruction via the transferfunction parameters, i.e. they are assumed to be independent of all other quantities. Hence, we can rewrite Eq. (??) and get~~

$$\mathbb{P}(C_p, \Theta \,|\, P_p, C_m, P_m) \propto \mathbb{P}(P_m \,|\, C_m, \theta) \; \mathbb{P}(P_p \,|\, C_p, \theta) \; \mathbb{P}(C_p \,|\, z) \; \mathbb{P}(z \,|\, \omega) \; \mathbb{P}(\omega) \; \mathbb{P}(\theta).$$

In paleoclimatology, the data stage is traditionally called ~~transferfunction~~transfer function, which in our case is formulated in a forward way. It probabilistically models the proxy data given climate variables and is ~~assumed to act locally, i.e. given the climate at location x, a pollen sample at x is conditionally independent of the climate and proxy data at all other locations. A more rigorous formulation is given~~ described in detail in Sect. ~~??.~~

3.2. The process stage stochastically interpolates the local climate information from the proxy data to a spatial domain ~~. In our model, this interpolation is performed using a mixture of Gaussian distributions calculated from the ensemble of PMIP3 simulations, which~~ and is described in ~~detail in Sect. ??. Note that our approach is not restricted to interpolation between proxy samples, but allows structural extrapolation through the eigenvectors of the covariance matrix and the weights of the ensemble members. Hence, we can infer climate beyond the domain of the proxy synthesis~~Sect. 3.3. The prior stage defines prior distributions for the model parameters Θ. These prior distributions are necessary to get a closed Bayesian model which ensures that the posterior is a valid ~~probality~~ probability distribution (Gelman et al., 2013).

3.2 ~~Preprocessing of PMIP3 simulations~~

~~We model the process stage in Eq. (??) and the prior distributions for ω and z based on the ensemble of PMIP3 simulations described in Sect. ?? following the ensemble kernel dressing approach of Schölzel and Hense (2011). Six steps are applied successively: All simulations are projected to a 2° by 2° grid using bilinear interpolation. We calculate the climatological means for MTWA and MTCO from the full timeseries of each simulation as described in Sect. ??. This minimizes high frequency internal variability which cannot be resolved by the pollen data because it is a time integrated quantity (Annan and Hargreaves, 2013) . The resulting climatologies from step two define the means μ_k, $k = 1, ..., K$, of the mixture components, where K is the number of ensemble members. Each mixture component is assumed to be multivariate Gaussian. The dimension N of each Gaussian vector is either the number of grid boxes in the case of single variable reconstructions or twice the number of grid boxes in the case of joint reconstructions of MTWA and MTCO. Following the kernel approach in Silverman (1986), we calculate the covariance matrix $\tilde{\Sigma}_{\text{prior}}$ as the empirical covariance of the inter-model differences scaled by the Silverman factor~~

$$f := \left(\frac{4}{K \cdot (N+2)} \right)^{\frac{2}{N+4}}.$$

~~Hence, $\tilde{\Sigma}_{\text{prior}}$ is given by~~

$$\tilde{\Sigma}_{\text{prior}} = f \cdot \frac{1}{K-1} \sum_{k=1}^{K} (\mu_k - \bar{\mu}) (\mu_k - \bar{\mu})^t,$$

~~where $\bar{\mu}$ is the mean over all mixture components, and the superscript t denotes the matrix transpose. To get a non-singular covariance matrix and avoid spurious correlations, we apply the graphical lasso algorithm (Friedman et al. (2008), implemented in the R-package glasso, https://cran.r-project.org/package=glasso) to $\tilde{\Sigma}_{\text{prior}}$. This algorithm approximates the precision matrix (inverse covariance) by a positive definite, symmetric, and sparse matrix $\Sigma_{\text{prior}}^{-1}$. Therefore, Σ_{prior} is a valid N-dimensional covariance matrix that we use as covariance matrix of the mixture components. The sparseness of the precision matrix has the additional advantage that computationally efficient Gaussian Markov random field techniques (Rue and Held, 2005) can be used, which reduces the computational burden significantly.Glasso maximizes the penalized log-likelihood~~

$$\log\det \Sigma_{\text{prior}}^{-1} - \text{trace}(\tilde{\Sigma}_{\text{prior}} \Sigma_{\text{prior}}^{-1}) - \rho \|\Sigma_{\text{prior}}^{-1}\|_1,$$

where ρ is the penalty parameter, $\|\cdot\|_1$ is the vector L_1-norm, and the first two terms are the Gaussian log-likelihood. To determine ρ, we first recognize that for values smaller than $\rho = 0.3$ the resulting matrices become numerically unstable in our application due to the small ensemble size. Five values for ρ were tested: 0.3, 0.5, 0.7To further structure the framework, we split the model parameters Θ into θ, which are parameters associated with the data stage, and ϑ, which are parameters that influence the process stage. We assume that θ and ϑ are a priori independent of each other and that the data stage is conditionally independent of ϑ given C_p. Furthermore, by construction, P_m and C_m only contribute to the reconstruction via the transfer function parameters, 1.0, and 2.0. Larger values lead to sparser precision matrices and therefore to smaller spatial correlations. This in turn means that the local information from the proxy data is spread less into space. For each of the five parameters we perform a cross-validation experiment and compare the resulting Brier scores (see Sect. **??**). While the smallest penalty parametershave the best mean Brier scores, the differences are generally small (Fig.**??**). On the other hand, the influence of the penalty term in i.e. they are assumed to be independent of all other quantities. Hence, we can rewrite Eq. (**??**) on the overall regression increases from 79.5% for $\rho = 0.3$ to 98.5% for $\rho = 2.0$. Based on these diagnostics, we choose $\rho = 0.3$ for the reconstructions in Sect. **??** to get a numerically stable covariance which performs at least as good as other choices of ρ in cross-validations, and is comparably little influenced by the penalty term. The sensitivity of the reconstruction with respect to ρ is further studied in Sect. **??**. Note that for other applications of our framework different values of ρ can produce the best results. The result of the previous steps is a mixture distribution

$$\mathbb{P}(C_p \,|\, \omega, \mu_1, ..., \mu_K, \Sigma_{\text{prior}}) = \sum_{k=1}^{K} \omega_k \, \mathcal{N}(C_p \,|\, \mu_k, \Sigma_{\text{prior}}),$$

where $\omega = (\omega_1, ..., \omega_K)$ are the prior weights. We define a Dirichlet distributed prior for ω with parameter $\frac{1}{2}$ for each of the K components. This corresponds to an objective prior (Jeffrey's prior; Gelman et al., 2013) in the Dirichlet-multinomial model. The variable z is added as a help parameter, which simplifies the inference algorithm as described in Sect. **??**. 1) and get

$$\mathbb{P}(C_p, \Theta \,|\, P_p, C_m, P_m) \quad \propto \quad \underbrace{\mathbb{P}(P_m \,|\, C_m, \theta)}_{\text{Calibration stage}} \quad \underbrace{\mathbb{P}(P_p \,|\, C_p, \theta)}_{\text{Observation stage}} \quad \mathbb{P}(C_p \,|\, \vartheta) \quad \mathbb{P}(\theta) \quad \mathbb{P}(\vartheta). \tag{2}$$

Here, $\mathbb{P}(P_m \,|\, C_m, \theta)$ is called calibration stage and $\mathbb{P}(P_p \,|\, C_p, \theta)$ is called observation stage (Parnell et al., 2015). The structure of the Bayesian model can be expressed by a directed acyclic graph as shown in Fig. 2. In the graph, each node represents a variable and the arrows indicate dependences of variables.

3.2 TransferfunctionsTransfer function

The Bayesian model uses probabilistic transferfunctions to model the pollen-climate relationtransfer functions to model proxy data, in our case occurrence information on taxa, given climate and transfer function parameters for which prior distributions have to be defined. From all the terms in Eq. (**??**2), the calibration layer $(P_m \,|\, C_m, \theta)$stage, the observation layer $\mathbb{P}(P_p \,|\, C_p, , \theta)$stage, and the prior distribution of the transferfunction transfer function parameters $\mathbb{P}(\theta)$ are related to the transferfunction (Parnell et al., 2015) .

~~To reconstruct climate from the Simonis et al. (2012) synthesis, we need a transferfunction that converts presence-absence information on taxa to climate variables~~transfer function. As described above, our main reconstruction target is the bivariate climate $C = (C_1, C_2)$, where C_1 is MTWA and C_2 is MTCO. ~~Previously, reconstructions~~

To reconstruct climate from the Simonis et al. (2012) ~~dataset were performed using PITM, orginially named "pdf method " (Kühl et al., 2002)~~synthesis, the probabilistic indicator taxa method (PITM) model is used, which is ~~an extension of the classical indicator species method (Iversen, 1944) . PITM fits probability distributions to presence-absence maps of taxa (Schölzel et al., 2002), which are plotted in the bivariate climate space . The reference climatology for this mapping is derived from the CRU TS v.4.01 1961 to 1990 monthly averages following the definitions of MTWA and MTCO described in Sect. **??**~~a well established transfer function to quantitatively constrain past climate states by occurrence information on taxa extracted from pollen and macrofossil samples. It uses taxa which are sensitive to MTWA and MTCO and determines the climatic niche, where they occur, by fitting response functions. The classical indicator taxa method (Iversen, 1944) estimates binary limits, e.g. a taxa occurs above a certain temperature but not below it. PITM, also named pdf method in the literature (Kühl et al., 2002) , is an extension of this method where probability distributions are fitted to acknowledge that most taxa have a preferred climate space but the transitions between climates where they usually occur and those where they do not grow is soft. Initially, Gaussian distributions were used for calibration (Kühl et al., 2002) against vegetation maps (Schölzel et al., 2002). Later, the model was extended to mixtures of Gaussians (Gebhardt et al., 2008) ~~. Taxa are treated as conditionally independent given climate. A statistical pre-selection of taxa, which are present in a sample, is applied to avoid over-fitting originating from violations of the independence assumption between taxa, i.e. due to co-occurrence of taxa (Kühl et al., 2002). This procedure uses the Mahalanobis distance (Mahalanobis, 1936) between the fitted distributions~~and quadratic logistic regression (Stolzenberger, 2011, 2017).

We ~~reformulate PITM in a forward way using quadratic logistic regression similar to Stolzenberger (2017)~~ integrate the forward formulation of PITM from Stolzenberger (2017) in our Bayesian framework. For each taxa, we fit a quadratic logistic regression model (response function) describing the probability of ~~observing the taxa~~ taxa occurrence for a given value of C. The idea of using quadratic logistic regression stems from the BIOMOD (BIOdiversity MODelling) software which is a model to predict species distributions (Thuiller, 2003). The regression for taxa T contains linear and quadratic terms for each of the climate variables as well as an interaction term:

$$\mathbb{P}(T = 1 \,|\, C = (C_1, C_2)) \;=\; \text{logit}\left(\beta_1^T + \beta_2^T C_1 + \beta_3^T C_2 + \beta_4^T C_1 C_2 + \beta_5^T C_1^2 + \beta_6^T C_2^2\right). \tag{3}$$

Here, logit denotes the logistic function and $\beta_1^T, ..., \beta_6^T$ are regression coefficients. This regression leads to a unimodal response function which is anisotropic but has two symmetry axes, as can be seen for dwarf birch (Betula nana) and European ivy (Hedera helix) in Fig. 3.

To fit response functions, vegetation data is used instead of modern pollen or macrofossil data, because it contains more accurate information on the occurrence of a taxa on the spatial scales of interest compared to modern pollen samples. Moreover, the nature of macrofossils makes it impossible to create a modern calibration dataset that contains pollen and macrofossils. The disadvantage of using vegetation data for the calibration is that the probability of presence of a taxa is only valid in vegetation

space on the spatial scale taken for the training data but not in the pollen or macrofossil space, where an absence of a taxa in a pollen or macrofossil sample can have multiple non-climatic reasons like local plant competition or pollen transport effects, as well as local climatic effects below the resolution of our reconstruction such that the taxa did not grow in the immediate surrounding of the core location.

For the calibration against the modern dataset, we use presence (T=1) as well as absence (T=0) information ~~of~~ on the taxa which can be justified by assuming that the vegetation maps contain accurate information on ~~the presence or absence of taxa. On the other hand, for the fossil pollen samples, we do not include absent taxa in the reconstruction, as the absence of a taxain a pollen assemblage can have multiple non-climatic reasons like changing biologic competition or pollen transport effects (Gebhardt et al., 2003)~~presence as well as absence of taxa. From the definition given in Sect. ~~??~~2.3, it follows that at any location MTWA is larger or equal than MTCO. Formally incorporating this constraint in the inference leads to a non-linear condition on the regression parameters, which is very hard to implement. Therefore, we choose the more practical way of adding artificial absence information for combinations of MTWA and MTCO such that MTCO $>$ MTWA. While this leads to ~~transferfunctions~~transfer functions, which do not preclude reconstructions of MTCO values larger than MTWA, it is at least very improbable. To apply the response functions for individual taxa to a set of proxy data, we assume that proxy samples $P(s)$, where $s = 1, ..., S$ subscripts the proxy samples, are conditionally independent given a climate field and that, conditioned on $C(x_s)$, where x_s is the location of the s-th sample, $P(s)$ is independent of the climate at all other locations. This leads to the following probabilistic model for the set of modern vegetation samples

$$\mathbb{P}(P_m \,|\, C_m, \theta) \;=\; \prod_{s=1}^{S_m} \prod_{T \in T(P)} \mathbb{P}\left(P_m^T(s) \,|\, C_m(x_s), \beta_1^T, ..., \beta_6^T\right). \tag{4}$$

Here, $P_m^T(s)$ is the presence or absence of taxa T in the s-th calibration sample, $T(P)$ is the set of all Taxa occurring in the fossil pollen ~~samples~~and macrofossil synthesis, and $\theta := (\beta_i^T,\, i = 1, ..., 6, T \in T(P))$.

As described above, the absence of a taxa in a pollen or macrofossil sample can have reasons that are not included in the absence probability estimated from Eq. (4), as this calibration is only valid in the vegetation space. As information on the absence of a taxa in the vegetation space (i.e. absence of the taxa in the grid box of interest at the respective time slice) is not available from pollen and macrofossil data, the only reliable occurrence information of a taxa in the respective grid box in the past is the presence of the taxa in a pollen or macrofossil sample (Gebhardt et al., 2003). Hence, only occurring taxa are included in the reconstruction step.

Violations of the assumption that taxa are treated as conditionally independent given climate, i.e. due to co-occurrence of taxa (Kühl et al., 2002), can lead to over-fitting and subsequently underestimation of uncertainty in the transfer functions. Therefore, a statistical preselection of taxa, which are present in a sample, is applied. This procedure uses the Mahalanobis distance (Mahalanobis, 1936) between the fitted distributions and is described in detail in Kühl et al. (2002) and Gebhardt et al. (2008) . For the pollen and macrofossil synthesis used in this study, the preselection was carried out by Simonis (2009) and we follow their results.

It is assumed that the transfer function acts locally, i.e. given the climate at location x, a pollen sample at x is conditionally independent of the climate and proxy data at all other locations. Following these considerations, $\mathbb{P}(P_p\,|\,C_p,\theta)$ is given by

$$\mathbb{P}(P_p\,|\,C_p,\theta) := \prod_{s=1}^{S_p} \prod_{T\in T(s)} \mathbb{P}\left(P_p^T(s)\,|\,C_p(x_s),\beta_1^T,...,\beta_6^T\right), \tag{5}$$

where $T(s)$ are the taxa occurring in sample s and picked by the preselection procedure of Simonis (2009).

Finally, we define a prior distribution for θ. We use a Gaussian distribution centred at 0 and a marginal variance of 10 for each parameter β_i^T. Due to the absence of prior information on the correlation structure, we assume independence between the taxa as well as within a taxa. ~~Hence, we get~~

$$\mathbb{P}(\theta = (\beta_i^T, i=1,...,6, T\in T(P))) = \prod_{T\in T(P)} \prod_{i=1}^{6} \mathcal{N}\left(\beta_i^T\,|\,0,10\right).$$

Due to the high information content in the calibration data set, the influence of the prior ~~(Eq. ??)~~ on the parameter estimates is negligible. Using a flat prior for $C_p(x_s)$ and removing spatial correlations, local climate reconstructions at the locations of the proxy samples can be calculated. These reconstructions depend only on the proxy data in grid box x_s. Results of local MH reconstructions for each grid box with proxy data are shown in Fig. 4, where the local reconstruction means and the marginal 90% ~~CIs~~ credible intervals (CIs) are plotted for MTWA and MTCO.

Local reconstructions can also be used to evaluate the ability of the ~~transferfunctions~~ transfer functions to reconstruct modern climate which provides a reference for possible regional biases. For the PITM model such evaluations have been performed by Gebhardt et al. (2008) and Stolzenberger (2011). Both evaluations show that the model tends to underestimate north-south gradients leading to positive biases in Fennoscandia, and slightly negative biases in the Mediterranean. The biases as well as the uncertainties are larger for winter temperature than for summer. Therefore, results for MTCO in northern Fennoscandia should be treated with caution, while for all other regions biases of the reconstruction means are within reconstruction uncertainties.

3.3 ~~Inference and computational performance~~

~~Because the PITM model is~~ A disadvantage of the PITM version used in this study, is the inconsistent use of calibration and fossil data by using presence and absence information on taxa for the calibration but only occurring taxa in the reconstruction. Despite this inconsistency, the reconstructions in this study are in agreement with previous versions of PITM, where this inconsistency with the calibration did not appear as previously only occurrence information were used to fit the probability density functions. However, there is no simple solution for the problem that the calibration is in vegetation space whereas the absence of taxa in the fossil samples is an information in the pollen or macrofossil space. A promising idea might be to model the absence due to non-climatic reasons as zero-inflation by adding a latent variable to estimate the detection probability of a taxa (MacKenzie et al., 2002). But the estimation of detection probabilities is a very challenging task because it depends on many factors like pollen influx area, local topography, soil properties, and plant competition which might change over time. It is a priori unclear which of these factor can be marginalized and which have to be included as covariates. In addition, the

Simonis et al. (2012) synthesis combines macrofossils with pollen data. The processes that influence the detection probability of macrofossils are very different than for pollen. Therefore, a different detection probability has to be estimated for pollen than for macrofossils. Resolving the described issues is an interesting direction for future research, which requires extensive cooperation of (paleo)climatologists, (paleo)botanists, and statisticians, but is beyond the scope of this study.

3.3 Process stage

Similar to data assimilation approaches in numerical weather and climate prediction, the ensemble of climate simulations is used to control the spatial structures of the reconstruction and to constrain the range of physically possible climate states for a given external forcing by computing a spatial prior distribution from the ensemble members. This distribution is combined with interpolation parameters ϑ to facilitate a more flexible adjustment to the proxy data in the process stage of Eq. (2). The estimation of the prior distribution is hampered by the small number of ensemble members $K = 7$.

It is not obvious which method for estimating the prior distribution is best suited for the problem on hand and which additional model parameters are appropriate to preserve as much physical consistency contained in the climate simulations as possible but to correct for climate model inadequacies. Therefore, we perform a comparison study of six process stage models, that are composed of three techniques to formulate the process stage and two choices for the involved spatial covariance matrix. All those models are inspired by methods used in data assimilation, postprocessing of forecasts, and climate change detection and attribution. The use of climate simulations in the process stage allows not just interpolation between proxy samples but also structural extrapolation through the eigenvectors of the spatial covariance matrix and the process stage parameters.

3.3.1 Gaussian model

The most common approach in the data assimilation literature is to assume that the ensemble members are independent and identically distributed (iid) samples from an unknown Gaussian distribution of possible climate states (Carrassi et al., 2018) . In the following, the climatological means of the K ensemble members are denoted by μ_k. Subsequently the spatial prior distribution is given by $\mathcal{N}(\bar{\mu}, \Sigma_{\text{prior}})$, where $\mathcal{N}$ denotes a Gaussian distribution, $\bar{\mu}$ is the ensemble mean, and Σ_{prior} is a spatial covariance matrix, which is given by a regularized version of the empirical covariance

$$\Sigma_{\text{emp}} = \frac{1}{K-1} \sum_{k=1}^{K} (\mu_k - \bar{\mu}) (\mu_k - \bar{\mu})^t . \tag{6}$$

The superscript t denotes the matrix transpose. Hence, the covariance matrix is based on the inter-model differences as an estimate of epistemic uncertainties. The regularization techniques of Σ_{emp} are specified below. The Gaussian distribution is multivariate and the state vector has dimension N, where N is the number of grid boxes times the number of jointly reconstructed variables.

The main advantage of this Gaussian model (GM) is that inference becomes simpler than in more complex probability density estimation techniques because the prior distribution is unimodal and Gaussian. The disadvantage is that it relies on the strong assumption that the samples μ_k are iid samples from an unknown Gaussian distribution. This assumption tends to

be more realistic for samples from just one ESM, whereas statistics of multi-model ensembles are often not well described by purely Gaussian distributions (Knutti et al., 2010). A second disadvantage of this model is that the absence of additional parameters limits the possibilities to adjust the posterior distribution to the proxy data and correct for climate simulation inadequacies. The third disadvantage of the GM is that spatial structures of the individual models, which are directly derived from the physical equations solved in the ESMs, get lost by averaging over all ensemble members. Nevertheless, in many climate prediction applications multi-model averages outperformed each individual ESM (Krishnamurti et al., 1999, e.g.).

3.3.2 Regression model

A relaxation of the assumptions of the GM is the second model, that we call the regression model (RM) because it is inspired by regression based models popular in postprocessing and climate change detection and attribution (Hegerl and Zwiers, 2011) . In the RM, the iid assumption is dropped for the first moments of the process stage by introducing weighted averages of the ensemble members with variable weights $\lambda_k, k = 1, ..., K$. This means, that samples, which fit better to the proxy data, are weighted higher in the posterior. The sum of the weights is set to one such that unrealistically warm or cold state are prevented. This leads to the process stage model

$$\mathbb{P}(C_p|\lambda_1, ..., \lambda_K) = \mathcal{N}\left(C_p \,\middle|\, \sum_{k=1}^{K} \lambda_k \mu_k, \Sigma_{\text{prior}}\right), \tag{7}$$

and an additional prior distribution for the model weights

$$\mathbb{P}(\lambda) = \text{Dir}\left(\lambda_1, ..., \lambda_K \,\middle|\, \tfrac{1}{2}, ..., \tfrac{1}{2}\right). \tag{8}$$

Dir denotes a Dirichlet distribution, which is chosen as prior distribution because is guarantees that the weights take values between zero and one and sum up to one. Conditioned on the ensemble member weights, the process stage distribution is Gaussian, but non-Gaussianity is permitted through the variable weights.

In addition, the RM has the advantage of possessing more degrees of freedom compared to the GM. The inference process becomes a little more involving than for the GM because the ensemble member weights have to be estimated, too, but the conditional Gaussian distribution of C_p helps designing efficient inference algorithms. Similar to the GM, spatial structures of the individual models can get lost, as we average over different ESM climatologies.

3.3.3 Kernel model

The third model has been introduced in the data assimilation literature by Anderson and Anderson (1999) to combine particle and Gaussian filtering approaches, where the particle part helps capturing non-Gaussian and non-linear features but the efficiency of Gaussian approximations in high dimensional filtering situations is still exploited. This kernel model (KM) assumes that each ensemble member is a sample from an unknown distribution of possible climate states given a set of forcings, but it does not assume that this unknown distribution is Gaussian. Instead, non-parametric kernel density estimation techniques

(Silverman, 1986), where the probability distribution is given by a mixture of multivariate Gaussian kernels, are used. Each ensemble member climatology corresponds to the mean of a kernel.

Ideally, the covariance matrix of each kernel would correspond to the respective ESM, such that the spatial autocorrelation of that ESM is preserved when we sample from its kernel. Unfortunately, there is only one MH run available for each ESM and the internal variability in those runs is much smaller than the inter-model differences. Using the internal variability of those runs would thus lead to very distinct kernels and allow too few climate states. Therefore, the covariance of each kernel is estimated from the inter-model differences as a compromise that allows to sample from a much broader range of states even though autocorrelation of the individual models is lost. This compromise is a very common choice in kernel based probability density approximations (Liu et al., 2016; Silverman, 1986, Chapter 3 and 4) if there is no good estimate for the covariance corresponding to each kernel available.

Compared to the GM, the empirical covariance matrix Σ_{emp} is scaled by the Silverman factor (Silverman, 1986)

$$f := \left(\frac{4}{K\cdot(N+2)}\right)^{\frac{2}{N+4}}, \tag{9}$$

which optimizes the variances of the kernels, Hence, in the KM the scaled empirical covariance matrix $\tilde{\Sigma}_{\text{emp}}$, given by $f \cdot \tilde{\Sigma}_{\text{prior}}$, is regularized leading to the spatial covariance matrix $\tilde{\Sigma}_{\text{prior}}$. Note that the small number of ensemble members compared to the dimension of the probability distribution leads to a standard deviation reduction of only around 2% in our applications.

Each kernel gets an assigned weight ω_k, ~~the posterior climate follows again a mixture distribution~~$k = 1, ..., K$, which is inferred in the Bayesian framework. The weights sum up to one. The resulting process stage is a mixture distribution

$$\mathbb{P}(C_p \,|\, \omega_1, ..., \omega_K) \;=\; \sum_{k=1}^{K} \omega_k \, \mathcal{N}\left(C_p \,|\, \mu_k, \tilde{\Sigma}_{\text{prior}}\right). \tag{10}$$

A Dirichlet distributed prior is used for ω with parameter $\frac{1}{2}$ for each of the K components. A computational disadvantage of the KM is that the process stage is multi-modal and non-Gaussian. We augment the model by an additional parameter z, which follows a categorical distribution, denoted by Cat, to restore that C_p is Gaussian conditioned on the ω and z. z selects a kernel k according to its weight ω_k, i.e. z is defined such that

$$\mathbb{P}(\omega) = \text{Dir}\left(\omega_1, ..., \omega_K \,|\, \tfrac{1}{2}, ..., \tfrac{1}{2}\right) \tag{11}$$

$$\mathbb{P}(z \,|\, \omega) = \text{Cat}\left(z_1, ..., z_K \,|\, \omega_1, ..., \omega_K\right) \tag{12}$$

$$\mathbb{P}(C_p \,|\, z) = \prod_{k=1}^{K} \left(\mathcal{N}\left(C_p \,|\, \mu_k, \Sigma_{\text{prior}}\right)\right)^{z_k}. \tag{13}$$

Integrating out z yields the mixture distribution Eq. (10).

Two advantages of the KM are that it is not assumed that the unknown prior distribution is Gaussian and that the kernels do not rely on an iid assumption for their first moment properties. On the other hand, the KM relies on an iid assumption for the

second moment properties, which is the compromise described above due to the absence of a suitable estimate of the uncertainty structures corresponding to each ensemble member. The KM preserves the spatial structures of each ESM in the first moments of the kernels and when sampling from Eq. (13), the mean of the sample belongs to one ESM and is not a weighted average over all ensemble members. This preservation of physical consistency reduces the degrees of freedom compared to the RM. For example, when the true climate state lies exactly between μ_1 and μ_2 and far away from all other ensemble members, the weights of μ_1 and μ_2 can be increased compared to the other members in the KM, but the ~~components do not belong to a standard probability distribution, that could be used for sampling~~mode cannot be changed to $\frac{1}{2}(\mu_1 + \mu_2)$, which is possible in the RM. Another disadvantage of the KM is that the multi-modality makes the design of efficient inference algorithms a lot more challenging.

3.3.4 Glasso based covariance matrices

The first technique to regularize the empirical covariance matrix (the scaled empirical covariance in the KM), which is applied in this study, is the graphical lasso algorithm (glasso, Friedman et al., 2008, implemented in the R-package glasso). This algorithm approximates the precision matrix (inverse covariance) by a positive definite, symmetric, and sparse matrix $\Sigma_{\text{prior}}^{-1}$. Therefore, Σ_{prior} is a valid N-dimensional covariance matrix. Glasso maximizes the penalized log-likelihood

$$\log\det\Sigma_{\text{prior}}^{-1} - \text{trace}(\Sigma_{\text{emp}}\,\Sigma_{\text{prior}}^{-1}) - \rho\,\|\Sigma_{\text{prior}}^{-1}\|_1, \tag{14}$$

where ρ is the penalty parameter, $\|\cdot\|_1$ is the vector L_1-norm, and the first two terms are the Gaussian log-likelihood. Because applying the glasso algorithm is computationally expensive, it is not feasible to formally include ρ in the Bayesian framework. Instead a suitable value of ρ has to be determined prior to the inference. In this study, ρ is chosen such that Σ_{prior} is a numerically stable covariance matrix and the performance in cross-validation experiments (CVEs) is optimized. Technical details of the determination of ρ are described in Appendix A.

The advantage of the glasso approach is that the empirical matrix can be approximated very closely and the sparseness of the precision matrix facilitates the use of efficient Gaussian Markov random field (GMRF) techniques (Rue and Held, 2005) in the inference algorithm. A disadvantage is that no new spatial structures are added to Σ_{emp}. Therefore, the effective number of spatial modes is much smaller than the dimension of the climate vector, which can lead to a collapse onto a very small subspace of the N-dimensional state space and subsequently biases and under-dispersion.

3.3.5 Shrinkage based covariance matrices

To overcome the deficiencies of the glasso approach, we propose an alternative covariance regularization technique, which combines the empirical correlation matrix of the climate simulation ensemble with a regular correlation matrix. A so-called shrinkage approach (Hannart and Naveau, 2014) is used, which is a weighted average of the empirical correlation matrix and a reference matrix, which in our case contains additional spatial modes such that the effective number of spatial modes in the covariance matrix is increased. This allows deviations from the spatial structures prescribed by the climate simulation ensemble and is therefore a strategy to account for climate model inadequacies.

Let Ψ_{emp} be the empirical correlation matrix of the climate simulation ensemble, which is related to Σ_{emp} by

$$\Sigma_{\text{emp}} = \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}} \, \Psi_{\text{emp}} \, \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}}, \tag{15}$$

where $\text{Diag}(\Sigma_{\text{emp}})$ denotes a diagonal matrix with the the same diagonal entries as Σ_{emp}, and the exponent $\frac{1}{2}$ means that the square root of each diagonal entry is taken. Replacing Ψ_{emp} by a weighted average of Ψ_{emp} and a shrinkage target Φ leads to the shrinkage covariance matrix

$$\Sigma_{\text{prior}} = \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}} \, (\alpha \, \Psi_{\text{emp}} + (1 - \alpha) \, \Phi) \, \text{Diag}(\Sigma_{\text{emp}})^{\frac{1}{2}}. \tag{16}$$

α is the weighting parameter, which takes values between zero and one. Φ is computed from a numerically efficient GMRF approximation of a stationary Matérn correlation matrix (Lindgren et al., 2011). Matérn correlation functions correspond to diffusive transport of spatial white noise forcing. The Matérn correlation matrix is controlled by three parameters, the smoothness, the range ρ, and the anisotropy ν. We fix the smoothness at a value which corresponds to the application of a standard Laplace operator. ρ controls the decorrelation length, and ν parameterizes the length of the meridional compared to the zonal decorrelation length. For joint reconstructions of multiple climate variables, independent correlation matrices for each variable are combined in a block structure. Details about the definition of Φ are given in Appendix B.

Ideally, the parameters α, ρ, and ν are estimated from the proxy data. But initial tests with weakly informative priors showed that the parameters cannot be constrained by the available proxy data, because the signal in the proxy data is not informative enough to infer second moment properties of the process stage. Therefore, an ensemble of parameter combinations is created from fitting the shrinkage model to each of the climate simulation ensemble members given all other members. This results in seven consistent sets of α, ρ, and ν. Those are passed to the reconstruction framework, such that each combination is chosen with a probability inferred from the proxy data. Thereby, each parameter combination is based on a fit against physically consistent structures, and uncertainty in the parameters is included in the inference, but the problem of non-identifiability of the parameters from proxy data alone is reduced. The parameter estimates depend strongly on the chosen ESM, which leads to combinations that cover a wide range of possible values.

The main advantage of the shrinkage approach over the glasso based matrices is that more spatial modes are included in the covariance matrix. Thereby, the collapsing of the reconstruction towards a very low-dimensional subspace is mitigated and stronger deviations of the reconstruction from the climate simulation ensemble are facilitated. A disadvantage compared to the glasso estimated covariance matrices is that neither the shrinkage covariance matrix Σ_{prior} nor its inverse are sparse matrices such that the numerical inference algorithms are more costly.

3.4 Inference strategy

Because PITM is non-Gaussian and non-linear, the posterior climate does not belong to a standard probability distribution. Therefore, ~~we use~~ Markov chain Monte Carlo (MCMC) techniques are used to asymptotically sample from the correct posterior distribution. These samples allow analyses beyond summary statistics like means and standard deviations. ~~We choose a~~ A Metropolis-within-Gibbs strategy is implemented, which means that ~~we sample alternately~~ in each update of the Markov

chain, we sample sequentially from the full conditional distributions (i.e. the distribution of the respective variable given all other variables) of θ, ~~ω, z~~ϑ, and C_p. This strategy is chosen because for many variables the full conditional distributions follow probability distributions for which efficient sampling algorithms exist, and for the remaining variables Metropolis-Hastings updates are used for sampling.

To sample the regression parameters θ in Eq. (??3) to (??5) efficiently, the data augmentation scheme of Polson et al. (2013) is used. For taxa T, the full conditional is only depending on C_p, C_m, P_p^T, and P_m^T, but not on other taxa. Therefore, we can sample $\beta_1^T, ..., \beta_6^T$ independently from the other taxa. Polson et al. (2013) introduce help variables $\gamma_l^T, l = 1, ..., L$, where L is the number of observations ~~(absence and presence)~~ of taxa T, such that $\mathbb{P}(\gamma_l^T \,|\, \beta_1^T, ..., \beta_6^T, C_m, C_p)$ is Pólya-Gamma (PG) distributed, and $\mathbb{P}(\beta_1^T, ..., \beta_6^T \,|\, P_m^T, P_p^T, \gamma_1^T, ..., \gamma_L^T)$ is Gaussian. Therefore, ~~we sample~~ the MCMC algorithm samples alternately from a PG distribution using the sampler of Windle et al. (2014) and from a ~~multivariate Gaussian~~ Gaussian distribution. The PG sampler is implemented in the R package BayesLogit (Windle et al., 2013).

~~An easy way to sample from the climate mixture distribution is to introduce a multinomially distributed augmentation variable $z = (z_1, ..., z_K)$ such that~~

$$\begin{aligned} \mathbb{P}(\omega) &= \text{Dirichlet}\left(\omega_1, ..., \omega_K \,|\, \tfrac{1}{2}, ..., \tfrac{1}{2}\right) \\ \mathbb{P}(z \,|\, \omega) &= \text{Multinomial}\left(z_1, ..., z_K \,|\, n = 1, \omega_1, ..., \omega_K\right) \\ \mathbb{P}(C_p \,|\, z) &= \textstyle\prod_{k=1}^{K} \left(\mathcal{N}\left(C_p \,|\, \mu_k, \Sigma_{\text{prior}}\right)\right)^{z_k}, \end{aligned}$$

~~i.e. C_p, conditioned on z selecting model k, is Gaussian. Integrating out z yields the mixture distribution Eq. (??). Equations (17) to (17) lead to full conditionals for ω and z, which are again Dirichlet and multinomially distributed but with updated parameters. Therefore, we can use Gibbs sampling to update ω and z. Preliminary tests revealed that in our application the prior distribution of ω has a negligible influence on the posterior distributions of z and C_p.~~To sample from the full conditional of C_p, we separate the grid boxes x_P with at least one proxy record from those without any proxy records denoted by x_Q. There is no closed form available for the full conditionals of $C_p(x_P)$. Therefore, we use a random walk Metropolis-Hastings algorithm to update $C_p(x_P)$ sequentially for all members of x_P. As ~~these updates are two-dimensional and the target distributions are in most cases close to Gaussian, the random walk proposals are very efficient. As the transferfunctions~~ the transfer functions act locally, $C_p(x_Q)$ is conditionally independent of P_p given $C_p(x_P)$ and ~~z~~ϑ. Therefore, we subsequently update $C_p(x_Q)$ by sampling from ~~$\mathbb{P}(C_p(x_Q) | C_p(x_P), z)$~~ $\mathbb{P}(C_p(x_Q) \,|\, C_p(x_P), \vartheta)$ which is Gaussian. ~~Detailed formulas for the full conditional distributions are given in Appendix C.~~

~~The mixture distribution Eq. (??) leads to a multimodal posterior. Due to the high dimensionality of C_p, the likelihood of choosing a new model z_k from one MCMC stepto the next one is very small. This leads to very slow mixing of the MCMC algorithm described above. To speed up the mixing, we apply a Metropolis coupled MCMC strategy (MC^3) , also called parallel tempering, which was developed by Geyer (1991) and adapted to parallel computer architectures by Altekar et al. (2004) and Werner and Tingley (2015). We run A MCMC chains in parallel, and after every M steps, we use an additional Metropolis-Hastings step to swap the states of the Markov chains a_1 and a_2 with probability $0 < p_{a_1,a_2} < 1$, where p_{a_1,a_2} is calculated from the~~ Sampling from ϑ depends on the particular process stage model. In models with shrinkage covariance matrix, α, ρ, and ν are

sampled from the K parameter combinations in a Metropolis-Hastings step. The weights λ in the RM are sampled from a random walk type Metropolis-Hastings ~~odds ratio. The Markov chains are created by exponentiating the process stage and the data stage by constants $\nu_1 = 1 > ... > \nu_A > 0$. The first Markov chain ($\nu_1 = 1$) asymptotically retains the original posterior distribution for all variables, whereas the subsequent chains sample from a flatter posterior distribution, in which it is easier to jump from one mixture component to another. Following empirical testing, we run the European reconstructions with $A = 8$ parallel chains, levels $\nu_1 = 1, \nu_2 = 1.25^{-1}, ..., \nu_8 = 1.25^{-7}$, and swaps after every $M = 30$ steps~~update. In the KM, Eq. (11) to (13) lead to full conditionals for ω and z, which are again Dirichlet and categorically distributed but with updated parameters. Therefore, Gibbs sampling can be used to update ω and z.

The multi-modality of the KM makes inference for this model a lot more challenging than for the GM and RM. The problem of efficient MCMC algorithms for multi-modal posterior distributions is a widely acknowledged issue in the literature (Tawn and Roberts, 2018) and in this study Metropolis coupled Markov chain Monte Carlo (MC^3; Geyer, 1991), which is also known as parallel tempering, is used to overcome this issue. Details of this procedure are provided in Appendix D.

~~As a second strategy to~~ To speed up the inference, ~~we treat the~~ grid boxes with proxy data and those without proxy data are treated sequentially. Because of the ~~Gaussian mixture components in~~ conditional Gaussian structure of the process stage and because the grid boxes without proxy data are not influencing the posterior ~~model weights, we can first integrate out of~~ ϑ, $C_p(x_Q)$ ~~and~~ is integrated out to get an estimate of the joint distribution of ~~ω, z,~~ Θ and $C_p(x_P)$. In a second step, we sample from $C_p(x_Q)$ conditioned on $C_p(x_P)$ and ~~z~~Θ, which leads to joint samples of C_p from the asymptotically correct posterior distribution. ~~In practice, this is done by drawing a sample of $C_p(x_Q)$ conditioned on each of the MCMC samples from $(C_p(x_P), z, \omega, \theta)$.~~ This strategy reduces the ~~number of MCMC chains needed in the MC^3 algorithm, and it reduces the~~ computation time of each MCMC update due to faster matrix operations. ~~Pseudo-code for the MC^3 algorithm is given in Appendix ??.~~

The remaining bottleneck in computation time is the estimation of the ~~transferfunction~~ transfer function parameters due to the large modern calibration set. While in theory the observation layer influences the updates of θ, in practice the influence of Eq. (~~??~~5) on the posterior of θ is negligible. Therefore, ~~we use~~ a modularization approach (Liu et al., 2009) is used similar to Parnell et al. (2015) in ~~cross-validation experiments~~CVEs, where a sequence of reconstructions with slightly changed proxy networks is computed~~(see Sect. ??)~~. This means that ~~we cut~~ feedback between Eq. (~~??~~4) and Eq. (~~??~~) 5) is cut by first drawing as many MCMC samples as necessary from θ using only Eq. (~~??) and Eq. (??~~4). Thereafter, ~~we reconstruct~~ C_p is reconstructed using these samples instead of sampling θ from its full conditional.

Detailed formulas for the full conditional distributions are given in Appendix C. Pseudo-code for the MCMC and MC^3 algorithms is provided in the Supplement. For a 798 dimensional climate posterior as it is the case in joint reconstructions of MTWA and MTCO, and 45 grid boxes that contain at least one proxy record, ~~we create~~ 75,000 MCMC samples ~~for each of eight parallel chains, where every chain runs on one central processing unit (CPU)~~are created. The first 25,000 samples are discarded as ~~burnin~~burn-in. To reduce the autocorrelation of subsequent samples, ~~we extract~~ every fifth sample is extracted to create a set of 10,000 posterior samples which is used for further analyses. On a standard desktop computer~~with at least eight CPUs~~, reconstructions with the modularized model can be computed in approximately 30 minutes. The convergence of

all MCMC variables is checked using the Gelman-Rubin-Brooks criterion (Brooks and Gelman, 1998) implemented in the R package coda (Plummer et al., 2006).

3.5 ~~Assessing the added value of reconstructions~~

~~To analyse the added value of combining proxy data and PMIP3 simulations compared to only using~~

4 Results

In this section, results from a comparison study of the six different process stage models are presented. Then, the MH reconstruction for Europe with the Simonis et al. (2012) synthesis and the PMIP3 ~~simulations and to assess the consistency~~ MH ensemble is presented. We study the mean and uncertainty structure of the posterior distribution, the fit of different ensemble members to the proxy data, the added value of the reconstruction, ~~we perform leave-one-out cross validation. Due to the sparseness of the proxy network, we do not leave out larger amounts of data. In addition, the cross-validations can be used to identify systematic mismatches between proxy data and simulations as well as potential outliers in the data.~~

~~The cross-validation is performed in the observation space and consists of three steps: A reconstruction with all proxy samples except for those in grid box x is computed. The forward model Eq. (??)is applied to the posterior and the prior distribution of $C_p(x)$ to get two probabilistic predictions for each proxy sample $P_p(s)$ with $x_s = x$. Proper scores (Gneiting and Raftery, 200 are computed for both probabilistic predictions . Then, the corresponding skill score, which is a measure of the added value from constraining the~~ and the results of joint compared to separate MTWA and MTCO reconstructions.

4.1 Comparison of different process stage frameworks

In this section, the reconstruction skill of the three process stage formulations (GM, RM, KM) and the two covariance models (glasso, shrinkage) are compared using two types of experiments. Identical twin experiments (ITEs) use the climate simulation ensemble by simulating pseudo-proxy data from one ESM and trying to reconstruct that reference climatology from the simulated proxies and the remaining ensemble members. These experiments facilitate the understanding of different modelling approaches for the process stage in a controlled environment. In particular, the ability of the Bayesian frameworks to reconstruct the climate can be evaluated without having to rely on indirect observations as it is the case in real paleoclimate applications where the true climate state is unknown. The second type of experiments are CVEs, where spatial reconstructions with the Simonis et al. (2012) synthesis are performed but the samples from one grid box are left out. Then, the reconstructions for this grid box are evaluated against the left-out sample in the vegetation space by applying the PITM forward model to the reconstruction. The advantage of these experiments is that the models are compared in a real-world setting. The disadvantage of CVEs is that the goal of a reconstruction is to reconstruct climate but there are no direct observations of paleoclimate available such that evaluations against observations have to be indirect. Evaluating prediction models against indirect observations in the observation space through forward modelling is also a recent way of model skill evaluation in weather forecasting.

4.1.1 Identical twin experiments

ITEs make use of the fact that the PMIP3 ~~ensemble by pollen data, is calculated. The PITM forward model maps a climate field to a binary field (presence or absence of a taxa), where the probabilistic prediction~~ ensemble is a multi-model ensemble such that the models can produce fairly different climatologies. Therefore, trying to reconstruct the climate state of one ESM given the others is a more realistic test environment than doing the same with single-model ensembles. The first step in an ITE is to choose a reference climate model with climate state C_p^{true}. Then, for each grid box, that contains samples from the Simonis et al. (2012) synthesis (denoted by x_P in accordance with the notation in Sect. 3.4), pseudo-proxies are simulated. The proxies are simulated according to a Gaussian approximation of the uncertainty structure of the local reconstructions depicted in Fig. 4, which means that it is described by a bivariate covariance matrix $\Sigma_p^{x_s}$ for each grid box x_s with proxy data. The pseudo-proxies are assumed to be unbiased with a bivariate Gaussian distribution, i.e.

$$P(x_s) \sim \mathcal{N}\left(C_p^{\text{true}}(x_s), \Sigma_p^{x_s}\right), \quad x_s \in x_P. \tag{17}$$

Using unbiased Gaussian pseudo-proxies is a common strategy to test climate field reconstruction techniques (e.g. Gomez-Navarro et al., 20 . It allows a direct study of the ability of the process stage methods to estimate spatial climate fields from sparse and noisy proxy data, without having to factor in potential biases in the transfer function. With the simulated proxies, the Bayesian framework is applied to compute a probabilistic spatial reconstruction, but the reference model climatology is removed from the climate model ensemble. ITEs are performed with each of the seven PMIP3 ensemble members as reference climate state, and the six different process stage configurations. For each of those 42 combinations, five randomized ITEs are computed to separate random from systematic issues.

The evaluation of the ITEs focuses on biases in the reconstructions, potential under-dispersion which is a typical phenomenon in data assimilation applications, and, as a combination of those two issues, the ability of the reconstruction to probabilistically predict past climate.

Averaged over all ITEs with the same process stage model and averaged in space, the mean deviation between reference climate and posterior mean as a measure for systematic biases is close to 0 K for all process stage models with values between -0.14 K for the shrinkage KM and +0.03 K for the shrinkage RM, but the variation across ITEs is larger for the glasso covariance models (standard deviation around 0.31 K) than for the shrinkage covariance models (standard deviation around 0.24 K). This shows that the additional spatial modes in the shrinkage covariances make the reconstructions more robust (Table 2, Fig. 5a). The standard deviations for MTCO reconstructions tend to be larger than for MTWA which can be explained by the larger noise level in the local MTCO reconstructions. This makes the spatial MTCO reconstructions more susceptible to biases. Concerning the spatial patterns of mean deviations, all ITEs with glasso covariance matrices have larger local biases than those with shrinkage matrices (see additional figures in Supplement). While the magnitude of biases for the GM and RM models with shrinkage covariances is much smaller, the magnitude of local deviations of the shrinkage KM model is just slightly smaller than for glasso covariances. This shows that the models with shrinkage matrix can reconstruct spatial patterns better than the models that use glasso due to more degrees of freedom in the covariance matrix. In addition, averaging over

different climate models in the mean of the process stage seems to be a more effective strategy as the GM and RM reproduce the spatial structures better than the KM.

The higher number of spatial modes in the shrinkage covariances leads to larger uncertainty estimates in the posterior distribution than for the glasso models, because the limited information contained in the proxy data can constrain only a small number of spatial modes (Table 2). To study dispersiveness of the reconstruction coverage frequencies for 50% and 90% CIs are calculated. This means that the frequency of the reference climate state to be included in the respective CIs is computed. For the 50% CIs, coverage frequencies below 50% indicate under-dispersiveness, whereas values above 50% indicate over-dispersion. Similarly, the target for the 90% CIs is 90%. In all ITEs, the glasso models are under-dispersive, and the shrinkage models are over-dispersive (Table 2, Fig. 5c). The coverage frequency for 50% CIs is below 41% in all ITEs with glasso covariance matrix (mean close to 30% in all three experiments) and above 56% in all ITEs with shrinkage covariance matrix (mean around 78%). Similarly, the coverage frequencies for 90% CIs are below 77% for all glasso ITEs and above 94% for all ITEs with shrinkage matrix (Fig. 5d).

At most grid boxes of the ITEs with glasso based covariance matrix, the coverage frequencies are below the target values, whereas they are above the desired values at almost all grid boxes in the ITEs with shrinkage matrix (see additional figures in Supplement). The values are closest to the target near grid boxes with proxy data in the glasso as well as the shrinkage matrix ITEs. This effect is more pronounced for the 50% coverage frequencies than the 90% coverage frequencies, which indicates that the reconstruction identifies the centers of the probability distribution better than its tails, which is not surprising considering the simplicity of all the process stage models and the small ensemble size. In particular, the process stage models do not contain parameters which control the tail behaviour.

To analyse the combined effect of biases and dispersiveness, the continuous ranked probability score (CRPS) is computed. This is a common strictly proper score function for evaluating probabilistic predictions (Gneiting and Raftery, 2007), in our case the ability of a reconstruction method to probabilistically predict past climate from sparse and noisy data. It is a generalization of the absolute error to probabilistic forecasts (Matheson and Winkler, 1976), given by

$$\mathrm{CRPS}\left(F, C_p^{\mathrm{true}}(x)\right) := \int_{-\infty}^{\infty} \left(F(y) - \delta_{(y \geq C_p^{\mathrm{true}}(x))}\right)^2 dy, \tag{18}$$

where F is the cumulative distribution function of the probabilistic reconstruction C_p at grid box x, and $C_p^{\mathrm{true}}(x)$ is the reference climate state at x. Defined in that way, the CRPS has a unique minimum at 0 and is positive unless F is a perfect prediction.

With a spatially averaged mean around 1 K for all three models, the ITEs with glasso covariance matrices have a higher CRPS than the ITEs with shrinkage matrices that have a spatially averaged mean around 0.4 K (Table 2, Fig. 5b). This is a result of larger biases on the grid box level and under-dispersiveness of the posterior distribution. In addition, the variability between the ITEs with the same process stage model is higher for models with glasso covariance, which shows that these models are less robust. The MTWA CRPS is slightly lower than the MTCO CRPS since the local reconstructions constrain MTWA more than MTCO. Among the process stage models with shrinkage matrix, the KM performs slightly worse than the GM and the RM which is a result of the larger biases on the grid box level described above. The spatial structures of CRPS reflect the

mean deviation patterns (Fig. 6). This is an effect of more pronounced spatial patterns in the mean deviations compared to dispersiveness.

4.1.2 Cross-validation experiments

CVEs are a way to understand the ability of a spatial reconstruction method to produce consistent estimates. In paleoclimatology, the issue is that all observations are indirect, which means that negative evaluations can result from errors in the process stage or the data stage. For example, even if the climate reconstruction at grid box x is accurate, the evaluation against the left-out data can be negative if the transfer function that translates data $C_p(x)$ into $P_p(x)$ is biased or its uncertainty estimate is misspecified. Hence, the assumption behind CVEs is that the data stage is unbiased or at least consistently biased among different proxy samples. Cross-validations are evaluated in the observation space. In this study, this is the vegetation space, i.e. the occurrence of taxa in a grid box. As the only reliable information that are available from the pollen and macrofossil synthesis on the vegetation composition in a grid box is the presence of certain taxa, this is also the only data that is used for the evaluation. Due to the sparseness of the proxy network, leave-one-out CVEs are performed and not more data is left out in each experiment.

In each CVE, a reconstruction with the Bayesian framework is computed with all proxy samples except for those in one grid box x. Then, the reconstruction $C_p(x)$ at grid box x is extracted and treated as probabilistic prediction of the climate at x. Next, the PITM forward model is applied to $C(x)$ for each sample P_p located in grid box x to produce probabilistic predictions of the occurrence of the taxa that are found in those samples. This prediction of the occurrence of a taxa T is represented by the probability of presence $p \in [0,1]$. A common ~~proper~~ score function for binary variables is the Brier score (BS; Brier, 1950) given by

$$BS(T) := \tfrac{1}{2}\left((\delta_T(1)-p)^2 + (\delta_T(0)-(1-p))^2\right) = \begin{cases} 1+p^2-2p & \text{if } T=1 \\ p^2 & \text{if } T=0 \end{cases} \tag{19}$$

where δ_T denotes the indicator function of taxa T. The BS takes values between zero and ~~two~~one, where zero corresponds to a perfect prediction and ~~two~~ one to the worst possible prediction. ~~In a recent study,~~ Stolzenberger (2017) used the BS to assess the skill of PMIP3 simulations for the MH using a network of pollen ~~records. To calculate the BS for a set of climate samples, either MCMC samples from the posterior or independent samples from the prior mixture distribution, the~~ and macrofossil records. The PITM forward model is applied to each ~~sample~~ MCMC sample, which leads to a set of probabilistic predictions $p_j(T), j = 1,...,J$ for taxa T. Predictions are calculated for each taxa which occurs in sample $P(s)$. The joint score of $P(s)$ is then calculated by averaging the BS of each taxa and prediction:

$$BS(P(s)) := \frac{1}{|T(s)|\,J} \sum_{T \in T(s)} \sum_{j=1}^{J} \left(21 + 2p_j(T)^2 - 42p_j(T)\right). \tag{20}$$

If multiple samples are assigned to one grid box, the mean score of those samples is taken. ~~Finally, the Brier skill score (BSS)is computed by comparing BS for posterior and prior:~~

$$\text{BSS} := \frac{\text{BS(Prior)} - \text{BS(Posterior)}}{\text{BS(Prior)}}.$$

5 ~~Results~~

~~In this section, results from the MH reconstruction for Europe with the Simonis et al. (2012) synthesis and the PMIP3 MH ensemble are presented. We study the mean~~ A problematic step in the methodology described above is that the BS is only evaluated for occurring taxa and not for those which are absent in the sample. This can make the BS improper when comparing statistical models that predict the presence or absence of taxa. However, the goal of the methodology described above is an indirect evaluation of predictions of past climate via transfer functions. In that context, it would lead to inconsistencies between the local reconstructions and the BS evaluations if taxa that are absent in the proxy synthesis were included as there is currently no model available to accurately estimate detection probabilities of pollen and macrofossil data (see the discussion in Sect. 3.2). Circumventing this issue is beyond the scope of this study. It should be noted that for each taxa the BS are a convex function of climate and minimal for a unique climate state. These two properties make the methodology described above useful for the comparison of climate reconstructions and the indirect evaluation of climate field reconstruction methods.

The models with glasso covariances perform slightly worse than those with shrinkage covariances, as the mean BS takes values of 0.186 (GM, RM) or 0.187 (KM) for the glasso based models compared to values between 0.161 and ~~uncertainty structure of the posterior distribution, the added value of the reconstruction, the performance of different ensemble members, the sensitivity with respect to the glasso penalty parameter,~~ 0.165 for models with shrinkage covariances (Table 2). Similar to the ITEs, the differences between models with different covariance types are larger than those with the same covariance model. Accordingly, the largest differences at individual gridboxes between models with different covariance matrix types are on the order of 10^{-1}, but only on the order of 10^{-2} between models with the same covariance type. The models with shrinkage covariance matrices tend to perform better in western Europe and Fennoscandia, whereas in central and eastern Europe, the magnitude of the differences is very small and the models with glasso covariance matrices perform slightly better in the majority of grid boxes.

The generally better performance of the models with shrinkage covariance matrices might be a result of the under-dispersive and less-robust behaviour of the models with glasso covariance matrices that was diagnosed in Sect. 4.1.1. This explanation is underlined by the spatial patterns of the performance differences. In western Europe and Fennoscandia, the local reconstructions tend to be less informative than in central and eastern Europe. Therefore, the reconstructions at the respective grid boxes are less constrained by nearby local reconstruction and more by far away proxy samples. This effect is enhanced in the glasso based models with less spatial degrees of freedom. Therefore, reconstructions with glasso matrices can be more biased if information are inadequately transferred over large differences. In central and eastern Europe this effect is less problematic, as the nearby proxy samples are the main contributors to the reconstructions at the left-out grid boxes. That the glasso based models perform even slightly better in central and eastern Europe could than be an effect of the less dispersive behaviour compared to the models with shrinkage covariances.

4.0.1 Conclusions from the comparison study

The ITEs show that the models with shrinkage matrix covariances are more dispersive, less biased, and more robust than those with glasso covariance matrices. These properties transfer to the CVEs where the models with shrinkage covariance matrix perform better, too. The results from models with the same covariance matrix are very similar except that the KM with shrinkage covariance matrix is on average more biased than the respective GM and RM. This shows that the covariance matrix choice determines the reconstruction skill more than the general formulation of the process stage as Gaussian, regression, or kernel model.

The better performance of shrinkage covariance models shows that the low number of spatial modes as a result of the small ensemble size is the main reason for the ~~results of joint compared to separate MTWA and MTCO reconstructions. Results from all reconstructions~~under-dispersiveness of the glasso based models. On the other hand, the over-dispersiveness of the shrinkage models should be an indicator that this model is not under-dispersed even in real world applications which face additional challenges from potential biases or under-dispersed transfer functions and from a potentially more sophisticated spatial structure of the climate state than in the ESM climatologies. Additionally, this over-dispersiveness shows that in most regions the ensemble spread is wide enough to lead to reconstructions which do not feature too narrow posterior distributions as long as enough spatial degrees of freedom are incorporated in the second moment properties of the process stage.

The larger biases of the KM with shrinkage covariance matrix compared to the GM and RM are a result of ensemble member weight degeneracy in the particle filter part of this model. The ensemble member weights tend to degenerate towards the least wrong model such that the mean values are biased towards that model. This tendency increases with the strength of the proxy data signal. This is a well-known issue of Bayesian model selection (Yang and Zhu, 2018), and therefore as well of particle filter methods (Carrassi et al., 2018), which hinders the use of KMs in data assimilation problems. To mitigate this issue, the particle filter part of the KM is combined with a Gaussian part that is more similar to Kalman type filters. The ITEs show that this adjustment is strong enough to avoid under-dispersiveness, but the degeneracy of the ensemble member weights still leads to larger biases than in the process stage models that rely on direct averaging of the ensemble members.

4.1 Spatial reconstruction of European MH climate

Based on the results presented in the previous section, the models with shrinkage matrix should be preferred over those with glasso covariance models. In addition, the smaller biases and more robust nature of the GM and RM with shrinkage covariance matrix compared to the KM model, makes them superior choices. Because the RM adjusts more flexible to the proxy data than the GM, this model is presumably better suited to deal with additional caveats of real world applications compared to the controlled test environment of ITEs. Therefore, this model is used for the spatial reconstructions, whose results are presented in this section. Reconstruction results are summarized in Table 1. Results from reconstructions with the other five process stage models are presented in the Supplement.

4.2 ~~Posterior mean and uncertainty structure~~

4.1.1 Posterior mean and uncertainty structure

The spatially averaged mean temperature of the reconstruction (posterior mean) is ~~17.77~~18.27°C (90% CI: (~~17.44~~17.79°C, ~~18.09~~18.75°C)) for MTWA and ~~2.24~~1.81°C (90% CI: (~~1.87~~1.22°C, ~~2.62~~2.45°C)) for MTCO, which is in ~~the former case equal to~~ both cases warmer than the the CRU reference climatology (~~+0.01~~ 0.51 K ~~) and in the latter 1.11~~ for MTWA and +0.69 K ~~warmer. Larger differences~~ for MTCO). Larger anomalies are found for subregions (Fig. 7a,~~7b):~~ b). For MTWA as well as MTCO, ~~cooler temperatures than today are found in most areassouth of 54° N, while north of this line the temperatures are~~ temperatures were cooler than today in many southern European areas, while in northern Europe the temperatures were predominantly higher than today~~with the exception of a cooling in parts of Fennoscandia. The on average higher MTCO anomalies compared to MTWA stem from higher anomalies in Fennoscandia. Many of the warming anomalies in the northern part as well as the cooling anomalies in the southern part~~ . More specifically, MTWA was warmer over Fennoscandia, the British Islands, and the Norwegian Sea. Most of these anomalies are significant on a 5% level. Here, a positive anomaly is called significant if the probability to exceed the reference climatology is at least 0.95. Significant negative anomalies are defined accordingly. The significance estimates are calculated point-wise. Negative MTWA anomalies are found in large parts of the Mediterranean and eastern Europe, but fewer anomalies are significant on a 5% level than in north-western Europe. The largest positive MTCO anomalies are found in Fennoscandia and off the Norwegian coast. In the other parts of the domain, the majority of MTCO anomalies are negative, but the spatial pattern is more heterogeneous than for MTWA. A lot fewer MTCO anomalies are significant on a 5% level ~~(see Sect. ?? for the definition of significance in the Bayesian context). In particular, using~~ compared to MTWA anomalies. Using the joint information ~~in the pollen~~ contained in the proxy synthesis and combining it with the spatial structure of the PMIP3 ensemble leads to more significant signals than in ~~any of~~ the individual data products.

Most of the taxa, which are used in the reconstruction, are stronger confined for MTWA than for MTCO because the growth of most European plants is more sensitive to conditions during the growing season. This results in more constrained local MTWA reconstructions (Fig. 4~~a~~c), which is in concordance with findings from Gebhardt et al. (2008). Hence, the uncertainty in the MTWA reconstruction is smaller ~~(spatially averaged point-wise 90% CI size of 2.90)~~ than in the MTCO reconstruction ~~(~~with spatially averaged point-wise 90% CI ~~size of 3.18~~ sizes of 4.15 K ~~)~~ and 5.84 K, respectively (Fig. 7c,~~7~~d). The uncertainty is smallest ~~for~~ at grid boxes with proxy records, and highest in the north eastern and north western parts of the domain where the PMIP3 ensemble spread is large and the constraint from proxy data is weak. For MTWA, additional regions with large uncertainties are found at the eastern and southern boundaries of the domain due to weak proxy data constraints. Besides, the reconstruction uncertainty has small spatial variations. The ratio of the CI size of spatially averaged temperatures and the spatially averaged point-wise CIs can be interpreted as a measure for the spatial degrees of freedom in the reconstruction.

The highest reduction of uncertainty due to the inclusion of proxy data is found at grid boxes with proxy data, as quantified by a spatially averaged reduction of point-wise CI sizes from prior to posterior of ~~69.5~~50.1% compared to ~~60.5~~26.0% for grid boxes without proxy data (Fig. 7e,~~7~~f). The uncertainty reduction for MTWA is higher for terrestrial grid boxes than marine ones, but the smaller PMIP3 ensemble spread over ~~sea~~ the British Islands, the North Sea, and the Bay of Biscay leads to

similar ~~CI sizes~~posterior CI sizes in these areas. For MTCO, the reduction of uncertainty is ~~highest for terrestrial areas and the Norwegian Sea and lowest for the Mediterranean Sea and off the French and Iberian coast. But since the PMIP3 ensemble spread is small in these areas, the reconstruction uncertainty is still lower than in some land parts of the domain~~generally smaller than for MTWA due to the weaker proxy data constraint.

To study whether the degree of spatial smoothing of the reconstruction is reasonable, ~~we calculate~~ a measure inspired by discrete gradients is calculated. For each grid box, ~~we calculate~~ the mean absolute difference between the value in the box and its eight nearest neighbours is computed. Then, ~~we compare~~ the spatial averages of this homogeneity measure H in the posterior, the climatologies of the PMIP3 ensemble members, and the reference climatology ~~. We expect a~~ are compared. A reconstruction with a good degree of smoothing is expected to have similar spatial homogeneity than the PMIP3 ensemble and the reference climatology, as H depends mainly on local features like orography or land-sea contrasts, and we expect these features to affect the local climate of the MH similarly than today's climate since reconstructions suggest only small changes in topography between the MH and today. For MTWA, ~~we get a posterior mean of 1.48~~ the posterior mean value is 1.41 K (90% CI: (~~1.40~~ 1.31 K, ~~1.57~~ 1.53 K)), which is in agreement with 1.39 K for the reference climatology and values between 1.08 K and 1.54 K for the PMIP3 climatologies. The heterogeneity of MTCO is higher than of MTWA, but ~~again,~~ the mean posterior value of ~~2.18~~ 2.54 K (90% CI: (~~2.09~~ 2.33 K, ~~2.27~~ 2.76 K)) is ~~in concordance with~~ of comparable magnitude as the reference climatology (2.02 K) and the PMIP3 climatologies (between 1.89 K and 2.41 K). From these results, ~~we deduce~~ it is deduced that the posterior has a reasonable degree of spatial smoothing.

4.2 ~~Added value of the reconstruction~~

~~The skill of the reconstruction is measured using the BSS, defined in Eq. (??), in cross-validation experiments. For positive BSS values,~~

4.1.1 Comparison of unconstrained PMIP3 ensemble and posterior distribution

By comparing the posterior with the prior and the local reconstructions, it can be seen that for most areas with nearby proxy records the posterior mean resembles the local reconstructions more than the PMIP3 ensemble mean. This shows that the uncertainty in the prior distribution is large enough to lead to a reconstruction which is mostly determined by proxy data, where available. The posterior MTWA mean is warmer in northern Europe than the prior mean and cooler in southern and eastern Europe. For MTCO, the posterior mean is much warmer than the prior mean in Fennoscandia and slightly cooler in southern Europe.

The posterior weights λ of the PMIP3 ensemble members are a combination of the prior distribution of λ and the likelihood of C_p for each combination of ensemble member weights (see Appendix C for details). λ provides information about which combination of ensemble members fits best to the proxy data. In our reconstruction, the ~~posterior distribution is superior to the prior, which means that constraining the~~ MPI-ESM-P climatology has the highest posterior weights (mean of 0.485) (see Fig. 8), followed by the EC-Earth-2-2 climatology (posterior mean of 0.154) and the Had-GEM2-CC climatology (posterior mean of 0.104). Note, that the weights of the MPI-ESM-P and the EC-Earth-2-2 are the only one that are on average higher than

the prior mean of 1/7. The large differences of the weights are a result of the large differences between the ensemble member climatologies. Because there is less uncertainty in the local MTWA reconstructions, it is the major variable for determining the posterior weights. Among all included models, the MPI-ESM-P simulation is closest to the dipole structure with MTWA warming in northern and cooling in southern Europe, which explains the high model weight.

4.1.2 Added value of the reconstruction

CVEs provide inside into the value that is added to the unconstrained PMIP3 ensemble~~by proxy data adds valueto the reconstruction.~~, represented by the process stage Eq. (7), by constraining it with the Simonis et al. (2012) synthesis. To quantify the added value, the BS from Eq. (20) is calculated for the unconstrained process stage, which is called BS(Prior), and compared to the BS of the posterior, BS(Posterior), calculated from leave-one-out CVEs. Then, the Brier skill score (BSS)

$$\mathrm{BSS} := \frac{\mathrm{BS(Prior)} - \mathrm{BS(Posterior)}}{\mathrm{BS(Prior)}} \tag{21}$$

is computed, which is a measure of the added value of the spatial reconstruction. For positive BSS values, the posterior distribution is superior to the prior. On the other hand, the posterior distribution is inferior to the prior for negative values. This would indicate inconsistencies in the local proxy reconstructions, a scaling mismatch of simulations and data, or the existence of spurious correlations in the prior covariance matrix.

For most left-out proxy samples, the BSS is positive (~~80~~68.9% of grid boxes~~with left-out pollen data~~) with a median of ~~0.39~~ 0.28 (Fig. 9). The BSS values are predominantly positive for all regions but the British Islands and ~~Norway~~the Alps. This indicates a high consistency of the reconstruction in large parts of the domain~~due to a good fit of the local reconstructions with the spatial correlation structure~~. In particular, ~~we see a~~ consistent MTWA cooling ~~south of 54° N~~ in southern and eastern Europe in the local reconstructions compared to the prior distribution ~~. This leads to a mean~~ leads to cooling and reduction of uncertainty in the posterior ~~which is transported to grid boxes without proxy data, including those with left-out proxy data, via spatial correlations and higher weights for cooler ensemblemembers~~compared to the unconstrained PMIP3 ensemble. Similarly, the consistent MTWA and MTCO warming of the local reconstructions in the north-eastern part of the domain lead to positive BSS values.

The persistent negative BSS values for the British Islands ~~and Norway~~ cannot be explained by data outliers but warrant a ~~more~~ systematic issue. For ~~these two regions~~this region, the uncertainty in the local reconstructions is larger than for other areas~~(Fig. 4)~~, such that the local proxy records constrain the posterior less than the posterior ensemble member weights and some of the more ~~distanced~~ distant proxy records. ~~The spatial correlations and degenerated weights (see Sect. ??) lead~~ This leads to a reduction of the posterior uncertainty compared to the unconstrained PMIP3 ensemble ~~(Fig. 7e, 7f)~~ but without improving the concordance of the mean state with the local reconstructions, which in turn results in negative BSS values. In and near the Alps, ~~BSS values are near zero indicating no added value from constraining the PMIP3 ensemble by proxy data. This~~ negative BSS might be a result of not accounting enough for orographic effects in the different sources of information.

4.2 Posterior ensemble member weights

The posterior weights of the PMIP3 ensemble members are determined by ω and z. The posterior of ω is a combination of its prior and the distribution of model choices z. As we defined a Jeffrey's prior and choose only one ensemble member in each MCMC step, the posterior of ω is a flattened version of the posterior of z (Fig. 8). z is more relevant for the posterior of the state C_p as it determines which ensemble member is chosen in each MCMC step. The posterior of z combines the distribution of ω and the likelihood of C_p for each of the ensemble members (see Appendix C). The latter term is more informative and therefore more important for determining z.

In our reconstruction, the posterior weights are dominated by the MPI-ESM-P climatology, with a posterior mean of z of 0.98 (blue diamonds in Fig. 8). The degeneracy of weights can be explained by large differences between PMIP3 climatologies. Because there is less uncertainty in the local MTWA reconstructions, it is the major variable for determining the posterior ensemble member weights. Among all included models, the MPI-ESM-P simulation is closest to the dipole structure with MTWA warming in northern and cooling in southern Europe, which explains the high model weight.

The prior climate state is changed in two ways during the inference: The ensemble member weights are updated, and each of the mixture components is modified according to the proxy data. By comparing the posterior with the prior and the local reconstructions, it can be seen that for most terrestrial areas the posterior mean resembles the local reconstructions more than the PMIP3 ensemble mean. This shows that the uncertainty in the prior distribution is large enough to lead to a reconstruction which is mostly determined by proxy data wherever available. The main exceptions are the British Islands and Norway, where the local reconstructions are not informative enough to strongly affect the joint signal from posterior ensemble member weights and farther away proxy samples. The aforementioned changes of the prior lead to a posterior mean which is cooler than the prior mean for MTWA for most of the domain except northern areas. For MTCO, the posterior mean is much warmer in northern Europe than the prior mean and slightly cooler in southern Europe.

4.2 Sensitivity with respect to the glasso penalty parameter

As described in Sect. **??**, the glasso algorithm, which regularizes the covariance matrix estimated from the PMIP3 ensemble, requires the specification of a penalty parameter ρ. Values of at least 0.3 produce numerically stable matrices. Larger values lead to smaller correlations, and therefore higher posterior uncertainties due to less spatial transfer of information from the proxy data. To test the influence of ρ on the reconstruction results, we compute reconstructions and cross-validations for $\rho = 0.3, 0.5, 0.7, 1.0, 2.0$.

The fraction of the penalty term on the total cost given by Eq. (**??**) increases from 79.5% for $\rho = 0.3$ to 98.5% for $\rho = 2.0$ (Fig. **??**) showing the generally high influence of the penalty term due to the small ensemble size and the increasing importance for larger ρ. While the mean BS is lowest for $\rho = 0.3$ with 0.37 and increases with ρ, the differences are small in total with 0.44 being the highest value for $\rho = 2.0$ (Fig. **??**). The mean value of the posterior climate and the posterior ensemble member weights are insensitive to changes in ρ with differences in the mean spatial average of at most 0.3 (Fig. **??**) as well as no substantial regional differences, and mean z values for MPI-ESM-P of more than 0.98 in all cases. On the other hand, the

posterior uncertainty grows substantially for larger ρ values. While the spatially averaged size of point-wise 90% CIs is 3.03 for $\rho = 0.3$, the value increases to 6.34 for $\rho = 2.0$ (Fig. **??**).

4.2 Joint versus separate MTWA and MTCO reconstructions

To compare

4.1.1 Joint versus separate MTWA and MTCO reconstructions

To study the effect of reconstructing MTWA and MTCO jointly compared to separately, additional reconstructions with only one climate variable are computed. Note that the interactions of MTWA and MTCO are twofold in the joint reconstruction: (a) the response functions have an interaction term in the logistic regression Eq. (**??**3), and (b) the process stage Eq. (**??**) contains joint ensemble member weights for MTWA and MTCO as well as inter-variable correlations in the covariance empirical correlation matrix.

The separate MTWA reconstruction is on average around 0.5 K warmer than the joint reconstruction, with the spatially averaged posterior mean being 0.63 warmer. On the other hand, whereas the spatially averaged posterior mean of the separate MTCO reconstruction is 1.05 0.83 K cooler (Table 3). Hence, the seasonal difference is smaller in the joint reconstruction, due to smoothing from the PMIP3 ensemble and slightly positive correlations between MTWA and MTCO in most of the joint local reconstructions. The MTWA only reconstruction is warmer in most land areas, with largest differences in southern and eastern Europe, but most of the differences are not almost never significant on a 5% level (Fig. 10a). As this part of the domain is best constrained by proxy data, and because the posterior ensemble member weights are similar to the joint reconstruction (mean $\bar{z}$ of 0.93 λ_k of 0.419 for MPI-ESM-P, Fig. 8), it is likely that the additional warming is due to the missing interaction in the transferfunctiontransfer function. On the other hand, the posterior ensemble member weights are very different change a lot for the separate MTCO reconstruction, with HadGEM2-CC and MRI-CGCM3 being the models with the highest weights (mean $\bar{z}$ of 0.65 and 0.34λ_k of 0.426 and 0.199, respectively). Together with the less constrained transferfunctions transfer functions for MTCO than MTWA and the missing interaction term, this leads to a cooler reconstruction for all most areas but some parts of central Europe and the Mediterranean (Fig. 10b). The cooling is strongest in Scandinavia, the British Islands, the Norwegian Sea, and the Iberian peninsula, but almost never significant on a 5% level. As these are the regions which are least constrained by proxy data(Fig. 4d), choosing different PMIP3 ensemble members affects the reconstruction more than in other areas. Many of the differences in Fennoscandia and the south-western part of the domain are significant (5% level), which indicates that for these areas our method might underestimate the reconstruction uncertainty.

The missing interaction of MTWA and MTCO also leads to a higher uncertainty in the separate reconstructions, in particular in areas which are not well central Europe, where MTCO is best constrained by proxy data. For MTWA, the difference is relatively small with spatially averaged only 0.2 larger point-wise 90% CIs. The spatial average of the point-wise 90% CIs for MTCO is 0.96 larger than The reconstruction uncertainties are of similar magnitude in the joint reconstruction, showing again that reconstructing the variables jointly has a larger effect on MTCO than on MTWA. The larger MTCO uncertainties are mostly due to much larger CIs at the Norwegian Sea (Fig. 10d)and the separate reconstructions (Table 3), because the

interactions in the transfer functions do not reduce the marginal uncertainties and the shrinkage target Φ does not contain correlations of MTWA and MTCO.

The ~~sign of the BSS~~ BSS pattern in the MTWA only reconstruction is mostly the same than in the joint reconstruction ~~, but the magnitude is smaller for most grid boxes (~~except for slightly positive skill in the British Islands (Table 3, Fig. 10e). This shows that the added value of the joint reconstruction compared to the unconstrained PMIP3 ensemble is mainly determined by the MTWA reconstruction. ~~In particular, the negative BSS at the British Islands and Norway, which is found in the joint reconstruction, is also present in the MTWA only reconstruction but not in the MTCO only reconstruction. For most of the domain, the BSS values of the MTCO only reconstruction are small (Fig. 10f), which indicates no or only little added value from including proxy data. Particularly in central Europe, the values are very close to zero because~~ On the other hand, the added value of the MTCO only reconstruction is much smaller (Table 3, Fig. 10f) due to larger uncertainties in the local MTCO reconstructions.

The results show that the more constrained local MTWA reconstructions lead to a higher influence on the joint reconstruction than the local MTCO reconstructions. Therefore, the ~~comparably small PMIP3 ensemble spread (Fig. 1d) is just weakly constrained by local reconstructions~~ MTCO only reconstruction differs more from the MTCO estimate in the joint reconstruction. Reconstructing MTWA and MTCO jointly should in theory lead to a physically more reasonable reconstruction by creating samples drawn from the same combination of ensemble members. In the example of this study, this effect can be seen from the smaller seasonal differences in the joint than in the separate reconstructions. On the other hand, Rehfeld et al. (2016) show that multi-variable reconstructions from pollen assemblages can be biased when signals from a dominant variable are transferred to a minor variable. While the PITM model might be less sensitive to this issue than the weighted averaging transfer function used in Rehfeld et al. (2016) because it better respects the larger MTCO uncertainties and because it is unclear whether taxa occurrence is similarly susceptible to the issue than pollen ratios, it will be subject to future work to study whether joint or separate reconstructions lead to more reliable results.

5 Discussion and possible extensions

5.1 Robustness of the reconstruction

Our approach is designed with the goal of being more suitable for sparse data situations than standard geostatistical models. ~~In a subsequent paper, we plan to apply the method to data from the Last Glacial Maximum, for which considerably less proxy samples are available than for the MH.~~ To understand the robustness of ~~our method~~ the Bayesian framework with respect to the amount of data included in a proxy synthesis, ~~we perform~~ five experiments with only half of the samples are performed, which are either selected to retain the spatial distribution of proxy samples or chosen randomly. In all of the tests, the general spatial structure of the posterior distribution, including the anomaly patterns, is preserved~~. Only the northern part of the MTCO reconstruction, especially~~ , even though depending on the chosen proxy samples the local anomalies and magnitude of changes varies which should be expected when such a large portion of the already sparse data is left out. Only the Norwegian Sea ~~,~~ in the MTCO reconstruction changes substantially in some experiments. ~~While in each experiment MPI-ESM-P remains the favoured~~

ensemble member, other members can reach weights, which are higher than in the prior. The spatially averaged temperatures Plots from the experiments with reduced proxy samples are provided in the Supplement.

The mean spatial averages differ up to 0.6 K for MTWA as well as the spatial homogeneity H stay within the uncertainty ranges of the reference reconstruction. The uncertainty structure of all tests is similar to the reference reconstruction but the lower number of proxy records leads to an increase of the overall uncertainty. For example, the spatially MTCO, but none of the changes is significant on a 5% level. In contrast, the uncertainty estimates are consistent across all experiments with reduced proxy samples with averaged point-wise 90% CIs that grow by up to 0.5 0.4 K . This number is still lowcompared to the overall uncertainty keeping in mind that the number of proxy samples from the reconstruction with the full proxy synthesis. Considering that half the proxy samples are left out, this number is low. In all experiments, the spatial homogeneity H is not significantly different from the values reported in Table 3, which shows that the spatial homogeneity as a local feature is more controlled by the process stage than the proxy data. In all but one experiment, the MPI-ESM-P remains the ensemble member with the largest weight λ_k, and the three ESMs which are favoured neither in the MTWA nor in the MTCO reconstruction retain very low weights in all experiments. But depending on whether proxy samples in which MTCO is much less constrained than MTWA are removed or not, the weights of the four models with the highest values in the joint and separate reconstructions can vary such that in one case the average weight of the EC-Earth-2-2 is reduced by 50%0.1 higher than the weight of MPI-ESM-P. These changes of the weights explain the MTCO changes in the Norwegian Sea as this is the region which is most influenced by the ensemble member weights. The experiments show that our method the reconstruction is robust with respect to the number of proxy samples as long as the remaining samples are informative and relatively uniformly distributed across space. In our example, this is not the case for the Norwegian Sea. Combining pollen records with sea surface temperature proxies could potentially overcome this issue.

The more constrained local MTWA reconstructions lead to a higher influence on the joint reconstruction than the local MTCO reconstructions. Therefore, the MTCO only reconstruction differs more from the MTCO estimate in the joint reconstruction. Reconstructing MTWA and MTCO jointly should in theory lead to a physically more reasonable reconstruction by creating samples drawn from the same ensemble member and therefore corresponding to a physically consistent MTWA and MTCO simulation. On the other hand, Rehfeld et al. (2016) show that multi-variable reconstructions can be biased when signals from a dominant variable are transferred to a minor variable. While we believe that the PITM model is less sensitive to this issue than the weighted averaging partial least squares (WA-PLS) method used in Rehfeld et al. (2016) because it better respects the larger MTCO uncertainties, it will be subject to future work to study whether joint or separate reconstructions lead to more reliable results.

The large PMIP3 ensemble spread for most grid boxes shows that the prior distribution, which is calculated from the ensemble, contains a wide range of possible states. In areas which are well constrained by proxy data, this large total uncertainty leads to a reconstruction which depends little on the climatologies of the ensemble members. Hence, in these areas, the reconstruction is not much influenced by the ensemble member weightssensitive to the particular formulation of the process stage (compare with Supplement). This shows that our method is applicable despite well-known model-data mismatches for the MH (Mauri et al., 2014). On the other hand, the spatial correlation structure estimated from the ensemble controls the smoothness

of the reconstruction and the controls the spread of local information into space. As the correlation structure becomes more robust for larger ensemble sizes, it would be desirable to have larger ensembles available for future applications of our method. In addition, the current ensemble covers only modelling uncertainty, while internal variability and uncertainties in forcings are not explicitly included. A possible extension of our method would use ensembles which account for all these types of uncertainty in a structured way allowing the estimation of a better constrained prior distribution. For most of terrestrial Europe, the posterior mean is more similar to the local proxy reconstructions than to the prior mean from the unconstrained ensemble. This shows that the spatial structure is the more important part of the prior distribution compared to the mean state as long as the prior uncertainty is larger than the proxy uncertainty . Therefore, our method is applicable despite well-known model-data mismatches for the MH (Mauri et al., 2014)Different formulations of the spatial correlation matrix can lead to substantially different reconstructions in regions that are not well constrained by proxy samples and in particular a spatial covariance with too few spatial modes can lead to overly optimistic uncertainty estimates.

5.2 Comparison with previous reconstructions

Several reconstructions of European climate during the MH have been compiled previously. Here, we compare our reconstructions to those of Mauri et al. (2015), Simonis et al. (2012), and Bartlein et al. (2011).

Mauri et al. (2015) use a plant functional type modern analogue transferfunction transfer function and a thin splate spline interpolation for pollen samples stemming mostly from the European pollen database. The simpler interpolation method allows the treatment of the samples as point data and the interpolation of the local reconstructions to an arbitrary grid. Among other variables, summer and winter temperatures are reconstructed. We find a dipole anomaly structure similar to Mauri et al. (2015) in our reconstructions, with mostly positive anomalies in northern Europe and negative anomalies in southern Europe. In Mauri et al. (2015) as well as in our reconstruction, the Alps are the only region with significant warming in central and southern Europe for summer temperature. Generally, the amplitude of summer anomalies in the two reconstructions are is similar, although locally there are differences with cooler anomalies at the British Islands and southern over south-western Fennoscandia in our reconstruction as well as warmer anomalies in south-eastern Europe and Finland. For winter temperatures, we do not find the cooling over the cooling in the Mediterranean and the British Islands that is reported is less pronounced and spatially less consistent in our reconstruction than in Mauri et al. (2015). As for summer temperatures, we find smaller anomalies in southern Fennoscandia. In contrast, our reconstruction shows higher anomalies in northern Scandinavia. In all other regions, the amplitude of anomalies are similar between the two reconstructions despite some local disagreements.

The same pollen dataset and a closely related transferfunction than in our reconstruction another version of PITM are used in Simonis et al. (2012) to reconstruct July and January temperature, such that differences between the two reconstructions are mostly related to the different smoothing technique. Simonis et al. (2012) minimize a cost function which combines pollen samples with an advection-diffusion model that is driven by insolation changes between the MH and today. In Simonis et al. (2012), the dipole structure is not found in the same way than in our reconstruction, which might be due to the different way how regions which are not well constrained by proxy data are treated. Both reconstructions share positive summer temperature anomalies in northern Europe as well as negative anomalies in central Europe and the Iberian peninsula. Unlike our

reconstruction, Simonis et al. (2012) find positive anomalies in ~~western and~~ south-eastern Europe. For winter temperatures, the reconstruction of Simonis et al. (2012) shows an east-west dipole in contrast to the north-east to south-west dipole in our reconstruction. This different structure might be due to the smaller proxy data control of the winter reconstructions, which leads to a higher importance of the interpolation schemes.

A reconstruction designed to evaluate the PMIP3 simulations was provided by Bartlein et al. (2011). They combine a large number of pollen based local reconstructions from the literature to produce a gridded product of six climate variables including MTWA and MTCO. In contrast to our reconstruction, the used local reconstructions are not smoothed across space but only within a grid box. Their results show a dipole structure but less pronounced than in our reconstruction and in Mauri et al. (2015). In particular, they find a cooling for eastern Fennoscandia in summer, a much smaller warming of northern Fennoscandia than our reconstruction, and a warming in Germany and France. On the other hand, the reported anomalies in Bartlein et al. (2011) for the Mediterranean and eastern Europe are similar to our results.

The comparisons show that patterns like the dipole type anomaly structure, which are not present in the PMIP3 ensemble, seem to be consistent across pollen ~~transferfunctions~~transfer functions. While some of the differences between the existing literature and our results can be explained by the used ~~transferfunctions~~ transfer functions and proxy syntheses, the choice of an appropriate interpolation method plays an important role, too, especially in areas with very sparse and weakly informative proxy data and to determine the degree of smoothing.

~~To facilitate more quantitative comparisons of reconstructions and allow a fair testing of modelling assumptions, we plan to expand our current framework to~~

5.3 Climate model inadequacy and process stage structure

To account for inadequacies of climate models to simulate past climate states, we introduced flexible ensemble member weights λ and the shrinkage matrix approach which combines the empirical covariance matrix of the climate ensemble with a matrix that is derived from an independent physically motivated model. Combining ensemble filtering methods with additional techniques to correct model biases in a physically consistent way is an important but also challenging direction of future work on climate field reconstructions as a balance has to be found between under-dispersion of the posterior distribution by inducing physcial structure and overfitting to noisy proxy data by enhancing the degrees of freedom. Beyond the strategies implemented in this work, some directions that can be envisaged are the increase of permitted spatial structures in the prior mean by adding patterns calculated from alternative physically motivated models, and the introduction of multiple shrinkage targets in the spatial covariance matrix (Gray et al., 2018).

In addition to those extensions of the current filtering framework, different types of ~~transferfunctions and interpolation techniques. In particular, an implementation of a parametric process stage using the model of Simonis et al. (2012), which is independent of climate simulation output, is envisaged~~process stages can be envisaged which are independent of simulations with ESMs but still computed from physical principles. One example are stochastic energy balance models, which are simple stochastic climate models that simulate the energy fluxes between external forcing, atmosphere, and surface. These types of models have become important tools in studies of climate variability Rypdal et al. (2015) and could be reformulated as process

stage in BHMs. The advantage of such a process stage is that it can be fully integrated in the BHM instead of using an offline run with an ESM which does not learn from the proxy data. In addition, spatial reconstructions that are independent of ESM output, are more suitable for comparisons of proxy data and climate simulations than the Bayesian filtering approach.

5.4 ~~Towards global and spatio-temporal reconstructions~~

~~A~~ Another valuable extension of our approach would be the computation of reconstructions on hemispheric to global scales. In this case, a more flexible structure for the ensemble member weights is desirable to account for the varying skill of climate models in different regions. On the other hand, the weights should not change too much within small domains to avoid unreasonable spatial heterogeneity. ~~A way to combine those two aspects would be a cost function for the model weights, which combines a spatial smoothing term and the local fit of the ensemble members with proxy data. The balancing of those two terms to optimize the degree of smoothing and to respect the local proxy reconstructions is a challenging task itself.~~ An additional problem is that the ~~penalty parameter ρ in the glasso algorithm is applied locally, such that~~ stationary shrinkage target matrix would no longer be a good approximation as different regions require different ~~values of ρ to result in a numerically stable covariance matrix and to optimize the criteria described in Sect. **??**. While glasso allows spatially varying penalty parameters, optimizing the values in an objective way is a complicated issue. An alternative would be the reformulation of the penalty term in Eq. (**??**), such that not only local properties are optimized but also the structure between subregions. One possibility to achieve this could be the use of an additional L_2-penalization as in Zou et al. (2006)~~target matrices. This means that non-stationarities have to be integrated in the shrinkage target. A straightforward to accomplish this way is to introduce non-constant parameters in the stochastic partial differential equation behind the Matérn model (Lindgren et al., 2011).

~~Computing spatio-temporal reconstructions with our approach currently faces two challenges: The increasing dimensionality due to the additional temporal component, and the small number of available transient paleosimulations with comprehensive ESMs~~The implementation of these different types of process stages would facilitate more quantitative comparisons of reconstructions and allow a fair testing of modelling assumptions by using ITEs and CVEs. In particular, ~~for time scales beyond the last millennium, there is currently no multi-model ensemble available which is comparable to the PMIP3 simulations for the MH. To deal with the higher dimension a method to reduce the complexity is necessary. One way could be a separation of time and space but this might lead to a breakup of the temporal structure in the climate simulations resulting in inconsistencies in the reconstructions. More generally, any method to reduce the degrees of freedom in the reconstruction can lead to a significant underestimation of uncertainty if either the dimensionality is reduced too much or if the retained spatio-temporal patterns are too far from reality. A possibility to deal with both problems would be a Bayesian framework that combines a flexible parametric interpolation method featuring a large number of degrees of freedom with additional constraints from transient as well as time slice simulations to increase the physical consistency~~the reformulation of existing climate field reconstruction techniques as BHM (Tingley et al., 2012) offers many model comparison techniques that are currently not available as the borders between very different statistical techniques especially to estimate reconstruction uncertainties have to be crossed.

6 Conclusions

We presented a new method for probabilistic spatial reconstructions of paleoclimate. The approach combines the strengths of pollen records, which provide information about the climate state in a small scale domain, and of climate simulations, which downscale forcing conditions to physically consistent regional climate patterns. Thus, we reconstruct physically reasonable spatial fields, which are consistent with a given proxy synthesis. ~~Bayesian modelling combined with MCMC methods is well suited for paleoclimate reconstructions as it can include multiple sources of information together with the corresponding uncertainties, and facilitates the separate modelling of proxy-climate transferfunctions and stochastic interpolation between proxy samples.~~ Our framework can deal with probabilistic ~~transferfunctions~~transfer functions, which are non-linear and non-Gaussian, such that an extension to a wide range of proxies and associated ~~transferfunctions~~ transfer functions is possible.

Using ITEs and CVEs, we showed that robust spatial reconstructions with Bayesian filtering methods that exhibit small biases and are not under-dispersed are possible as long as the statistical framework is flexible enough to account for deficiencies of climate simulations and to avoid filter degeneracy, which can emerge due to small ensemble sizes and biases in climate simulations. However, all these properties can be lost when the model does not adequately account for those two issues. The resulting model, which is used for spatial reconstructions of European MH climate, uses a weighted average of the involved ensemble member climatologies and a shrinkage matrix approach for spatial interpolation and structural extrapolation of the proxy data. A strong over-dispersion is found in ITEs which is less harmful than under-dispersion because it avoids the drawing of overconfident conclusions.

We apply our framework to reconstruct MTWA and MTCO in Europe during the MH using the proxy synthesis of Simonis et al. (2012) and the PMIP3 MH ensemble. Brier scores from cross-validations reveal that the spatial reconstruction predominantly adds value to the unconstrained PMIP3 ensemble, and analyses of the spatial homogeneity of the posterior distribution indicate a reasonable degree of smoothing through the induced correlation structure. The large scale spatial patterns of our reconstruction are in agreement with previous work (Mauri et al., 2015; Bartlein et al., 2011). As the posterior mean is more similar to the local proxy reconstructions than to the prior mean for most terrestrial areas, we see that the main role of the simulation ensemble is to provide a spatial ~~correlation~~ covariance structure and that a reconstruction, which is in line with reconstructions that do not include simulation output, is possible despite well-known model-data mismatches (Mauri et al., 2014). Our framework provides a way to quantitatively test hypotheses in paleoclimatology and to assess the consistency of a given proxy synthesis. This includes the fit with large scale structures, the spatial homogeneity, and the quantitative quality control of the proxy data by identification of potential outliers.

~~In future work, we plan to apply our framework to new multi-proxy syntheses like the one currently compiled in the German PalMod project (www. palmod. de). This includes the incorporation of new proxies and transferfunctions as well as the use of larger spatial domains. In particular, the lower degree of uncertainty reduction for the marine parts of the domain suggests an inclusion of supplementary marine proxies. Moreover, the addition of new high resolution paleosimulations for example from the ongoing PalMod and Paleoclimate Modelling Intercomparison Project Phase 4 (PMIP4) projects (Kageyama et al., 2018)~~

~~will lead to a better quantification of uncertainties in the prior mixture distribution and, subsequently, to more robust reconstructions.~~

~~The hierarchical structure of the framework permits the replacement of the spatial interpolation module by other types of stochastic interpolation, e. g. geostatistical models (Tingley et al., 2012) or simple physical models (Gebhardt et al., 2008). The comparison of our current framework with these different modelling assumptions will lead to an improved understanding of the spatial patterns of past climate .~~

Code and data availability. R code for computing reconstructions with the presented Bayesian framework is provided in a Bitbucket repository available under https://bitbucket.org/nils_weitzel/spatial_reconstr_repo. The pollen and macrofossil synthesis is published in Simonis (2009). It is available as an R list object in the Bitbucket repository. The PMIP3 MH simulations are available in the CMIP5 archives. In this study, they were downloaded from the DKRZ long term archive CERA (https://cera-www.dkrz.de). The modern climate data was downloaded from the University of East Anglia Climate Research Unit, available at http://www.cru.uea.ac.uk/data/. The vegetation data for transfer function calibration was provided by Thomas Litt and Norbert Kühl.

Appendix A: Determination of glasso penalty parameter

To determine the glasso penalty parameter ρ, we first recognize that for values smaller than $\rho = 0.3$ the resulting matrices become numerically unstable in our application due to the small ensemble size. Five values for ρ were tested: 0.3, 0.5, 0.7, 1.0, and 2.0. Larger values lead to sparser precision matrices and therefore to smaller spatial correlations. This in turn means that the local information from the proxy data is spread less into space. For each of the five parameters, we perform CVEs with the RM and compare the resulting BS (see Sect. 4.1.2). While the smallest penalty parameters have the best mean BS, the differences are generally small (see Supplement). On the other hand, the influence of the penalty term in Eq. (14) on the overall regression increases from 79.5% for $\rho = 0.3$ to 98.5% for $\rho = 2.0$. Based on these diagnostics, we choose $\rho = 0.3$ for the reconstructions in Sect. 4 to get a numerically stable covariance which performs at least as good as other choices of ρ in CVEs, and is comparably little influenced by the penalty term. The sensitivity of the reconstruction with respect to ρ is further studied in the Supplement.

Appendix B: Shrinkage target matrix

The shrinkage target Φ is defined on a regular lat-lon grid following the SPDE approach of Lindgren et al. (2011). This approach allows a computationally efficient approximation of Matérn covariances with parameters that are physically motivated in the context of stochastic Laplace equations, which model diffusive transport of a stochastic forcing. Setting $\zeta^2 := \frac{8}{\rho^2}$, the range parameter ρ is rewritten as a scale parameter ζ^2. Moreover, we let the anisotropy parameter ν parameterize a diagonal diffusion matrix $\upsilon := \mathrm{Diag}\left(\sin(\nu\pi/2), \cos(\nu\pi/2)\right)$. Then, the stochastic partial differential equation from which Φ is deduced

is given by

$$\left(\zeta^2 - \nabla \cdot (\upsilon \nabla)\right) X(x) = \mathcal{W}(x). \tag{B1}$$

$\mathcal{W}$ denotes white noise, and X is the stationary Gaussian random field that solves Eq. (B1). Discretizing this equation with linear finite elements, and using a diagonal approximation of the involved mass matrix leads to a GMRF with correlation matrix $\hat{\Phi}$ (Lindgren et al., 2011). We use degree as distance unit on the regular lat-lon grid instead of m which means that the decorrelation length depends on the latitude. While this is counterintuitive on first glance, it better reflects the mostly shorter decorrelation lengths in higher latitudes and leads to a better model fit. Φ is constructed from $\hat{\Phi}$ by combining spatial correlation matrices of type $\hat{\Phi}$ for each climate variable, that is jointly reconstructed, but with different parameters ρ and ν for each climate variable, in a block diagonal structure.

Appendix C: Full conditional distributions

The Metropolis-within-Gibbs approach samples (asymptotically) from the full conditional distributions of each variable, i.e. the distribution of the variable given all other variables. Some variables are treated block-wise to account for correlations between them, while others are updated sequentially if they are independent from each other or the joint distribution is too complicated for efficient sampling. In this appendix, we detail the conditional distributions that are used for sampling.

To sample the ~~transferfunction~~ transfer function parameters, we introduce PG distributed augmented variables γ_l^T, where $T \in T(P)$ and $l = 1, ..., L(T)$ are the number of observations for taxa T (Polson et al., 2013). γ_l^T is PG distributed given $\beta^T = (\beta_1^T, ..., \beta_6^T)$ and climate data $C(l) = (C_1(l), C_2(l))$, where C_1 and C_2 denote MTWA and MTCO. More precisely, the full conditional distribution is given by

$$\gamma_l^T \mid \beta^T, C(l) \quad \sim \quad \text{PG}(n = 1, X_{C(l)} \cdot \beta^T), \tag{C1}$$

$$\text{where } X_{C(l)} := \left(1, C_1(l), C_2(l), C_1(l)C_2(l), C_1(l)^2, C_2(l)^2\right). \tag{C2}$$

Including the Gaussian prior defined in Sect. ??3.2, the full conditional of β^T is ~~multivariate normal~~ Gaussian distributed:

$$\beta^T \mid P_m^T, P_p^T, \gamma_1^T, ..., \gamma_{L(T)}^T \quad \sim \quad \mathcal{N}\left(V_\gamma X^t \kappa^T, V_\gamma\right), \tag{C3}$$

$$\text{where } V_\gamma := \left(X^t \Gamma X + B^{-1}\right)^{-1}. \tag{C4}$$

Here, X is a matrix with rows $X_{C(l)}$ for $l = 1, ..., L(T)$, Γ is a diagonal matrix with entries γ_l^T, B is the 6×6 prior covariance matrix of β^T, and κ^T is a vector with entries $\left(P^T(l) - \frac{1}{2}\right)$, where $P^T(l)$ is the presence or absence of taxa T in observation l. In our case, B is a diagonal matrix with all values equal to 10. Details on the definition of PG variables and the augmented Gibbs sampler can be found in Polson et al. (2013).

~~The~~ Sampling from ϑ depends on the specific version of the process stage which is used. In the RM, $\lambda = (\lambda_1, ..., \lambda_K)$ is influenced by its prior and by the Gaussian distribution of C_p given λ:

$$\lambda | \, C_p \quad \sim \quad \text{Dirichlet}\left(\tfrac{1}{2}, ..., \tfrac{1}{2}\right) \mathcal{N}\left(C_p \,\middle|\, \sum_{k=1} \lambda_k \mu_k, \Sigma_{\text{prior}}\right). \tag{C5}$$

This full conditional does not follow a probability distribution from which we can sample directly. Therefore, a random walk type Metropolis-Hastings update is used for updating λ.

In the KM, the full conditional of $\omega = (\omega_1, ..., \omega_K)$ is Dirichlet distributed given $z = (z_1, ...z_K)$ and its Dirichlet prior:

$$\omega \,|\, z \quad \sim \quad \text{Dirichlet}\left(\tfrac{1}{2} + z_1, ..., \tfrac{1}{2} + z_K\right). \tag{C6}$$

Given ω and C_p, z is ~~multinomially distributed:~~ categorically distributed:

$$z \,|\, \omega, C_p \quad \sim \quad \text{MultinomialCat}\left(n = 1, \alpha_1, ..., \alpha_K\right), \tag{C7}$$

$$\text{where} \;\; \alpha_k := \frac{\omega_k \cdot \exp\left(-\frac{1}{2}(C_p - \mu_k)^t \, \Sigma_{\text{prior}}^{-1} (C_p - \mu_k)\right)}{\sum_{i=1}^{K}\left(\omega_i \cdot \exp\left(-\frac{1}{2}(C_p - \mu_i)^t \, \Sigma_{\text{prior}}^{-1} (C_p - \mu_i)\right)\right)} \tag{C8}$$

If shrinkage covariance matrices are used, the parameters (α, ρ, ν) have to be chosen in each MCMC step from one of the $K = 7$ predefined parameter sets. We use a uniform prior, i.e. all seven parameter sets have the same prior probability. Then, the full conditional of τ which indexes the parameter set, is given by

$$\tau \,|\, C_p, \hat{\mu} \quad \sim \quad \mathcal{N}\left(C_p \,\middle|\, \sum_{k=1} \hat{\mu}, \Sigma_{\text{prior}}(\alpha_\tau, \rho_\tau, \nu_\tau)\right), \tag{C9}$$

where $\hat{\mu}$ is given according to conditioning on the process stage parameters of the GM, RM, or KM. We update τ using an Metropolis-Hastings step with independent proposals, which choose $\tau = k$ with probability $\frac{1}{K}$.

We update $C_p(x)$ for $x \in x_P$ sequentially using random walk Metropolis-Hastings sampling. The set of all grid boxes ~~but~~ besides x is denoted by Y_x, and let $\Sigma_{\text{prior}}^{-1}(a, b)$ be the block of the inverse covariance matrix containing the rows a and columns b. Then, the full conditional distribution of $C_p(x)$ depends on the pollen samples $P_p(s)$ with location $x_s = x$, the climate $C_p(Y_x)$ at the other locations, and ~~the chosen model z with $z_k = 1$~~process stage parameters ϑ. It does not follow a standard distribution:

$$C_p(x) \,|\, P_p, C_p(Y_x), z\vartheta \;\sim\; \mathcal{N}\left(\tilde{\mu}_k(x), \left(\Sigma_{\text{prior}}^{-1}(x, x)\right)^{-1}\right) \prod_{\substack{s \\ \text{with} \\ x_s = x}} \prod_{T \in T(s)} \text{logit}\left(X_{C_p(x)} \cdot \beta^T\right), \tag{C10}$$

$$\text{where} \;\; \tilde{\mu}_k(x) := \mu\hat{\mu}_k(x) - \left(\Sigma_{\text{prior}}^{-1}(x, x)\right)^{-1} \Sigma_{\text{prior}}^{-1}(x, Y_x)\left(C_p(Y_x) - \mu\hat{\mu}_k(Y_x)\right). \tag{C11}$$

The step size of the random walk proposals is controlled by the conditional covariance $\left(\Sigma_{\text{prior}}^{-1}(x, x)\right)^{-1}$.

Conditioned on $C_p(x_P)$ and ~~z with $z_k = 1$~~ϑ, $C_p(x_Q)$ follows a Gaussian distribution:

$$C_p(x_Q) \mid C_p(x_P), z\vartheta \quad \sim \quad \mathcal{N}\left(\tilde{\mu}_k(x_Q), \left(\Sigma^{-1}_{\text{prior}}(x_Q, x_Q)\right)^{-1}\right), \tag{C12}$$

$$\text{where } \tilde{\mu}_k(x_Q) := \mu\hat{\mu}_k(x_Q) - \left(\Sigma^{-1}_{\text{prior}}(x_Q, x_Q)\right)^{-1} \Sigma^{-1}_{\text{prior}}(x_Q, x_P)\left(C_p(x_P) - \mu\hat{\mu}_k(x_P)\right). \tag{C13}$$

Appendix D: ~~Pseudo-code for MC^3~~ Metropolis coupled Markov chain Monte Carlo algorithm

~~For our reconstructions, we use $J = 2500$ iterations of $M = 30$ steps to draw altogether 75, 000 samples from $A = 8$ MCMC chains (see Sect. ??). The~~ As described in Sect. 3.4, the multi-modality of the KM in combination with the high dimensionality of the posterior makes the standard MCMC algorithm very inefficient. In our specific formulation the inefficiency is manifested in a very slow mixing of z, because conditioned on C_p the likelihood of choosing a new model z_k, from one MCMC step to the next one is very small. This problem could be shifted to other variables by integrating out z, but than the conditional Gaussian structure of C_p would be lost which would lead to new challenges for generating efficient MCMC strategies. Therefore, we apply a MC^3 ~~algorithm follows the pseudo-code in Algorithm ??~~or parallel tempering strategy. The original strategy from Geyer (1991) was adapted to parallel computer architectures by Altekar et al. (2004) and applied in a paleoclimate reconstruction problem by Werner and Tingley (2015).

~~If the modularized version is used, 75, 000 samples of the transferfunction parameters θ are drawn first. These are used for subsequent reconstructions which follow the pseudo-code without sampling the transferfunction parameters~~We run A MCMC chains in parallel, and after every M steps, we use an additional Metropolis-Hastings step to swap the states of the Markov chains a_1 and a_2 with probability $0 < p_{a_1,a_2} < 1$, where p_{a_1,a_2} is calculated from the Metropolis-Hastings odds ratio. The Markov chains are created by exponentiating the process stage and the data stage by constants $\nu_1 = 1 > ... > \nu_A > 0$. The first Markov chain ($\nu_1 = 1$) asymptotically retains the original posterior distribution for all variables, whereas the subsequent chains sample from a flatter posterior distribution, in which it is easier to jump from one mixture component to another. Following empirical testing, we run the European reconstructions with $A = 8$ parallel chains, levels $\nu_1 = 1, \nu_2 = 1.25^{-1}, ..., \nu_8 = 1.25^{-7}$, and swaps after every $M = 30$ steps. Pseudo-code for the MC^3 algorithm is given in the appendix.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This work was supported by German Federal Ministry of Education and Research (BMBF) as Research for Sustainability initiative (FONA, www.fona.de) through the Palmod project (FKZ: 01LP1509D). N. Weitzel was additionally supported by the National Center for Atmospheric Research (NCAR). NCAR is funded by the National Science Foundation. We acknowledge all groups involved in producing and making available the PMIP3 multi-model ensemble. We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for ~~CMIP. For CMIP~~ CMIP5. For CMIP5 the US Department of Energy's Program for

Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. We thank Douglas Nychka for helpful ideas to speed up the MCMC algorithm. We thank two anonymous referees and the editor for their interesting and helpful comments which facilitated a substantial improvement of this manuscript.

References

Altekar, G., Dwarkadas, S., Huelsenbeck, J. P., and Ronquist, F.: Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference, Bioinformatics, 20/3, 407–415, 2004.

Anderson, J. L. and Anderson, S. L.: A Monte Carlo Implementation of the Nonlinear Filtering Problem to Produce Ensemble Assimilations and Forecasts, Monthly Weather Review, 127, 2741–2758, 1999.

Annan, J. and Hargreaves, J.: A new global reconstruction of temperature changes at the Last Glacial Maximum, Climate of the Past, 9, 367–376, 2013.

Bartlein, P., Harrison, S., Brewer, S., Connor, S., Davis, B., Gajewski, K., Guiot, J., Harrison-Prentice, T., Henderson, A., Peyron, O., Prentice, I., Scholze, M., Seppä, H., Shuman, B., Sugita, S., Thompson, R., Viau, A., Williams, J., and Wu, H.: Pollen-based continental climate reconstructions at 6 and 21 ka: a global synthesis, Climate Dynamics, 37, 775–802, 2011.

Birks, H. J. B., Heiri, O., Seppä, H., and Bjune, A. E.: Strengths and Weaknesses of Quantitative Climate Reconstructions Based on Late-Quaternary Biological Proxies, The Open Ecology Journal, 3, 68–110, 2010.

Braconnot, P., Harrison, S. P., Otto-Bliesner, B., Abe-Ouchi, A., Jungclaus, J., and Peterschmitt, J.-Y.: The Paleoclimate Modeling Intercomparison Project contribution to CMIP5, CLIVAR Exchanges, 56/16/2, 15–19, 2011.

Bradley, R. S.: Paleoclimatology - Reconstructing Climates of the Quaternary, Academic Press, Oxford, 3 edn., 2015.

Brier, G.: Verification of Forecasts Expressed in Terms of Probability, Monthly Weather Review, 78, 1–3, 1950.

Brooks, S. P. and Gelman, A.: General Methods for Monitoring Convergence of Iterative Simulations, Journal of Computational and Graphical Statistics, 7/4, 434–455, 1998.

Carrassi, A., Bocquet, M., Bertino, L., and Evensen, G.: Data assimilation in the geosciences. An overview on methods, issues and perspectives, WIREs Climate Change, Available at: https://arxiv.org/abs/1709.02798, 2018.

Dee, S., Steiger, N. J., Hakim, G. J., and Emile-Geay, J.: On the utility of proxy system models for estimating climate states over the common era, Journal of Advances in Modeling Earth Systems, 8, 1164–1179, 2016.

Friedman, J., Hastie, T., and Tibshirani, R.: Sparse inverse covariance estimation with the graphical lasso, Biostatistics, 9/3, 432–441, 2008.

Gebhardt, C., Kühl, N., Hense, A., and Litt, T.: Multi-Scale Processes and the Reconstruction of Palaeoclimate, pp. 325–336, In: Neugebauer, Horst J. and Simmer, Clemens, Dynamics of Multiscale Earth Systems, Springer, Berlin, 2003.

Gebhardt, C., Kühl, N., Hense, A., and Litt, T.: Reconstruction of Quaternary temperature fields by dynamically consistens smoothing, Climate Dynamics, 30, 421–437, 2008.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D.: Bayesian data analysis, vol. 3, Chapman & Hall, CRC Press, Boca Raton, 2013.

Geyer, C.: Markov chain Monte Carlo maximum likelihood, pp. 156–163, In: Keramidas (ed.), Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, Interface Foundation, Fairfax Station, 1991.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102/477, 359–378, 2007.

Gomez-Navarro, J. J., Werner, J., Wagner, S., Luterbacher, J., and Zorita, E.: Establishing the skill of climate field reconstruction techniques for precipitation with pseudoproxy experiments, Climate Dynamics, 45, 1395–1413, 2015.

Gray, H., Leday, G. G., Vallejos, C. A., and Richardson, S.: Shrinkage estimation of large covariance matrices using multiple shrinkage targets, arXiv:1809.08024v1, pp. 1–32, 2018.

Hannart, A. and Naveau, P.: Estimating high dimensional covariance matrices: A new look at the Gaussian conjugate framework, Journal of Multivariate Analysis, 131, 149–162, 2014.

Harris, I. and Jones, P.: CRU TS4.01: Climatic Research Unit (CRU) Time-Series (TS) version 4.01 of high-resolution gridded data of month-by-month variation in climate (Jan. 1901- Dec. 2016), https://doi.org/doi:10.5285/58a8802721c94c66ae45c3baa4d814d0, 2017.

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

Haslett, J., Whiley, M., Bhattacharya, S., Salter-Townshend, M., Wilson, S. P., Allen, J., Huntley, B., and Mitchell, F.: Bayesian paleoclimate reconstruction, Journal of the Royal Statistical Society, Series A (Statistics in Society), 169, 395–438, 2006.

Hegerl, G. and Zwiers, F.: Use of models in detection and attribution of climate change, WIREs Clim Change, 2, 570–591, 2011.

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, Geoscientific Model Development, 10, 483–498, 2017.

Holmström, L., Ilvonen, L., Seppä, H., and Veski, S.: A Bayesian Spatiotemporal Model for Reconstructing Climate from Multiple Pollen Records, The Annals of Applied Statistics, 9/3, 1194–1225, 2015.

Huntley, B.: Reconstructing palaeoclimates from biological proxies: Some often overlooked sources of uncertainty, Quarternary Science Reviews, 31, 1–16, 2012.

Iversen, J.: Viscum, Hedera and Ilex as climate indicators, Geologiska Foereningens i Stockholm foerhandlingar, 66, 463–483, 1944.

Jones, P., New, M., Parker, D., Martin, S., and Rigor, I.: Surface air temperature and its variations over the last 150 years, Reviews of Geophysics, 37, 173–199, 1999.

Kageyama, M., Braconnot, P., Harrison, S. P., Haywood, A. M., Jungclaus, J., Otto-Bliesner, B. L., Peterschmitt, J.-Y., Abe-Ouchi, A., Albani, S., Bartlein, P. J., Brierley, C., Crucifix, M., Dolan, A., Fernandez-Donado, L., Fischer, H., Hopcroft, P. O., Ivanovic, R. F., Lambert, F., Lunt, D. J., Mahowald, N. M., Peltier, W. R., Phipps, S. J., Roche, D. M., Schmidt, G. A., Tarasov, L., Valdes, P. J., Zhang, Q., and Zhou, T.: The PMIP4 contribution to CMIP6 – Part1: Overview and over-arching analysis plan, Geoscientific Model Development, pp. 1033–1057, 2018.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G. A.: Challenges in combining projections from multiple climate models, Journal of Climate, 23, 2739–2758, 2010.

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science, 285, 1548–1550, 1999.

Kühl, N., Gebhardt, C., Litt, T., and Hense, A.: Probability Density Functions as Botanical-Climatological Transfer Functions for Climate Reconstruction, Quarternary Research, 58, 381–392, 2002.

Kühl, N., Litt, T., Schölzel, C., and Hense, A.: Eemian and Early Weichselian temperature and precipitation variability in northern Germany, Quarternary Science Reviews, 26, 3311–3317, 2007.

Li, B., Nychka, D. W., and Ammann, C. M.: The Value of Multiproxy Reconstruction of Past Climate, Journal of the American Statistical Association, 105/491, 883–895, 2010.

Lindgren, F., Rue, H., and Lindström, J.: An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach, Journal of the Royal Statistical Soviety, Statistical Methodology, Series B, 73/4, 423–498, 2011.

Liu, B., Ait-El-Fquih, B., and Hoteit, I.: Efficient Kernel-Based Ensemble Gaussian Mixture Filtering, Monthly Weather Review, 144, 781–800, 2016.

Liu, F., Bayarri, M., and Berger, J.: Modularization in Bayesian analysis, with emphasis on analysis of computer models, Bayesian Analysis, 4, 119–150, 2009.

MacKenzie, D. I., Nichols, J. D., Lachman, G. B., Droege, S., Royle, J. A., and Langtimm, C. A.: Estimating site occupancy rates when detection probabilities are less than one, Ecology, 83, 2248–2255, 2002.

Mahalanobis, P.: On the generalized distance in statistics, Proceedings of the National Institute of Sciences of India, 12, 49–55, 1936.

MARGO Project Members: Constraints on the magnitude and patterns of ocean cooling at the Last Glacial Maximum, Nature Geoscience, 2, 127–132, 2009.

Matheson, J. and Winkler, R.: Scoring Rules for Continuous Probability Distributions, Management Science, 22, 1087–1096, 1976.

Mauri, A., Davis, B., Collins, P., and Kaplan, J.: The influence of atmospheric circulation on the mid-Holocene climate of Europe: a data-model comparison, Climate of the Past, 10, 1925–1938, 2014.

Mauri, A., Davis, B., Collins, P., and Kaplan, J.: The climate of Europe during the Holocene: a gridded pollen-based reconstruction and its multi-proxy evaluation, Quaternary Science Reviews, 112, 109–127, 2015.

Ohlwein, C. and Wahl, E. R.: Review of probabilistic pollen-climate transfer methods, Quaternary Science Reviews, 31, 17–29, 2012.

Parnell, A. C., Sweeney, J., Doan, T. K., Salter-Townshend, M., Allen, J. R., Huntley, B., and Haslett, J.: Bayesian inference for palaeoclimate with time uncertainty and stochastic volatility, Journal of the Royal Statistical Society, Applied Statistics, Series C, 64/1, 115–138, 2015.

Parnell, A. C., Haslett, J., Sweeney, J., Doan, T. K., Allen, J. R., and Huntley, B.: Joint Palaeoclimate reconstruction from pollen data via forward models and climate histories, Quarternary Science Reviews, 151, 111–126, 2016.

Plummer, M., Best, N., Cowles, K., and Vines, K.: CODA: Convergence Diagnosis and Output Analysis for MCMC, R News, 6, 7–11, https://www.r-project.org/doc/Rnews/Rnews_2006-1.pdf, 2006.

Polson, N. G., Scott, J. G., and Windle, J.: Bayesian inference for logistic models using Polya-Gamma latent variables, arXiv:1205.0310v3, pp. 1–42, 2013.

Rehfeld, K., Trachsel, M., Telford, R., and Laepple, T.: Assessing performance and seasonal bias of pollen-based climate reconstructions in a perfect model world, Climate of the Past Discussions, 13, 1–22, 2016.

Rue, H. and Held, L.: Gaussian Markov random fields : theory and applications, Chapman & Hall / CRC (Taylor & Francis Group), Boca Raton, 2005.

Rypdal, K., Rypdal, M., and Fredriksen, H.-B.: Spatiotemporal Long-Range Persistence in Earth's Temperature Field: Analysis of Stochastic–Diffusive Energy Balance Models, Journal of Climate, 28, 8379–8395, 2015.

Schölzel, C. and Hense, A.: Probabilistic assessment of regional climate change in Southwest Germany by ensemble dressing, Climate Dynamics, 36, 2003–2014, 2011.

Schölzel, C., Hense, A., Hübl, P., Kühl, N., and Litt, T.: Digitization and geo-referencing of botanical distribution maps, Journal of Biogeography, 29, 851–856, 2002.

Silverman, B.: Density Estimation for Statistics and Data Analysis, vol. 26 of *Monographs on Statistics and Applied Probability*, Chapman & Hall / CRC, Boca Raton, 1986.

Simonis, D.: Reconstruction of possible realizations of the Late Glacial and Holocene near surface climate in Central Europe, Dissertation, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2009.

Simonis, D., Hense, A., and Litt, T.: Reconstruction of late Glacial and Early Holocene near surface temperature anomalies in Europe and their statistical interpretation, Quaternary International, 274, 233–250, 2012.

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., and Roe, G. H.: Assimilation of Time-Averaged Pseudoproxies for Climate Reconstruction, Journal of Climate, 27, 426–441, 2014.

Stolzenberger, S.: Untersuchungen zu botanischen Paläoklimatransferfunktionen, Diploma thesis, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2011.

Stolzenberger, S.: On the probabilistic evaluation of decadal and paleoclimate model predictions, Dissertation, Meteorologisches Institut der Rheinischen Friedrich-Wilhelms-Universität Bonn, 2017.

Tawn, N. G. and Roberts, G. O.: Accelerating Parallel Tempering: Quantile Tempering Algorithm (QuanTA), arXiv:1808.10415v1, pp. 1–39, 2018.

Thuiller, W.: BIOMOD – optimizing predictions of species distributions and projecting potential future shifts under global change, Global Change Biology, 9, 1353–1362, 2003.

Tingley, M. P. and Huybers, P.: A Bayesian Algorithm for Reconstructing Climate Anomalies in Space and Time. Part I: Development and Applications to Paleoclimate Reconstruction Problems, Journal of Climate, 23, 2759–2781, 2010.

Tingley, M. P., Craigmile, P. F., Haran, M., Li, B., Mannshardt, E., and Rajaratnam, B.: Piecing together the past: statistical insights into paleoclimatic reconstructions, Quaternary Science Reviews, 35, 1–22, 2012.

Werner, J. P. and Tingley, M. P.: Technical Note: Probabilistically constraining proxy age-depth models within a Bayesian hierarchical reconstruction model, Climate of the Past, 11, 533–545, 2015.

Windle, J., Polson, N. G., and Scott, J. G.: BayesLogit: Bayesian logistic regression, URL http://cran.r-project.org/web/packages/BayesLogit/index.html, 2013.

Windle, J., Polson, N. G., and Scott, J. G.: Sampling Pólya-Gamma random variates: alternate and approximate techniques, arXiv:1405.0506v1, pp. 1–26, 2014.

Yang, Z. and Zhu, T.: The good, the bad, and the ugly: Bayesian model selection produces spurious posterior probabilities for phylogenetic trees, arXiv:1810.05398v1, pp. 1–6, 2018.

Zou, H., Hastie, T., and Tibshirani, R.: Sparse Principal Component Analysis, Journal of Computational and Graphical Statistics, 15/2, 265–286, 2006.

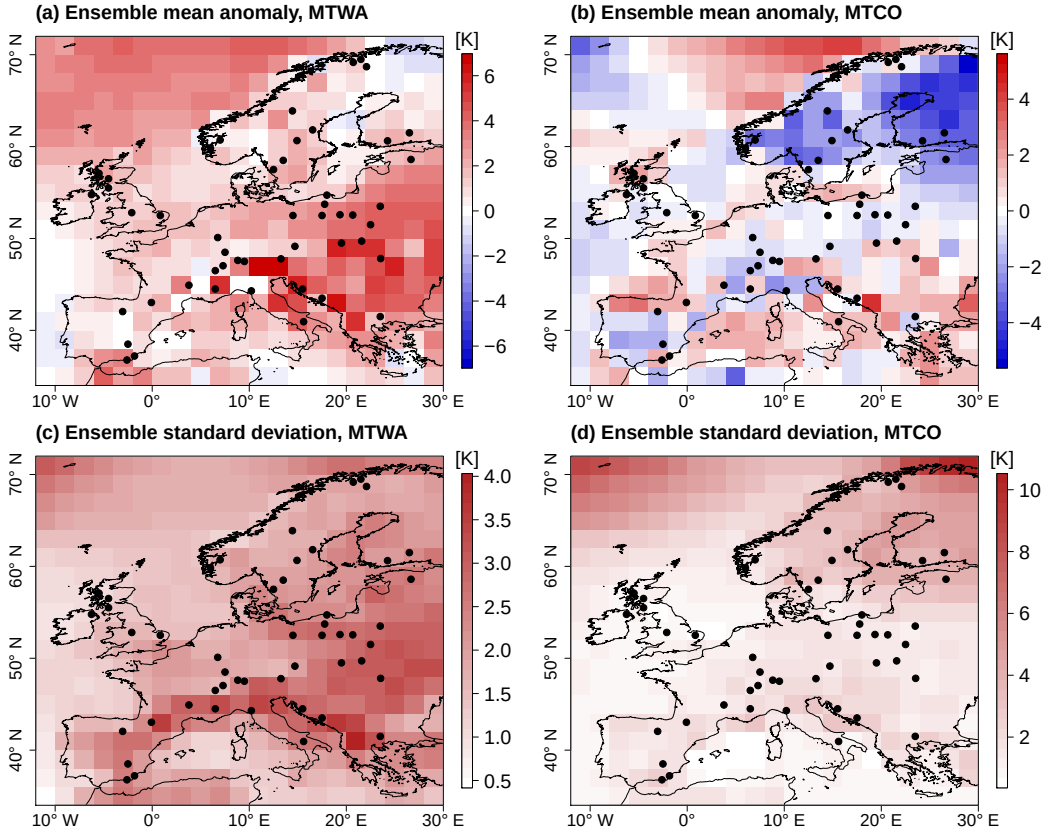

Figure 1. PMIP3 MH ensemble mean anomaly from CRU reference climatology, (a) MTWA, (b) MTCO, and ensemble spread as ~~size of point-wise 90% CIs~~ empirical standard deviation of the ~~prior mixture distribution~~ensemble members, (c) MTWA, (d) MTCO. Black dots depict proxy samples from Simonis et al. (2012).~~In the top row, point-wise significant anomalies (5% level) are marked by black crosses.~~

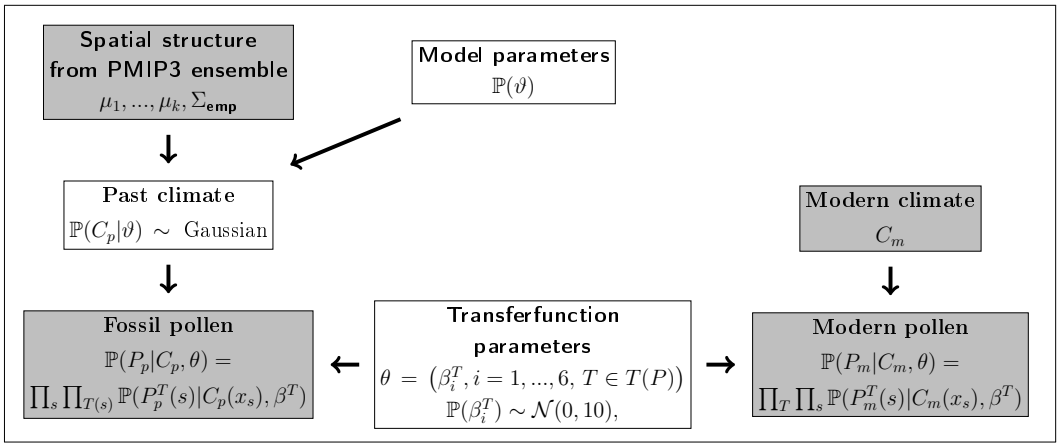

Figure 2. Directed acyclic graph corresponding to the Bayesian framework Eq. (??2). Involved quantities are given by nodes and arrows indicate dependences of variables. Details of the formulas are explained in Sect. ??3.2 and ??3.3. White: inferred quantities; gray: input data.

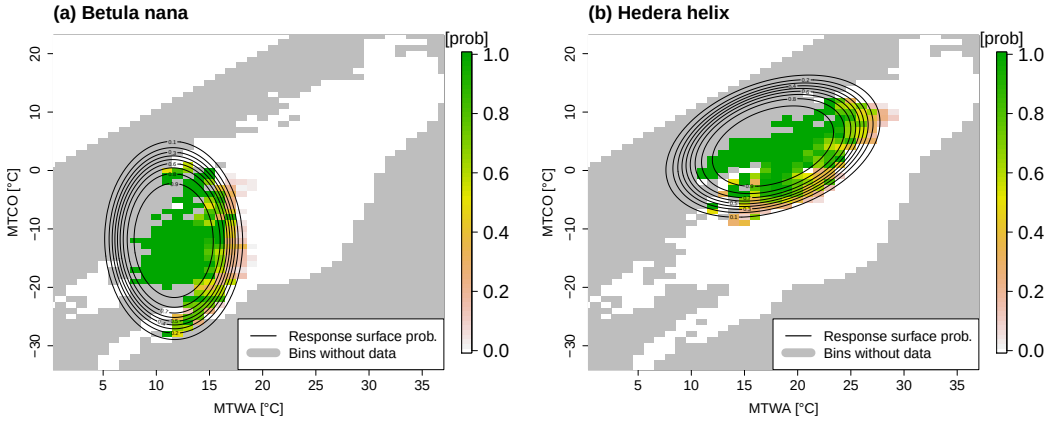

Figure 3. Response functions for Betula nana (a) and Hedera helix (b). The ~~probability of presence~~ relative frequency of ~~the taxa~~ occurrence in 1K bins is shown in colours, ~~black dots denote presence of the taxa,~~ and ~~gray dots denote absence of~~ the ~~taxa in~~ contours depict the probability of presence as estimated by the logistic response function. Gray boxes denote bins without calibration ~~dataset~~data. In the climate space, combinations of MTWA and MTCO ~~above the gray line at~~ with MTWA ~~=~~ < MTCO cannot occur by definition. ~~Gray dots above this line are~~ White bins in the upper left depict artificial absence information added to account for this constraint.

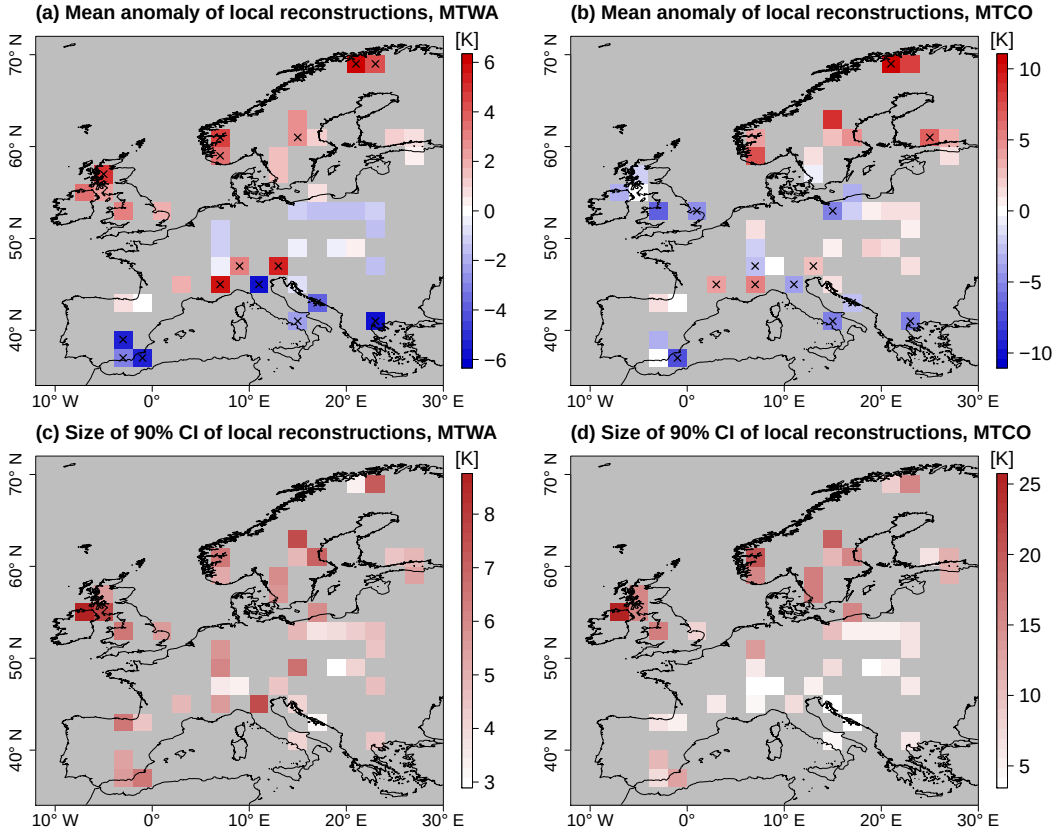

Figure 4. Summary statistics of local reconstructions using the PITM forward model. Top row: mean anomaly from CRU reference climatology (left: MTWA, right: MTCO), bottom row: uncertainty measured by the size of marginal 90% CIs (left: MTWA, right: MTCO). In the top row, significant anomalies (5% level) are marked by black crosses.

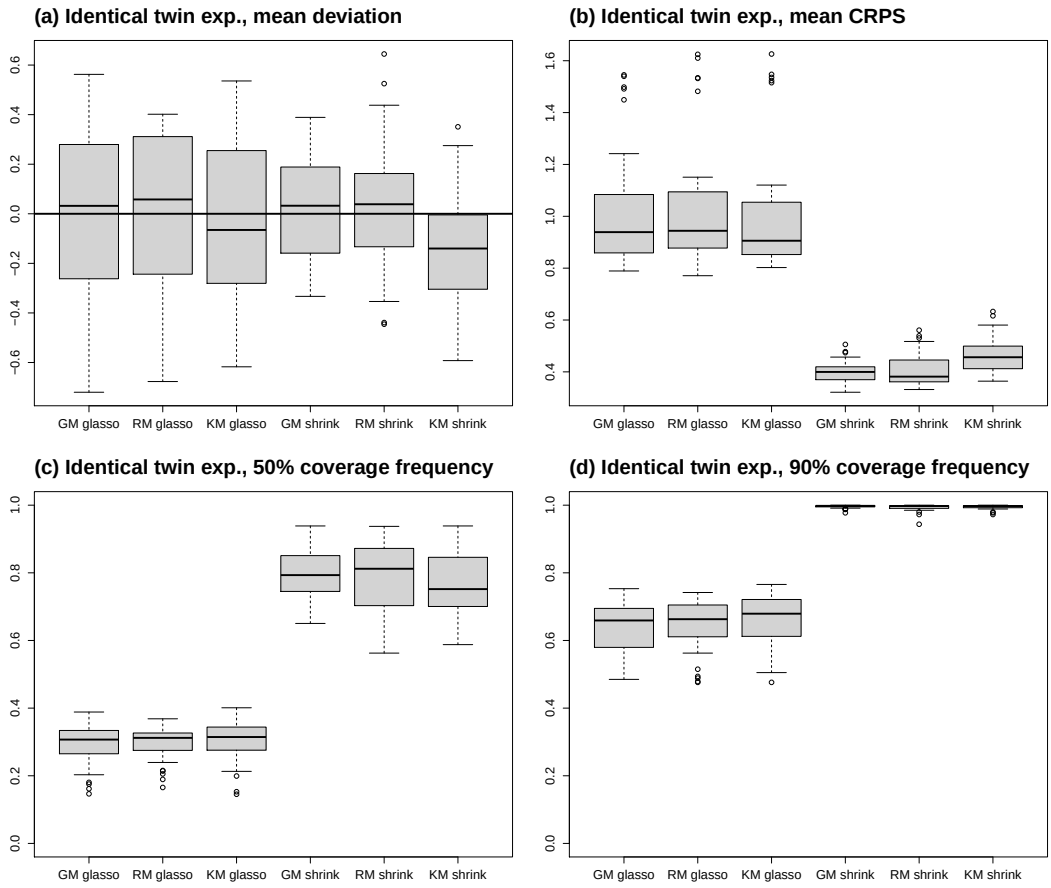

Figure 5. ~~Spatial reconstruction for MH~~Results from ITEs. ~~Top row: Posterior mean anomaly~~ The boxplots depict the distribution of experiments with the same process stage model. (a) Mean deviation from ~~CRU~~ reference ~~climatology~~ climate, (~~left: MTWA~~b) mean CRPS, ~~right: MTCO~~(c) coverage frequency of 50% CIs, ~~middle row: reconstruction uncertainty plotted as size~~ (d) coverage frequency of ~~point-wise~~ 90% CIs~~(left~~.

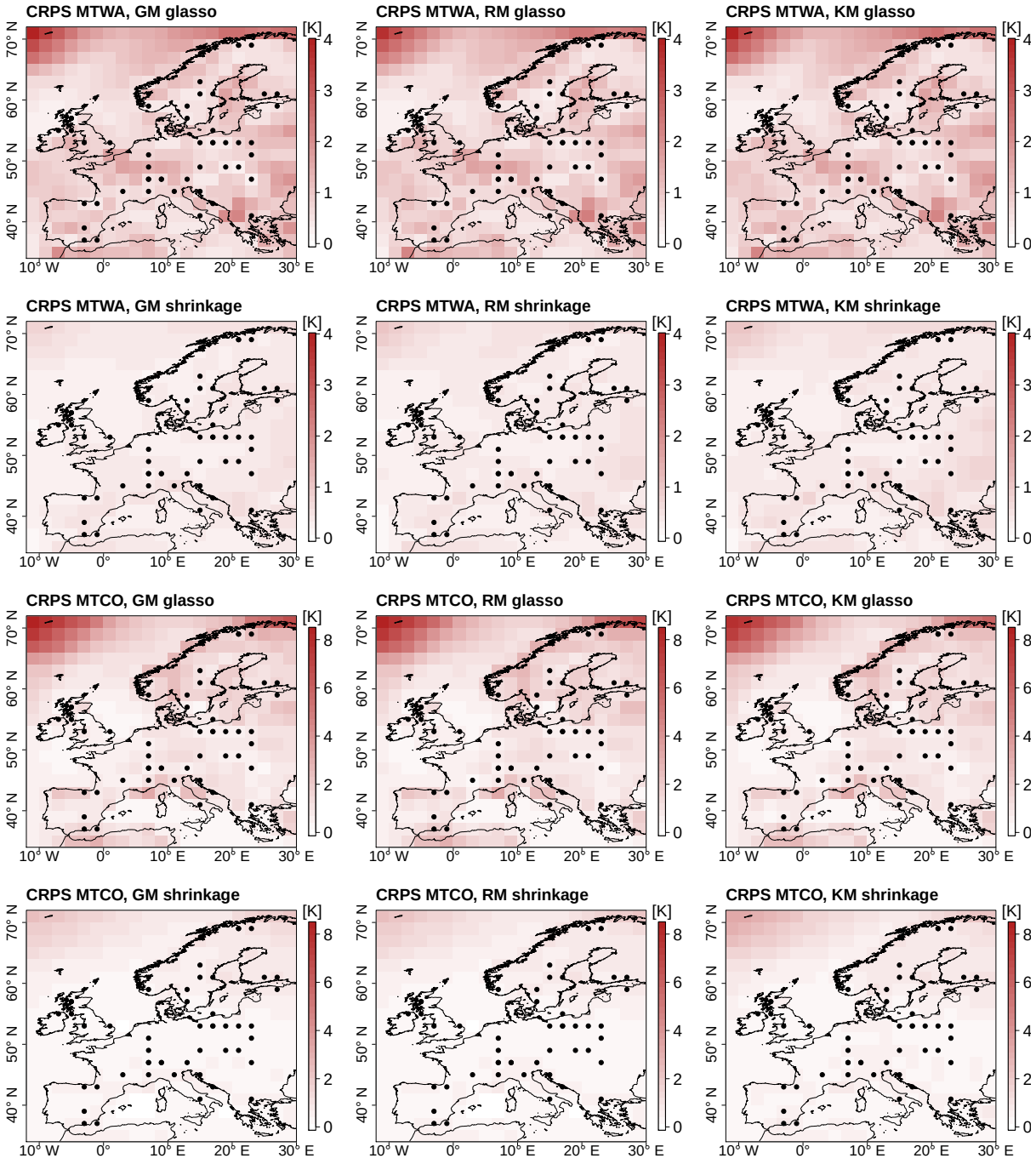

Figure 6. Mean CRPS in ITEs for GM, RM, and KM. Top row: Models with glasso covariance matrix, MTWA. Second row: Models with shrinkage covariance matrix, ~~right~~MTWA. Third row: Models with glasso covariance matrix, MTCO~~)~~, ~~bottom~~ Bottom row: ~~reduction of uncertainty from posterior to prior measured by ratio of posterior to prior point-wise 90% CI sizes (left: MTWA~~Models with shrinkage covariance matrix, ~~right:~~ MTCO~~)~~. ~~Black dots depict~~ Grid boxes with simulated proxy ~~samples~~data are depicted by black dots.~~In~~

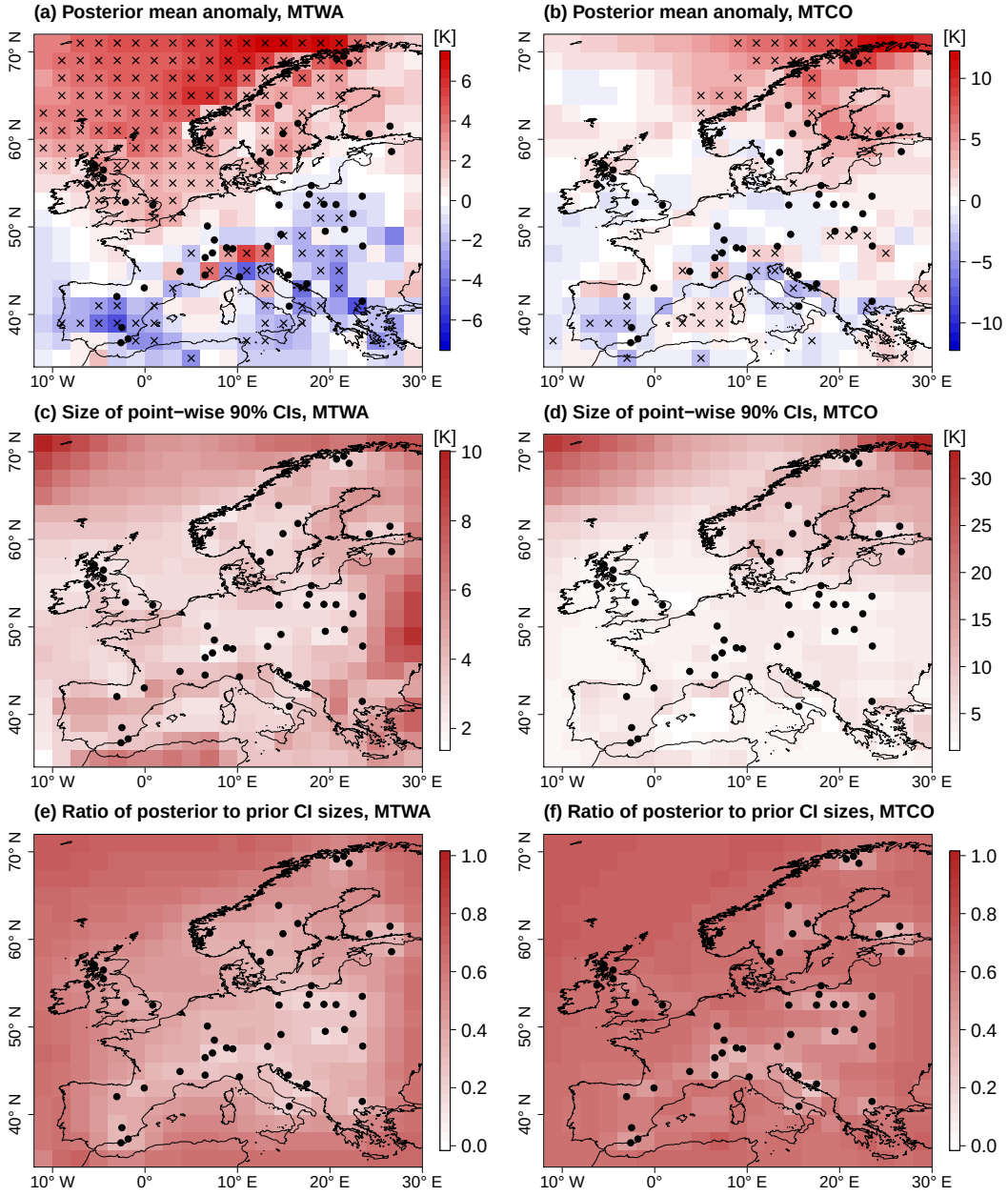

Figure 7. Spatial reconstruction for MH. Top row: Posterior mean anomaly from CRU reference climatology (left: MTWA, right: MTCO), middle row: reconstruction uncertainty plotted as size of point-wise 90% CIs (left: MTWA, right: MTCO), bottom row: reduction of uncertainty from posterior to prior measured by ratio of posterior to prior point-wise 90% CI sizes (left: MTWA, right: MTCO). Black dots depict proxy samples. In the top row, point-wise significant anomalies (5% level) are marked by black crosses.

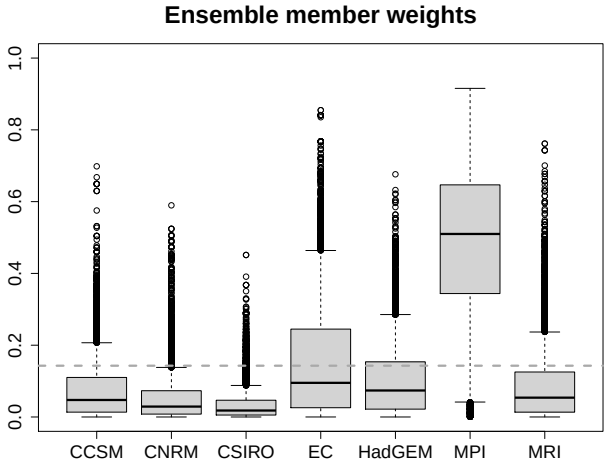

Figure 8. Posterior ensemble member weights (λ) of the ~~top row, point-wise significant anomalies~~ reconstruction. Prior weights (~~5% level~~mean of λ) are ~~marked~~ denoted by ~~black crosses~~the dashed line.

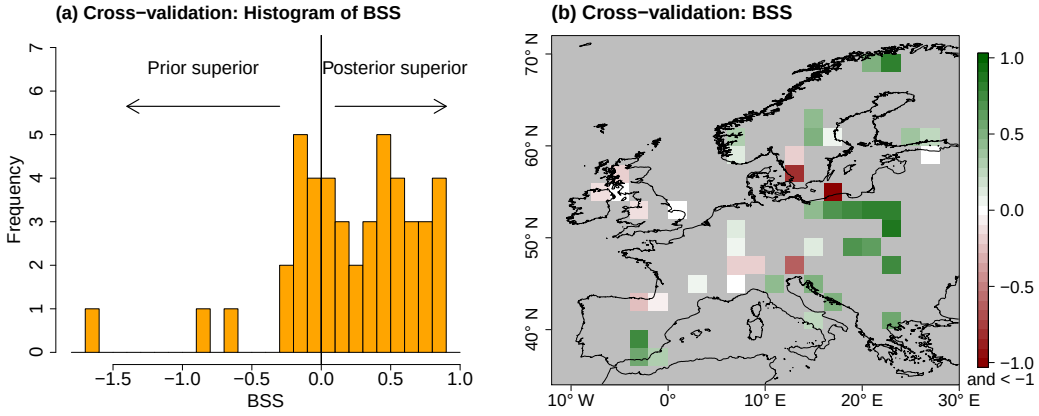

Figure 9. BSS from leave-one-out cross-validation. For positive values the posterior is superior to the prior distribution from the unconstrained PMIP3 ensemble, while for negative values the posterior is inferior to the prior. (a) histogram, (b) spatial distribution.

Posterior ensemble member weights of the reconstructions. The posterior of ω in the joint reconstruction is shown as boxplot. The diamonds represent the mean of z in the joint, the MTWA only, and the MTCO only reconstruction. Prior weights (mean of z) are denoted by the dashed line.

Sensitivity of reconstructions to changes in the glasso penalty parameter ρ. Portion of the penalty term on the total cost in Eq. (??), mean BS of cross-validations, spatially averaged posterior mean anomaly from CRU reference climatology, and spatially averaged size of point-wise 90% CIs. All quantities are plotted as percentage of the respective values for $\rho = 0.3$. Note the non-linear scaling of the x-axis.

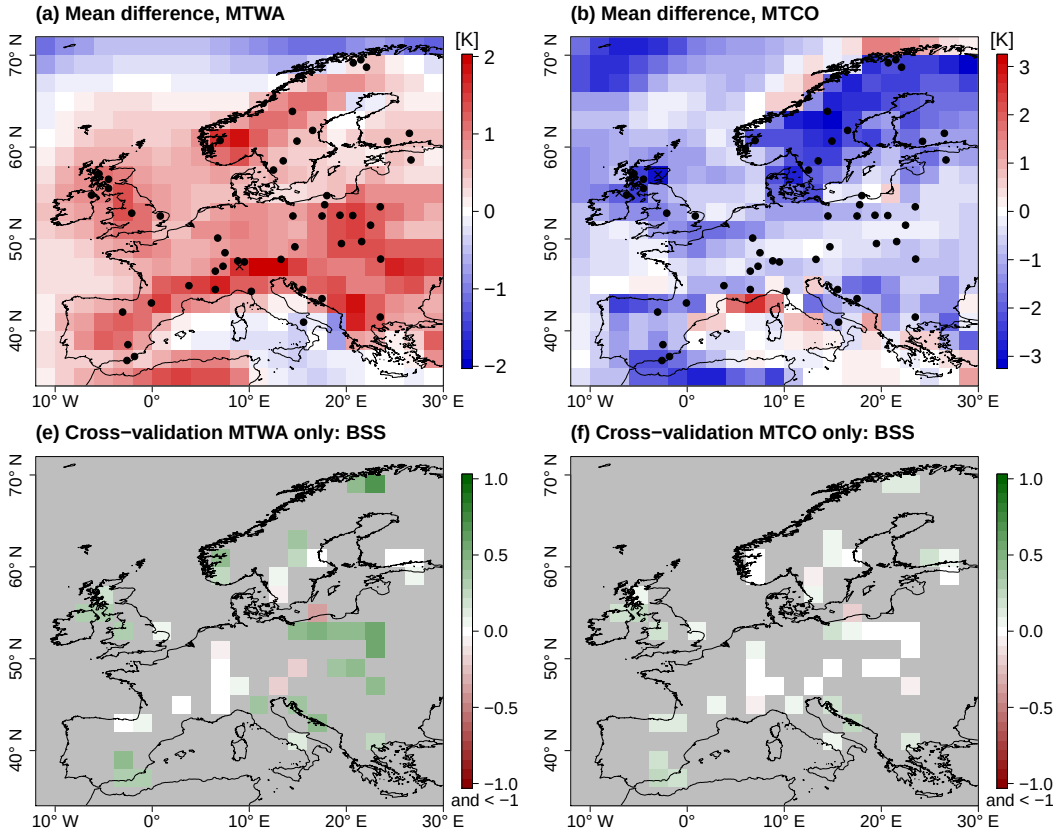

Figure 10. Differences of joint and separate reconstructions of MTWA and MTCO. Top row: posterior mean difference (left: MTWA, right: MTCO); ~~middle row: difference of size of point-wise 90% CIs (left: MTWA, right: MTCO);~~ bottom row: BSS of the separate reconstructions (left: MTWA, right: MTCO). Black dots depict proxy samples. In the top row, point-wise significant differences (5% level) between the separate and the joint reconstructions are marked by black crosses.

Table 1. Basic information on the PMIP3 climate simulations used to construct the ~~prior~~ process stage in the Bayesian framework (from https://pmip3.lsce.ipsl.fr)

~~Model~~ Model	~~Institute~~ Institute ~~Model yearsMH~~	~~Atmospheric grid~~ Atmospheric grid	~~Ocean grid~~ Simulated years
CCSM4	NCAR	288x192xL26	~~320x384xL60~~ 301
CNRM-CM5	CNRM-CERFACS	256x128xL31	~~362x292xL42~~ 200
CSIRO-Mk3-6-0	CSIRO-QCCCE	192x96xL18	~~192x195xL31~~ 100
EC-Earth-2-2	ICHEC	320x160xL62	~~362x292xL42~~ 40
HadGEM2-CC	MOHC	192x144xL60	~~360x216xL40~~ 35
MPI-ESM-P	MPI-M	196x98xL47	~~256x220xL40~~ 100
MRI-CGCM3	MRI	320x160xL48	~~364x368xL51~~ 100

Table 2. Summary measures for ITEs and CVEs with the six models described in Sect. 3.3: Spatially averaged mean deviation from reference climatology, spatially averaged size of point-wise 50% CIs, spatially averaged size of point-wise 90% CIs, coverage frequency of 50% CIs, coverage frequency of 90% CIs, spatially averaged CRPS, mean BS.

	GM glasso	RM glasso	KM glasso	GM shrink	RM shrink	KM shrink
Mean deviation [K]	0.004	0.000	-0.032	0.026	0.028	-0.140
50% CI size [K]	1.189	1.194	1.238	1.583	1.592	1.755
90% CI size [K]	2.906	2.916	3.032	3.880	3.901	4.356
50% coverage frequency [%]	29.2	29.6	30.5	79.9	78.8	76.5
90% coverage frequency [%]	64.1	64.4	66.1	99.6	99.3	99.5
CRPS [K]	1.03	1.032	1.010	0.399	0.408	0.468
BS [p^2]	0.186	0.186	0.187	0.165	0.163	0.161

Table 3. Summary measures for the joint MTWA and MTCO reconstructions (rows 1 and 2) and the separated reconstructions of MTWA (row 3) and MTCO (row 4). Numbers in brackets are minima and maxima of the corresponding 90% CIs. Additional explanations for all the columns can be found in Sect. ~~??~~ 4.1.1 to ~~??~~4.1.1.

Reconstruction name	Spatial mean anomaly	Spatially averaged 90% CI size	Point-wise uncertainty reduction	Spatial homogeneity	Median BSS	~~PMIP~~ PMIP3 model with highest weight
Joint reconstruction (MTWA)	(~~0.33 K~~0.03 K) ~~0.01 K~~ 0.51 K (~~-0.32 K~~0.99 K)	~~2.90 K~~ 4.15 K	~~64.5~~38.1%	(~~1.40 K~~1.31 K) ~~1.48 K~~ 1.41 K (~~1.57 K~~1.53 K)	0.28	MPI-ESM-P
Joint reconstruction (MTCO)	(~~1.50 K~~0.10 K) ~~1.11 K~~ 0.69 K (~~0.74 K~~1.32 K)	~~3.18 K~~ 5.84 K	~~58.6~~19.6%	(~~2.09 K~~2.33 K) ~~2.18 K~~ 2.54 K (~~2.27 K~~2.76 K)		
Separate MTWA reconstruction	(~~0.96 K~~0.55 K) ~~0.64 K~~ 1.04 K (~~0.30 K~~1.51 K)	~~3.10 K~~ 4.14 K	~~62.2~~38.1%	(~~1.38 K~~1.29 K) ~~1.47 K~~ 1.41 K (~~1.55 K~~1.51 K)	~~0.16~~ 0.27	MPI-ESM-P
Separate MTCO reconstruction	(~~0.92 K~~ 0.82 K) ~~0.06 K~~ 0.14 K (~~-0.79 K~~0.60 K)	~~4.14 K~~ 5.72 K	~~49.4~~20.6%	(2.45 ~~K~~K) ~~2.66 K~~ 2.69 K (~~2.86 K~~2.92 K)	~~0.04~~ 0.05	HadGem2-CC

~~Pseudo-code for MC^3 algorithm Initialize A MCMC chains $j = 1,...,Jm = 1,...,MT \in T(P)l = 1,...,L(T)$Sample from full conditional of γ_l^T Sample from full condition of β^T Sample from full conditional of ω Sample from full conditional of z $x \in x_P$Sample random walk proposal for $C_p(x)$ Calculate Metropolis-Hastings ratio r Accept proposal for $C_p(x)$ with probability $p = \max(1,r)$ $a = 1,...,(A-1)$Calculate Metropolis-Hastings ratio r of chains a and $a+1$ Swap chains a and $a+1$ with probability $p = \max(1,r)$ $i = 1,...,JM$Sample from full conditional of $C_p(x_Q)$~~