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.

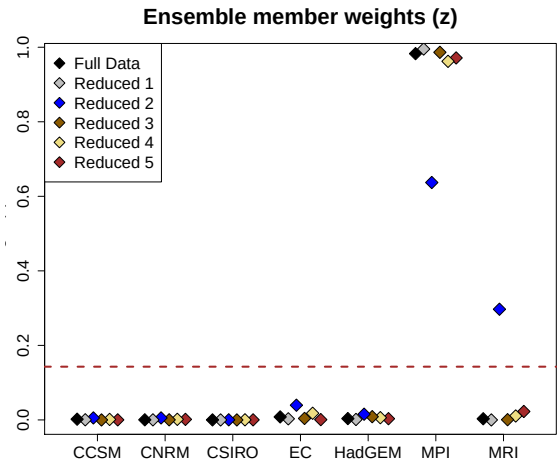

Figure 1: Posterior ensemble member weights (mean of z) of the the joint reconstruction and of five experiments where half of the proxy samples are left out. The black diamonds correspond to the reconstruction with the full dataset, whereas the coloured diamonds represent the experiments with reduced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

Fig. 1.