## BUMPER, Holden et al, GMDD
## Response to C. ter Braak (Referee)

**Referee comments black.**
**Author responses red.**
**Manuscript changes green.**

We are grateful for this careful review, which has certainly improved the mathematical rigour of the model description.

General comments

The first aim of this paper is to alleviate a difficult problem in every Bayesian analysis: that of setting priors. In this case for the Bayesian approach to palaeo environment reconstruction using Gaussian response curves. This approach was first proposed in Holden et al. 2008. The priors are partly based on the data to be analyzed (as in Empirical Bayes), so that the approach is not fully Bayesian in the strict sense; but it practical and appealing.

The second aim is to provide a more general evaluation of this approach, with the new priors, to simulated data sets and a number of (famous and newer) data sets. The approach is compared with the simple approach based on weighted averaging with deshrinking using inverse regression, aka WAPLS1.

The computation approach avoids the usual MCMC computation, or approximations thereof (e.g. INLA), in Bayesian analysis by limiting the approach to one-dimensional modelling and reconstruction and by discretising the parameters, so that in fact 2,560 possible parameter combinations remain. Thereby a fully Bayesian analysis is possible without MCMC. The posteriors for these models act as if they are weights in a model averaging exercise. It is well known that simple models when averaged can solve complex problems. In my view, the paper fits in the journal, has a clear aim and fulfils the claims.

Specific comments

In one place, it looks like the distinction between prior and posterior is lost in the notation/formulas. In section 2.2 prob(SRC_jk) is surely the posterior denoted by prob(SRC_jk|Y,X), where Y and X are the training data (as in eq (1)). Note that the model setup also belongs to the condition. In eq( 2), the posterior weights are meant, is it not? Instead of adapting all formulas (when I am right in this) state explicitly that "From now on prob(SRC_jk) is the posterior probability, the probability of the SRC given the training data Y and X.

Yes, we agree and have made this change (correcting Eq. 2, and after then adding the suggested sentence). The mathematics should also be simpler to follow now we have changed Eq. 1 to reflect the application of the entire training set (see response to the other referee, Andrew Parnell) . Eq 1 and 2 now read

$$prob(SRC_{jk}|Y_k, X) \propto prob(SRC_{jk}) \times \prod_i prob(y_{ik}|SRC_{jk}, x_i) \qquad \text{Eq. 1}$$

$\sum_j prob(SRC_{jk}|Y_k, X) = 1$ Eq. 2

The probability distributions in section 2.3 form a hurdle model (zero inflated distribution wit truncation at 0 of the count distribution. If I am right in this, please mention this.

Thank you, we have made this clarification.

"Species counts distributions are represented with a hurdle model (a zero inflated distribution with truncation at zero of the distribution of percentage counts)."

I numbered the pages from 1-20.

P2L4: in the model? It depends of course what you mean with model here. But in the natural sense, the model is fixed and only the parameters of the model are uncertain, and for discrete parameters their distribution (weights). So, say so. Even although the approach has aspects of model averaging, it is best viewed as defining one model..., which is then fixed.

Agreed. We have changed the text ("uncertainty in the data and in the model parameters") to clarify that we are referring only to the parametric uncertainty of the model here (not the structural uncertainty arising from choice of mathematical structure). Additionally, we have changed the description of the parameter selection process from "model selection" to "parameter estimation".

P4L14 (end section 2.2) At how many points is x evaluated?

New text added

"The likelihood functions are evaluated at 100 evenly spaced points across a range that comfortably spans the training set environmental range (see section 2.4), and are normalised to 1."

and

"The reconstruction is evaluated at the same 100 points as the likelihood function, and the probabilities normalised to 1."

and (in Section 2.4)

"We note that the indicative tolerance is also used to define the range of environment considered in the reconstruction (Section 2.2), from $(x_{min} - 6t')\ to\ (x_{max} + 6t')$. Significant probabilities beyond this range are unlikely given the constraints imposed upon the optima and tolerances. In any event, as with any transfer function, the model should not be applied under suspected extrapolation beyond the training set environment."

P4L19 The expected abundance follows a distribution??? This is not a distribution in any of the senses you use in this paper; it is a Gaussian response curve model; see ter Braak & Barendregt 1986 http://dx.doi.org/10.1016/0025 5564(86)90031-3 when it has aspects of a distribution.

Change made and citation added

"can be fitted by a Gaussian response curve (ter Braak and Barendregt, 1986)"

P4L39-41. Eq (11) Note it relation to the exponential distribution and geometric distribution. https://en.wikipedia.org/wiki/Exponential_distribution. Probably you treat is as a discrete distribution and truncate is at 0 (or y<1). On P5L1 we learn that you made an assumption on the data y: between 0 and 100. Please be more explicit and/or give a more general denominator in (11). Such things can lead to strange errors later on, when used without further scrutiny.

Clarified

"is expressed as a continuous distribution, truncated at 0 and 100:"

P13L40 To make the program even more user-friendly a wrapper in R and/or Python appears much wanted. Make it a priority.

Thank you for this suggestion, we will look into this.

Technical corrections

P2L21. The two -> Two (they were not mentioned before)

Change made