We thank the two reviewers for their positive and constructive comments on our manuscript.

We hereafter address the comments made by the reviewers. Our replies are given in italic.

Reviewer 1:

1. I was struck by the estimates of h for the real-world reported in Table 1 – they are larger than I expected. I've always been aware of spatial autocorrelation as an issue with calibration datasets, but I would have thought that the spatial autocorrelation is on the order of 10's of kilometers rather than 100's of kilometers. The results here (300 to 750km) implies a fairly hefty discarding of data – a 750 km radius around a point will remove a lot of data! Suggest adding a short paragraph to the discussion that notes these points and maybe speculates about the ecological or environmental processes that are creating such a large-scale spatial autocorrelation.

The values of h found in this study are indeed large. Different variables have different ranges of spatial autocorrelation. Mean annual air temperature, mean annual sea surface temperature or salinity have large spatial autocorrelation at the scale of 100's of kilometres. Hence values of h in the order of 100's of kilometres are not that surprising. However, as shown in the manuscript: only spatial autocorrelation in nuisance variables is unduly improving performance. Not the spatial autocorrelation in the variable of interest (samples are more similar than exclusively caused by the variable of interest).
The goal of h-block cross-validation is to have a validation set that is independent from the calibration dataset. We therefore have to remove spatial autocorrelation caused by influence of spatially autocorrelated nuisance variables.
We will add a short paragraph following the reviewer's suggestion.

2. The abstract and conclusions both emphasize the point that the three methods return the same value of h, but on p 4735, there seems to be a certain amount of fudging going on to ensure that the variance explained approach is returning a value of h that isn't 'excessively large.' A suspicious reader might wonder whether this approach had been tuned to meet the expectations set by the other approaches, and whether this tuning would be robust for other datasets. Maybe a sentence or two addressing this point, in Methods or Discussion, would help.

The sum of squares as a function of h show roughly an L shape. First a rapid decrease and then a part with low changes in sum of squares. We need a way of distinguishing between an important decrease (vertical part of the L) and an unimportant decrease (horizontal part of the L) in sum of squares. To find the divide between an important and an unimportant decrease in sum of squares we needed to define an indeed arbitrary rule. This rule, however, worked for the simulations as well as the three real-world datasets. We therefore think that this rule should be robust for other datasets.
Similar rules are used when pruning classification and regression trees or choosing the number of PLS or WA-PLS components to use for a transfer function .
We will add a few sentences addressing this point to the revised manuscript.

3. Suggest adding a conceptual or demonstration figure illustrating the three methods summarized on p4732.

4. These recommendations are all for cross-validation tests. Many paleoclimatologists, of course, use transfer functions to make down-core reconstructions of past climatic variables. When going downcore,

should paleoclimatologists still apply h-block winnowing, or is this only necessary for cross-validation?

We mainly suggest using h-block cv under cross-validation. One application of h-block cross-validation is to find out if it is to reconstruct a variable at all. The relation between modern RMSEP and downcore RMSE is largely unknown.

5. 'Spatially independent' – suggest defining this concept explicitly in the ms.

We will add a definition of 'spatially independent' to the manuscript.
If we have two points in space and measure a (climate) variable at these two points, it is not possible to predict the variable at one point using the variable at the other point and vice versa.

We will also incorporate all the line by line comments in a revised version of the manuscript.
We also reply to the most important comments directly.

P4731 L1-2: "most palaeolimnological transfer functions have little spatial structure in the calibration set, and thus are not affected by this problem (Telford and Birks, 2009)." Either modify this statement to make it less sweeping or add references to better support it. T&B2009 only showed that there wasn't much spatial autocorrelation in a single paleolimnological variable (pH) for a single region (NE US).

We will modify this statement following the reviewer' s suggestions.

L20-22: A sum of squares less than 2 is being established as a criteria the data been standardized at this point? Or, if not, does this create the problem that different variables and different units would imply different scalings here?

With the variance-explained method we are comparing squared correlations (r2, bounded between 0 and 1) Pearson' s product-moment correlation coefficient is a standardised covariance and is therefore independent of the units of the original variables.

4739 L3-8: This section is generally correct but is blurring a bit the distinction between taxonomic similarity and environmental similarity; specifically it implies that MAT choices are based on environmental similarity. MAT of course is based on taxonomic similarity, so environmental similarity matters only insofar as it determines taxonomic similarity.

We will edit the text slightly to make this clearer.
Indeed, MAT selects taxonomically similar samples based on an appropriate distance metric between species assemblages. This distance metric is a holistic measure of the similarity of all environmental variables affecting the species assemblage (Telford and Birks, 2005),-not only the taxonomic similarity caused by the variable of interest.