**Climate
of the Past
Discussions**

Interactive comment on "Clustering climate reconstructions" by G. Bürger

G. Bürger

gbuerger@uvic.ca

I would like to thank the reviewer for reading the paper and raising a couple of important points.

1) My major concern is about the significance levels. Not much is mentioned about this in the paper, only in the caption to Table 2 is there a brief comment (0.47 as the 99% level). It seems that the results of the paper depend very much on this value and it is therefore important that the method to calculate it is described in detail. One important parameter is the number of degrees of freedom used in the calculation of the threshold value (0.47). There is a lot of serial correlations in the reconstructions so the number of degrees of freedom is smaller than the number of years. How this is dealt with should be described in details in the paper. Also, I guess that only the reconstructions themselves and not their uncertainty bands are used when the significance

level is calculated. But the reconstructions would probably appear more coherent if the uncertainty bands are included.

Response: For a fairly large class of processes (e.g. linear processes) the significance levels for coherence are largely independent of frequency and solely depend on the amount of smoothing used for estimating the spectra (size of frequency window), as described e.g. in Brockwell and Davis, p451 (it all comes down to the periodogram estimates for different frequencies being independent). They do not depend on the memory character of the series, as demonstrated in the attached Fig.1 using white and red (d=0.8) noise: For a significance level of 90%, 95%, and 99%, and the red and blue curves represent the corresponding quantiles of 1000 random samples of corresponding noise series (straight lines are the estimates themselves).

With regard to the uncertainty bands, I am unaware of techniques to incorporate those into the coherence spectrum. But how can two series that are not coherent alone become so, in a significant way, when uncertainty is added? That is counterintuitive to me.

2) I am also confused by the way the significance threshold is used. You can reject the null-hypothesis of non-coherent reconstructions if the coherence is above the threshold. But does that mean that two reconstructions are non-coherent if their mutual coherence is below the threshold? If the reconstruction period had been short enough the threshold value would be large enough for none of the coherences to be significant. How should that be interpreted?

Response: For short periods, not only would the threshold go up but also the coherence values themselves. If the average number of significant frequencies would depend on the period length we would have a pretty bad estimator.

3) The author uses exclusively an integrated value of the coherence as a measure of "similarity". I would like to see this described in more detail. How is the integrated coherence related to the cross-correlation coefficient of smoothed data (as used in Juckes

2007)? The author mentions that coherence "present a better protection against e.g. spurious significance". Better than for the cross-correlation? The author should be much more detailed here.

Response: Because of the things mentioned above I think average coherence gives a better estimate of similarity than correlation if the data are smoothed. In that case, as indicated in the paper, special care has to be taken for the reduced degrees of freedom ('spurious significance'), which is not trivial. This will be analyzed more thoroughly in the revision.

4) I guess that the coherence does not depend on the amplitude and offset of the reconstructions (this is also not very clear from the description in the paper). But many reconstruction methods differ very much in these aspects (Christiansen et al. 2009). I think the author should repeat his analyses with other indices of "similarity" such as the cross-correlation and indices based on the low-frequency amplitude. Perhaps that will give different results. It could be that the coherence is not a good measure. From Fig. 3 it seems that a large part of the low-frequency variability is common to most reconstructions. Most of them have the coldest period in the beginning of 17th century and the warmest in the beginning of the millennium.

Response: This is an important point. The first stimulus to write this paper was actually the question whether the common graphic display of the reconstructions may actually be misleading and whether that stands up to statistical scrutiny. For example, only if one takes "the beginning of the millennium" to be the period from 1000 to 1200 then all reconstructions are warmest at that beginning. And "the coldest period" is better described to be some time between 1300 and 1700. The crucial question is: having 'softened' our criteria enough, can we not bring *any* two series into accord this way, and is not what we see possibly insignificant? - A common negative millennial trend (which, interestingly, Keith Briffa was quite skeptical about itself) is not enough to prove truly coherent behavior; that requires synchronous variations over a wider range of time scales. (The occasional detrending in calibrating temperature-proxy relations is

based on this argument.)

amplitude and phase: I explicitly neglected amplitudes since to me this is just part of the reconstruction post-processing (chronologies being rescaled to whatever target). Phase shifts may be an issue and will be taken into account for a revision, although I doubt that the intra-cluster reconstructions reveal strong shifts.

All minor comments will be fixed accordingly. A few deserve extra replies:

p660,l18: What is "reasoning under uncertainty: and "paraconsistent logic"? The author should be more explicit here.

Response: I only wanted to note the existence of these topics instead of thoroughly describing them. If all reviewers agree that these references are misleading or too short I will drop them completely.

p662,l25: Note that Christiansen et al. 2009 showed that the reconstruction methods all have serious biases regarding the low-frequency variability. So even if the methods are consistent they may all miss the target. Christiansen et al. 2009 also found that the low-frequency correlation (with the target) was well reconstructed by most methods in agreement with the speculations in the present paper that the major part of the problem of the weak coherence lies in the different proxy compilations.

Response: Very good point that will be mentioned in the revision.

Fig. 3: There are only four panels to show the five clusters.

Response: Ma99 (last figure) is a single cluster. This will be clarified.

p666,l13: But they do seem to agree about the low-frequency variability. Perhaps coherence is not the best measure (see above).

Response: I disagree. The two time series are shown in the attached Fig. 2 (with Butterworth smoothing >20y). Simple counting of the 'single events' shows that they disagree much more often than they agree. Other than the negative trend they have

not much in common. This will be further analyzed (see my response to 3).

p666,l15: A verification procedure could be based on pseudo-proxy experiments. Such experiments could give information about the influence of the number of proxies, their positions, the choice of regression method etc.

Response: The use of pseudo-proxies has been discussed in the introduction. But as the reviewer rightly points out (by referring to Christiansen et al. 2009), the methods may be appropriate but not the proxies. To test the impact of proxy quality one would need much more refined error models for the pseudo-proxies.

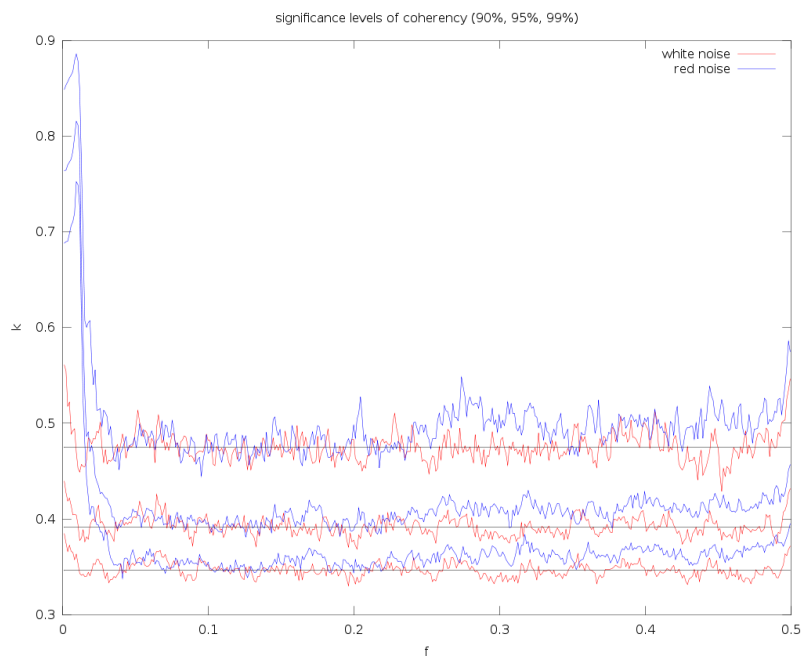Interactive comment on Clim. Past Discuss., 6, 659, 2010.
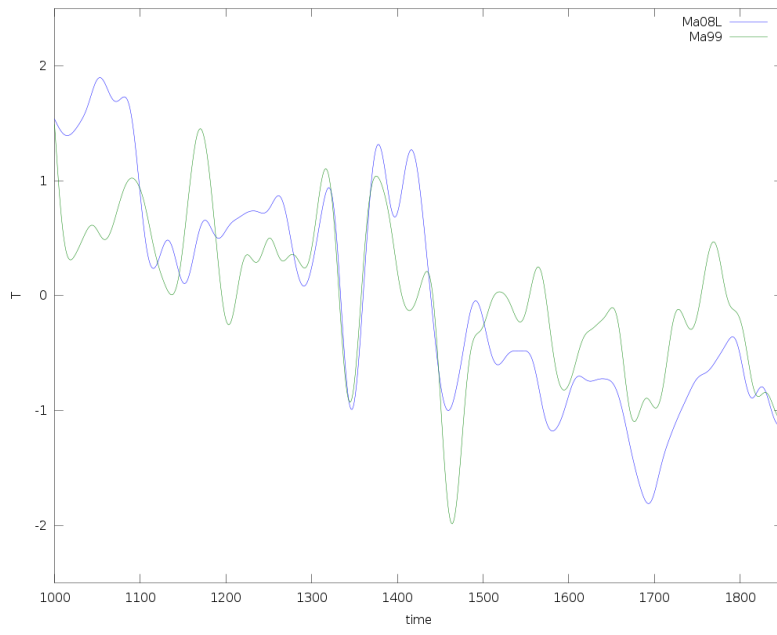
Fig. 1. Estimating significance levels

Fig. 2. The Ma08L and Ma99 reconstruction (smoothed)

C219