Clim. Past Discuss., 5, C1142–C1155, 2010
www.clim-past-discuss.net/5/C1142/2010/

Climate of the Past Discussions

Interactive comment on "Refining error estimates for a millennial temperature reconstruction" *by* M. N. Juckes

Anonymous Referee #2

Received and published: 5 February 2010

General Comments

A number of substantive issues must be addressed before this manuscript should be considered for publication in *Climate of the Past*. In addition, the manuscript should largely be re-written with a focus on clarifying both the presentation and the mathematical notation. The current presentation is such that a reader with more than passing familiarity with the mathematical and statistical techniques used in the manuscript cannot clearly understand what has been done. This must be remedied prior to the manuscript being considered for publication.

There is certainly a need for better uncertainty estimates in millennial scale temperature reconstructions, and it is good to see this problem receiving attention. In particular,

the impacts that modeling (or structural) errors, as distinct from observational errors (random imperfections in the proxy-temperature relationship), have on reconstructions is an area that requires further research. Several relatively recent publications have proposed more advanced techniques for estimating the uncertainty in reconstructions of spatial mean temperatures, and these must be referenced and discussed. Lee et al. (2008) is cited only as an example of a study that uses climate model output. It should be discussed further, as it, like this manuscript, attempts to better quantify errors. Li et al. (2007) is not mentioned, and is even more relevant to the current study, as it makes use of a resampling strategy, seeks to quantify structural uncertainty, and assess the probability that the late 20th Century warmth was unprecedented.

This manuscript suggests a method/technique for quantifying uncertainties, and also presents a (somewhat — cf. JAB2007) new reconstruction, with errors estimated according to the proposed technique. The author has not made the nature or scope of this contribution clear in the abstract and introduction, and I am not convinced that the author has sufficiently considered how this work contributes to climate science. By my reading, this paper serves one of three roles, and in each case, the manuscript should be adjusted to make the nature and scope of the contribution more clear.

- The main contribution of the paper is methodological, and the incremental improvement to an existing reconstruction is included an initial, largely demonstrative application of the method. If the author shares this view, than the paper should be re-written to clearly focus on the method, should include far more references to the statistics literature, more discussion linking these methodological developments to previous work, and the application should clearly be presented as an example. In this scenario, the introduction of new data is an unnecessary source of complication, and I would recommend that the author use an existing, published data set as a test-bed for the proposed method.
- The main contribution is the new reconstruction, and the refinement of error a

secondary contribution. If this is the case, then the abstract and introduction should emphasize the new reconstruction, and make more minor mention of the error estimates. More emphasis should be placed on the discrepancies between this work and JAB2007, and the methodological discussion should be confined more to the appendices.

- If the author views this paper as making two separate contribution, one methodological and one applied, then this division should be clearly stated in the abstract, and made clear throughout the paper. The author might consider (or the editor might suggest) breaking the paper into two separate publications, one methodological and one applied, which would allow for a more complete treatment of each aspect of the current manuscript.

The author elects to use quite a sophisticated statistical technique (the delete–*d* jackknife), but does not justify this decision. There are two references to the fact that this method is more robust to departures from normality, but no discussion of the extent to which the data violates that assumption. There is no discussion of why the jackknife was selected over the more well known bootstrap, or if the results are substantially different from a much simpler approach that assumes normality. The assumption that proxies are unbiased estimators of the true mean is implicit to the entire analysis, and this must be made explicit.

Abstract

The jackknife should be mentioned explicitly, as it is the main analysis tool presented in the manuscript.

P1L10: Grammatically incorrect.

1. Introduction

P3L3: The true, unkown proxy-temperature relationship is a statistical relationship, as the proxies add some form of noise to the true temperatures. If the form and parameters of this relationship were known, statistical methods would still be necessary. The problem is further complicated by the need to estimate both the form and the parameters of this relationship.

P3L17: “the uncertainty estimate given by the Jackknife method would be zero." This is misleading - the uncertainty estimated using ANY reasonable method (sample standard deviation, interquartile range) would be zero. Also - the meaning of the word “uncertainty” should be given a definition at this point. Is the “uncertainty” the standard error in the estimate of the mean of the proxies for any given year? The rmse between the “best estimate” from the proxies and the unknown truth? These quantities may be very different, and this issue must be clarified.

P3L19: “The uncertainty thus reflects inhomogeneities in the proxy data and the extent to which the true NH-mean temperature variations are captured by the retained data." What is meant by the word “retained”? The statement is misleading, if not incorrect. If each proxy time series was simply the true underlying NH mean temperature, plus iid white noise deviates, than the “uncertainty” in each proxy time series (taken as the variance) can be arbitrarily large, but with enough proxy time series, estimates of the NH mean can have very small errors. On the other hand, if each proxy suffers from the same, unknown systemic bias, then the error in the estimates of the NH mean do not converge to zero as the number of proxies increases. Basically, the variance across the proxy time series and the uncertainty in the estimate of the NH mean temperature are not the same, as the mean squared error is given by the sum of the variance and squared bias of the estimator. The notion of bias must be discussed, and the implicit assumption that each proxy is an unbiased estimator of the spatial mean must be made explicit.

P3L20: Two data sets are mentioned: a 13 proxy compilation from an earlier publication, and a new 15 proxy compilation that includes 12 of the proxies used in the earlier publication. These are clearly not independent, but P3L15 mentions two independent data sets. Please clarify this point.

2. Data

What impact doe leaving out the Arabian sea series have on the analysis? The justification that it is “at best an indirect indicator of temperature" seems weak, and applies to all proxies. Are the conclusions affected by this modeling decision?

Much of the proxy data section (2.1) discusses a Mann et al data set, and a review of a number of proxy series. What is the point of this information? How does this discussion relate to the choice of the proxy series used in this study? This section must be clarified.

P5L2: Sentence is long and unclear.

3. Temporal homogeneity of the proxy data

The phrase, “stationarity in time” would be more appropriate.

P6L16: This is not a covariance, not a correlation. The standard notation for a sample correlation is r, and this is used on page 9. More generally, the mathematical notation throughout this work needs to be improved.

4. Northern Hemisphere temperature

Why has a method-of-moments estimator been used to determine the slope? Why not MLE?

P7L23: Standardizing must be mentioned earlier — see previous comment.

P8L1: “Entirely independent of the data used here." This is misleading, given the normal definition of the word “independent” in statistics, and the fact that all proxies under discussion are assumed to reflect some aspect of spatially averaged temperatures. I assume the author means that no proxy time series is common to the two compilations.

P9: I assume that the spatial distribution of the Wilson compilation is different from that of the R15 compilation, and that this discrepancy impacts the estimates of the spatial mean. This issue warrants discussion.

P9L21: State r^2 rather than r, as it is easier to interpret.

P9L23: How was the significance level calculated? Provide a reference or show a formula.

P9L24: The standard null hypothesis is that the correlation between the two series is zero. The conclusion is misleading - what is meant by “Fluctuation?" What if the small scale “fluctuations" in two time series are independent, but each shares a deterministic trend? Then the random part of each would be independent, but the correlation still high. The only conclusion that can be made is that, at a high significance level, the two series are not independent. Please correct the discussion.

5. Uncertainty Estimates

5.1 The Jackknife

Why has the author elected to use a delete-d jackknife? Why was this chosen over the more conventional and well-known bootstrap, or even the delete-1 jackknife? An explanation of this choice must be provided, and I suggest consulting the classical literature on these techniques, such as Efron (1979) and Efron and Gong (1983), to inform both the choice of analysis techniques and to justify this choice.

The notation throughout this section *must* be improved. In particular, it is very difficult to distinguish between various population, sample, adjusted, partially adjusted, and fully adjusted variance parameters. At any given time point, a number of different variances come into play:

- The sample variance of the proxies.
- The population covariance matrix of the proxies.
- The sample estimate of the variance of the mean of the proxies.
- The population variance for the distribution of that mean.
- For a given jackknife ensemble member, the sample variance of the retained proxies.
- Each value of d results in a number of ensemble members, and each ensemble member gives an estimate of the mean. The sample variance of these means.
- Various corrected and uncorrected jackknife estimates of the variance of the distribution of the proxy mean.

The notation must be clarified to distinguish between these quantities. In statistics, it is common to identify sample estimates with a hat (), or by referring to population variances as σ^2 and sample estimates as s^2.

The end goal of this section is to quantify the uncertainty in the time series of sample means. Such an estimate requires an estimate of the variance of the proxies (assuming that, at each year, they are iid) or their covariance matrix. A jackknife procedure is used to estimate the variance in the distribution of the mean, and a correction is made to account for the covariance between proxies. These notions must be clarified, and the notation should be completely redesigned — currently, the mathematics obscures, rather than clarifies, the method. Section 5.1 cannot be followed — see in particular, P10L10 and the ensuing paragraph. Which mean? Each ensemble member produces a spread of proxy time series, and thus a mean time series. Each value of d produces an ensemble of such means, which in turn has a mean. I cannot discern which quantities are under discussion.

P10L7: Why are none of the ensemble members shown in a figure?

P10L13, P11L15: “Reduced assumptions of normality." Please clarify, provide additional references, and explain why there is reason to suspect non-normality in the estimate of the mean of a number of proxies. Conditional on the true mean temperature time series, is there reason to believe that the proxies are not normally distributed about that mean? The author has chosen to use a sophisticated statistical technique, and should include a justification for doing so. Are the results any different form simple procedures? I don’t understand the last two lines of this paragraph.

5.2 Estimating proxy error correlation

The notation in this section is confusing. The author basically derives the distribution of the mean of N_c correlated normal random variables. I suggest expressing these

formulas in terms of covariance matrices, and then noting how the formulas can be re-expressed in terms of the mean of the diagonal and off-diagonal elements of that covariance matrix. As an example, assume that,

$$\mathbf{X} \sim N(\mu, \mathbf{\Sigma}),$$

where $\mathbf{X}$ is a vector of length N. The mean of the elements of $\mathbf{X}$ is simply $\frac{1}{N} \cdot \mathbf{1}^T \mathbf{X} \equiv \mathbf{Y}$, and the resulting distribution is,

$$\mathbf{Y} \sim N(\frac{1}{N} \cdot \mathbf{1}^T \mu, \frac{1}{N^2} \cdot \mathbf{1}^T \mathbf{\Sigma} \mathbf{1}).$$

The covariance form simply sums all elements of Σ, and scales that sum by N^2, and this representation is more intuitive than that used by the author.

P12L5: How does this formula relate to Eq. 2? The logic is hard to follow. The relevant formula is derived in Appendix B, not Appendix A.

P12L9: Note that if $c = 1$ this formula reduces to $0/0$, so does not in fact adjust for this possibility. It is simply a correction to take into account the covariance structure of the proxies. Notation: $\sigma^2_{d:jack:c}$ is cumbersome. As mentioned above, the use of a hat (̂) or s^2 to clarify that this is a sample estimate would be helpful.

5.3 Structural Uncertainty

This section requires major revisions, and is currently not appropriate for publication. Equation. 4 is unnecessary and distracting — a vague definition of a pdf is not required. The key idea in this section, as far as I understand, is very simple: $P(X, Y) = P(X|Y)P(Y)$, and integrating out Y then gives the marginal distribution of X.

I don't understand the two uses of γ_{vm}, in both Eq.5 and Eq.6. The goal of the calculation is to integrate out the γ. I believe the idea is that, setting some reference slope,

the temperature distribution for a different slope is found by scaling by the ratio of these slopes inside the argument of the pdf for T. The notation obscures this idea.

P14L2: What is meant in this context by γ_{vm}? The quantity in Eq.2? In that case, make this connection explicit.

5.4 Summary

The logic here is flawed, and I don't see how the calculation is relevant to the problem. Indeed, the whole first paragraph of this section is superfluous, while the last two simply seem out of place (one sentence paragraphs are rarely appropriate). The question at hand is, "Given one particular realization of the climate system, and assorted measurements on that system, with what certainty was 1998 the warmest year." To simplify the issue, consider an experiment consisting of N independent coin tosses, and uncertain data about the outcome of each toss. The goal is then to determine if any of the tosses resulted in 'heads.' Each toss has some probability of being heads, but the relevant question is, "with what certainty can we say if there was *at least one* heads."

The uncertainty intervals calculated in this manuscript are point-wise frequentist confidence intervals. The correct interpretation, for one fixed time point, is that in repeated realizations of this experiment (both the climate system and the proxies), 95% of intervals calculated in this way will contain the true value. It is not correct to say that with 95% probability, the true value for a given year falls within the uncertainty bounds. Autocorrelation in time hinders an easy interpretation of the uncertainty envelopes through time, and the shape of the "true mean time series" might be very different from the estimated mean.

From the frequentist viewpoint, each year in the past either did or did not exceed the temperature in 1998, and the statements made by the author do not really make sense. If the author wishes to make probabilistic statements about particular years exceeding

the 1998 temperature, I suggest adopting a Bayesian approach to this problem. In the Bayesian framework, the unknown temperature in any year is assigned a probability distribution, whereas in the frequentist paradigm, the temperature for that year is treated as a fixed, unknown parameter.

The approach taken in the next section moves in this direction, where probabilities are assigned by calculating the proportion of ensemble members that meet some criterion.

6. Uncertainties in specific statements

Reference should be made in this section to Li et al. (2007). This section does not actually discuss the results presented in Table 3. Please include information directly in the text about the proportion of ensemble members that meet each criterion.

P45L8: First sentence is grammatically incorrect.

7. Conclusions

Calling this a new reconstruction is a stretch, as a previous study made use of all but four of the proxies, and the method used to calculate the point estimate of the mean time has not changed between these two studies.

This last section should emphasize methodological, as well as the scientific, advances.

P17L1: Combine with previous paragraph.

P17L10: Re-state the IPCC assumptions that are being discussed.

Appendix B

See comments with regards to the notation at P11L20. Using matrices and products with vectors of ones would clean up this presentation.

I suspect that this derivation is a well-known result. The author should search the statistics literature for a reference.

P19L1: Notation? Maybe use $\{\}$ notation to indicate sets.

P19L5: "Now consider the anomalies of the Jackknife ensemble elements relative to the mean of the ensemble." What is being referred to — for a given ensemble member, the spread of the proxies about the mean value, or the spread of the mean of each ensemble member about some grand mean? Please clarify the presentation.

Eq.B1: What is a_i? It is never defined, which makes this derivation impossible to follow.

I think there is also confusion with resect to α — is this the number of subsets of d elements, or the number of elements in the complement?

To clarify the derivation, I suggest first presenting an example with $N = 3$ and $d = 1$, in which case there are 3 possible jackknife ensemble members, and then generalizing.

Appendix C

Reference Fig. 7 for the max and min slope values.

The asymmetry seems awfully small — the three slope values are $(0.12, 0.18, 0.26)$. Is there a strong a priori reason to believe this is a real effect? What are impacts of using symmetric forms for the pdf?

Note that the three different regression lines correspond to three different assumptions

about the relative magnitudes of the variances of the errors. Standard regression assumes that all error is in the response variable (instrumental temperatures, in Fig. 7), "inverse" regression assumes all error is in the predictor (the proxy composite), while the "variance matching" looks (to my eye) a lot like what would result from orthogonal least squares, which assumes equal errors in the predictors and response.

In all likelihood, the errors associated with the proxies are larger, which should inform the choice of the pdf, and this is in line with the decision to put more weight in values greater than γ_{vm}. This issue should be discussed.

Tables and Figures

Table 1: Why are the numbers in column 4 different, and likewise for column seven? Why is the jackknife estimate sensitive to d? Table 3: The "statements" being tested should be stated in the caption, and the U/G notation should be explained.

Table 3: What is σ? Is this a corrected jackknife estimate? Please tie this estimate back to the description of the methodology.

Fig. 2: Explain in the caption that the black line is a measure of the average covariance between the tree ring and non tree ring proxies.

Fig. 3: The notation in the caption seems different from that used at P8L5, which refers to this figure. What are Union and Union++?

Fig. 5: Are the solid lines necessary? According to Eq. 2, they are incorrect. As the dashed lines are indistinguishable, I don't think the figure is necessary. A simple comment could be included in the text to the effect that after the correction of Eq. 2 is applied, the 5th and 95th percentiles are virtually identical for d=4,5,6.

Fig. 6: Many of the comments for Fig. 5 apply here, and I question if all lines are

necessary. The fact that the various corrections reduce the 5th percentile more than increasing the 95th is interesting, and worthy of more commentary.

Fig. 7: Identify the three different lines in the caption or using a legend.

Fig. 8: Again - are all lines necessary? If the point of the figure is to illustrate that 1998 was 'anomalously' (in some sense) warm, than the 10th, 90th, 33d, and 67th do not really add much.

Figs. 9,10: See comments for Figure 8.

References

B. Efron. Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1):36–48, 1983.

T.C.K. Lee, F.W. Zwiers, and M. Tsao. Evaluation of proxy-based millennial reconstruction methods. *Climate Dynamics*, 31(2):263–281, 2008.

B. Li, D. Nychka, and C. Ammann. The'hockey stick'and the 1990s: a statistical perspective on reconstructing hemispheric temperatures. *Tellus-Series A-Dynamic Meteorology and Oceanography*, 59(5):591–598, 2007.

Interactive comment on Clim. Past Discuss., 5, 2631, 2009.