Clim. Past Discuss.,
https://doi.org/10.5194/cp-2018-87-AC1, 2018
© Author(s) 2018. This work is distributed under
the Creative Commons Attribution 4.0 License.

Climate of the Past Discussions
Open Access
EGU

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,

Printer-friendly version

Discussion paper

CC BY

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-

CC BY

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-

CC BY

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

CC BY

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

CC BY

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

CC BY

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

Printer-friendly version

Discussion paper

CC BY

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

CC BY

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.

CC BY

"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

CC BY

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

CC BY

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

CC BY

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.

CC BY

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."

Printer-friendly version

Discussion paper

CC BY

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.

CC BY

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

Printer-friendly version

Discussion paper

CC BY

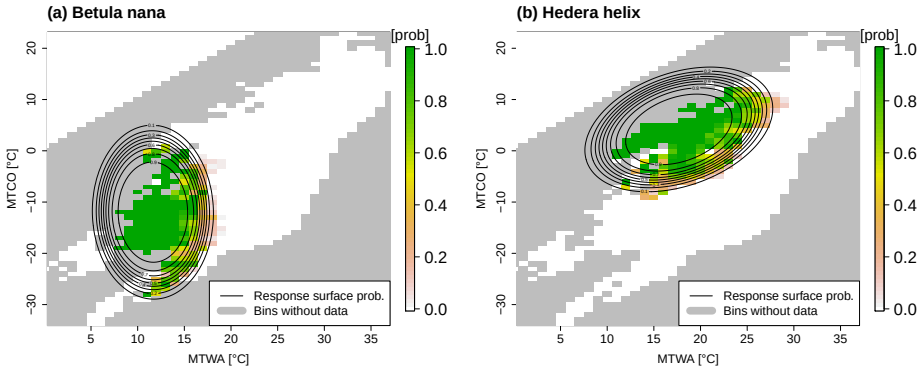

Figure 1: Response surfaces for Betula nana (a) and Hedera helix (b). The ratio of presence versus absence in the modern calibration data in each bin with at least one data point is shown in colours. Gray bins are bins without data. The response surfaces (probability of presence according to Eq. (7) 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.

Printer-friendly version

Discussion paper

CC BY