REPLIES TO REVIEWERS AND AUTHORS COMMENTS

GENERAL COMMENTS TO EDITOR:

The MM comment does not meet the standards for publication in a 'Communications Arising'. We have demonstrated that the **central conclusion** of MBH98 is in no way brought into question. We show that the results of MBH98 are robust, and that MM04's criticisms are specious and without merit. Moreover, in the **6 years since our paper was published**, numerous independent reconstructions have come to the same conclusions as MBH98. We show that the conclusions reached by MM04 **result from the inappropriate use of statistics, rather than anything that would enlighten the reader** about the true nature of temperature variations over recent centuries.

Specifically,

1) The two main claims of MM04 are demonstrated to be false. The results of MBH98 are, as we demonstrate, in no way dependent on the procedure used to calculate PCs of tree-ring networks, or the existence of missing (infilled) values in one series from AD 1400-1403.

2) The reconstruction of MM04 completely fails statistical verification, in contrast to that of MBH98. We note that Reviewer #2 stated ***"Should this validation be successful, I would recommend the publication of both manuscripts"***. However, McIntyre and McKitrick now clearly confirm our assertion that their reconstruction (unlike ours) fails objective statistical verification. Remarkably, they now promote the use of a statistic (R^2) which is known to be an inappropriate measure of verification skill as it does not consider changes in mean or variance outside the calibration interval. MM04 have also made mistakes in their calculation of verification statistics, as discussed below. Based on the specific standard clearly set by reviewer #2, the MM comment should be rejected.

3) Their own spurious reconstruction is reproduced only through the elimination of the Northern Treeline and ITRDB North American data sets. This is what they did in MM03, and have effectively done in MM04 by failing to apply the appropriate rule for the number of eigenvectors to retain.

4) Their reconstruction disagrees with all other northern hemisphere temperature reconstructions in its indication of anomalous early 15th century warmth (see Supplementary Information #4).

Below we provide specific responses to each of the referees/authors comments:

RESPONSES TO REFEREE #1

>I find merit in the arguments of both protagonists,

As we show in our revised manuscript, there is no merit at all to the main arguments put forward by MM, which are specious in nature.

> ...though MBH is much more difficult to read than MM. Their explanations are (at least superficially) less clear
> and they cram too many things onto the same diagram, so I find it harder to judge whether I agree with them.

We have shortened and clarified our response, and we have simplified the diagrams. We hope our revised response is clearer to the reviewer.

>There are two main points of dispute:
>
>1. The principal component technique used.
>
>2. The quality of the early data.
>

Both involve false claims on the part of MM, as we show in our revised reply.

>I deal with 1. first. It is an area where I have expertise, but it is not at all clear what exactly is being done. MM talk
>about 'scaling to 1902-1980 mean and standard deviation', whereas MBH's phrase is 'restandardization by
>detrended standard deviation'. These suggest different things to me. The latter seems more appropriate than the
>former, but I am still uneasy about applying a standardization based on a small segment of the series to the whole
>series, if that is what is being done.

There are several different issues here. The method of MBH98, by its nature, requires that proxy data be standardized over the calibration period. Each predictor must have the same mean (zero) over the calibration period, for the regression approach to be applied. Furthermore, each predictor must be normalized over some reference period, because different proxy records in general have different units. Thus, the proxy data, as the instrumental record, have been standardized (subtraction of mean, division by standard deviation) over a 20th century overlap interval. MBH98 chose, as mentioned above, to calculate the standard deviation after detrending over the calibration period, precisely so that series with large trends were not overly emphasized. In the analyses presented here, however, we have chosen to normalize by the nominal standard deviation (i.e., not the detrended standard deviation), to address any criticism of this convention, and to demonstrate that the result is insensitive to the particular normalization convention. We have also used a 1902-1971 calibration interval to avoid a moderate number of infilled proxy values between 1972 and 1980 to which MM have objected. Most importantly, we have demonstrated the falsehood of the claim by MM04 that the centering convention used by MBH98 in the PCA of the North American ITRDB data has any influence on the main features of the MBH98 reconstruction (see also 'Supplementary Information' #1 and #2). We furthermore present the results of analyses in which all individual North American ITRDB data (rather than PC representations of these data) are used, and with equal weight. Thus, as noted below, the claim by MM that it our convention for calculating PCs of the North American ITRDB data that governs the character of the MBH98 reconstruction *cannot possibly be true*.

>MBH seem to be too dismissive of MM's red noise simulations. Even if red noise is not the best model for the
>series, they should have reservations about a procedure that gives the 'hockey stick' shape for all 10 simulations,
>when such a shape would not be expected. Having said this, MM's corrected series in their Figure 4c still has the
>upward trend towards the end of the series, so this trend is not just an artifact of MBH's PCA procedure.

The first point is a fair one. We should have addressed this in more detail in our original reply. The MM simulations are inappropriate because they assume stationary data (i.e., data which, by construction, do *not*, like many of the actual tree-ring data, have a statistically significant 20th century trend) and they ignore spatial correlation within the dataset (the 70 North American ITRDB tree-ring series do not represent N=70 statistical independent noise processes but, like the climate, exhibit large-scale spatial correlation structure). We have repeated their simulations, and shown that the claimed result is not true. Changing the centering of the data does not produce a spurious pattern of variance (in particular, a 'hockey stick' pattern) that does not exist in the actual data. It simply, in these somewhat artificial examples, alters the prominence of different patterns of variance in the data, or decomposes them into linear combinations of other leading patterns (see our "Supplementary Information 2"). Using a modern reference period indeed gives greater weight to those patterns of variance in the data exhibiting long-term departures from the recent mean in these experiments, but it does not produce those patterns. Because the eigenvalues are degenerate in these experiments (i.e., there are no eigenvalues distinct from a noise floor--see our "Supplementary Information" #1 and #2), it is easy to obtain a different basis set when the reference period is changed even slightly. But the basis vectors are still linear combinations of each other. This is all that is going on here.

In the case of the actual ITRDB data, the situation is very different. The data exhibit large-scale spatially-correlated structures that are highly significant relative to a red noise null hypothesis. There are at least 2 statistically significant eigenvectors using the MBH98 centering, and at least 5 significant eigenvectors using the MM04 centering, relative to a red noise hypothesis--see Supplementary Information #1 and #2).

Thus, in the actual ITRDB data, the low-frequency patterns of the individual PCs are far more robust with respect to any re-centering of the data. For example, the PC#1 pattern of the actual North American ITRDB data based on the MBH98 centering convention *appears in essentially identical form* in the PCA analysis based on the MM04 centering convention, albeit slightly farther down (PC#4) in the eigenvalue spectrum ("see Supplementary Information #1"), something that was not disclosed by MM04. Applying the selection criterion used by MBH98 for these data, as discussed above, the first 5 PCs of the re-centered PCA analysis should be retained. **A reconstruction employing a network including these 5 indicators gives essentially the same result as MBH98 ('Supplementary Information #1')**

More fundamentally, however, we have demonstrated that our reconstruction is not sensitive to the use of any PC representation of the ITRDB data at all. We have now provided the results of an analysis in which all 95 individual proxies available back to AD 1404 are used with equal weight, and an additional analysis in which the "Gaspé" series challenged by MM (see below) was removed, giving 94 indicators back to AD 1400. A 1902-1971 calibration period was used that avoids a modest number of infilled missing proxy data between 1972-1980 objected to by MM03. **The resulting reconstructions are remarkably similar to that of MBH98, with the same characteristic 'hockey stick' pattern, and with no evidence of the spurious early 15th century warmth obtained by MM04.** We show that this latter spurious feature can only be obtained through the elimination of the low-frequency information in the ITRDB dataset (which MM03 achieved through the explicit censoring of the ITRDB data prior to AD 1600, while MM04 achieve this through an inappropriate truncation of their EOF analysis, as discussed above).

We thus assert that the principal objection of MM04, that the main features of the MBH98 reconstruction are somehow dependent on the details of the PCA representation used to represent certain tree-ring networks, is absolutely false.

>I am not qualified to say much on 2. but it seems to be the crucial point. Both sets of authors agree that the
>omission of some early data changes the early reconstruction considerably. MBH say that the omitted data are
>reliable; MM say they are not.
>Does anyone know who is correct? If there is disagreement among experts, then the true behaviour of the series
>must be very uncertain.

MM dwell on the 4 missing years in the 'Gaspé' series. We have demonstrated that the hemispheric reconstruction is not sensitive to the use of this indicator anyway (see above), so this is a moot point. The remaining data issue then involves the ITRDB North American tree-ring data. One of us (Malcolm Hughes) is among the foremost experts on dendroclimatology, and, in particular, the ITRDB data. Hughes applied a thorough quality control and screening procedure to this data set for our analysis, emphasizing issues of replication within each chronology, correlation between constituent samples in each chronology, nature of biological growth trend removal, and minimum segment length (see Mann, M.E., Gille, E., Bradley, R.S., Hughes, M.K., Overpeck, J.T., Keimig, F.T., Gross, W. , Global Temperature Patterns in Past Centuries: An interactive presentation, *Earth Interactions*, 4-4, 1-29, 2000. An online link is available here: http://holocene.evsc.virginia.edu/Mann/articles/articles.html). MM raise the issue of possible 'CO_2 fertilization' influences on growth, yet this was already dealt with by Mann et al (1999). Any apparent influence of this effect was removed, as described in that paper, and the corrected data were used in the hemispheric temperature reconstruction. There was no significant influence on the post AD 1400 reconstruction.

>Incidentally, I am not entirely convinced by MBH's dismissal of the MM model reconstruction on the basis of RE.
>I suspect that a lot of the difference is due to the much larger variance in the MM model reconstruction compare to
>MBH's. This is probably inevitable, given the reduced sample size for the early data.

The reviewer has put his/her finger on the issue. We have tried to clarify this with an appropriate graphic in the revised reply, that shows that it is indeed the consistently greater variance in the MM04 reconstruction noted by the reviewer that gives both the anomalous 15th century warmth and the unrealistic 19th century warmth/ increased variance so at odds with the pre-calibration (1854-1901 in MBH98, 1856-1901 in MM04) instrumental data. The large negative *RE* score (-0.8) is a faithful representation of this mismatch and of the failure of their statistical model. We have provided references to the literature, back to Lorenz (1956), favoring *RE* as the preferred statistical diagnostic of predictive (or reconstructive) skill. We stand by the assertion that the *RE* score for the MM model, which is essentially equivalent to that expected for a purely random estimate, is in fact grounds for dismissal of the model. Indeed, another reviewer has essentially agreed that this is the standard by which the merit of the MM04 comment should be judged. On these grounds, they have clearly failed to meet this standard, even to their own admission (though their *RE* values are not correct--see below). The issue of statistical calibration and verification is discussed in more detail in our "Supplementary Information #3".

RESPONSES TO REFEREE 2

>The technical criticisms raised by McIntyre and McKritik (MM) concerning the temperature reconstructions by
>Mann et al (MBH98), and the reply to this criticism by Mann et al is quite difficult to evaluate in a short period of
>time, since they are aimed at particular technical points of the statistical methods used by Mann et al, or at the use
>of particular time series of proxy data. A proper evaluation would require to redo most of the calculations
>presented in both manuscripts, something which is obviously out of reach in two weeks time. Furthermore, both
>manuscripts seem to contradict each other in some basic facts.

The reviewer's comments are well taken. However, it should be noted that the basic result of MBH98 has been reproduced by other groups, using independent methods and proxy data (see "Supplementary Information #4"). So the outlier, or the contradiction, is clearly in the MM04 result, and the burden of proof through verification, as this reviewer notes later on, was indeed on them. As discussed below, they have clearly failed in meeting that burden.

>Therefore, my comments are based on my impression of the consistency of the results presented, but there is a
>wide margin of uncertainty that could be resolved only by by looking in detail into the whole data set and the
>whole software used by the authors.
>In general terms I found the criticisms raised by McIntyre and McKritik worth of being taken seriously. They
>have made an in depth analysis of the MBH reconstructions and they have found several technical errors that are
>only partially addressed in the reply by Mann et al.

As detailed in our revised reply, there are no technical errors in the MBH98 reconstructions (in another paper in review, we address each of the individual criticisms put forward originally by MM03). The criticisms by MM04, as outlined below, are specious.

>1)Mann et al assert that important features in the reconstruction by MM, for instance the increased warmth in the
>15th century, is due to the fact that they completely ignore the time series from the NOAM tree ring data sets.
>However, MM explicitly state that they have used the two leading PCs of this data sets. Of course, it is impossible
>to ascertain who is right and who is wrong in this particular point, but I feel that Mann et al should have taken into
>account in their reply the statement by MM concerning the NOAM time series.

We have clarified this point, on which there may have been some confusion. MM03 eliminated these data entirely. MM04 have effectively eliminated the low-frequency component of variability in this data set (the pattern of variability in the data that serves as a skillful indicator of large-scale surface temperature changes) through an indefensible truncation of the eigenvector basis representation of these data. As discussed above, by re-centering the data relative to the MBH98 convention as MM04 have done, the MBH98 PC#1 pattern shifts to PC#4. However, while the selection rules dictate that only 2 PCs should be retained in the PCA analysis using the MBH98 centering convention, the same selection rules indicate that 5 PCs should be retained using the MM04 centering convention (see "Supplementary Information #1"). Thus, the same key pattern of low-frequency variability in the dataset should have been retained as an indicator. However, MM04 incorrectly truncated at the 2nd eigenvector. As shown in "Supplementary Information #1", this decision is indefensible. Through this decision, MM04 effectively filtered out the key low-frequency pattern of variability in the data. As discussed above, inclusion of the correct 5 PCs for the MM04 centered PCA as indicators yields a very similar reconstruction to MBH98 ("Supplementary Information #1").

>2)My doubts expressed in point 1 are strengthen by Figure 1b in Mann et al. reply. Mann et al. have tried to
>replicate in this figure the MM reconstruction (MM04c, green line). But one can clearly see that the variance of
>this reconstruction is much larger than the observations even in the calibration period. Although it might be
>possible, it seems a very awkward result. Any linear regression method that I am aware of must produce a
>reconstructed predictand with less variance than the observed predictand (the rest of the variance being the
>residuals). It seems to me that the something is technically not correct in the replication by Mann et al of the MM
>reconstruction. If the main difference between the original MBH98 and the MM04c reconstruction is just the
>elimination of the NOAM data set, why is the variance of the reconstruction in the calibration period inflated by a
>factor of 2-3?.

The MBH98 method, if followed correctly, indeed insures the correct variance (and positive calibration *RE*), in the reconstructed PC series over the calibration interval (thought not *necessarily* the NH series--only the PCs are actually statistical predictands!). This is insured by MBH98 by scaling the reconstructed PCs to have the same variance as the annual instrumental PCs over the calibration period. As we understand the MM04 implementation, they did not follow this latter step--thus, overfitting in the calibration (and negative calibration *RE* scores) is actually possible, and in fact observed. If we censor the predictors as in MM04, but apply the MBH98 protocol which assures the proper scaling of the PCs during the calibration interval, we indeed obtain positive calibration *RE* scores for the NH series (*RE*=0.21), but still a significantly negative verification *RE* score (-0.16), and the same spurious 15th century warm peak, for the NH reconstruction (see "Supplementary Information #3").

>3)The reply by Mann et al is in my opinion correct when requiring MM to present some validation statistics in a
>validation period, and the RE statistics seems to me adequate in this context.

We are glad that the reviewer recognizes that the *RE* statistic is an appropriate metric for evaluating reconstructive skill. MM04 remarkably argue for using instead an R^2 metric, even though it is well known that this does not account for changes in mean and variance outside the calibration interval (we have added some discussion of this point to our reply).

>Since this is the main argument in
>Mann et al reply I would urge MM to address this criticism. The low value of the RE statistics in the replicated
>MM reconstruction (MM04c) indicated by Mann et al. seems to be due to the erroneous replication of the MM
>reconstruction. An inspection by eye of the MM04c reconstruction seems to indicate that both reconstructions, the
>replicated MM04c and the MBH98 method (1400-1500 model) are more or less equally correlated with the
>instrumental temperature in the calibration period, and that the low RE value of the MM04c reconstruction stems
>from its much larger variance. I think that this large variance is unrealistic and therefore the real RE value should
>be positive.

We respectfully assert that the reviewer is mistaken. The spurious features of the MM04 reconstruction are intrinsic to the insufficient nature of the predictor network to produce a skillful reconstruction. The increased variance in the calibration period, as discussed above, arises when the MBH98 protocol is not followed correctly (as we believe is the case with MM04 with regard to the scaling of the reconstructed PCs). This is reinforced by the fact that the verification score they themselves calculate for their fully censored network (*RE*=-1.04) is similar to that obtained when we use the same censored network, and do not enforce the constraint that the PCs have the correct scaling (*RE*=-0.76). If we enforce the correct scaling, as discussed above, a positive NH calibration *RE* score is indeed obtained, and the calibration period variance does not exceed the variance in the instrumental NH record, but a negative verification score is still obtained (-0.16). The verification period variance still exceeds that of the instrumental record during the verification interval, and there is a substantial mean (warm) bias relative to the instrumental record. The reconstruction of spurious 15th century warmth is similar. We suspect that MM04 *did not* enforce the correct scaling, however, given the much more negative *RE* score they themselves calculate for their reconstruction! However, this doesn't really make any difference. In either case, the same combination of positive bias in both mean and variance evident in the verification interval (and responsible for the negative verification *RE* statistics) is presumably responsible for the spurious anomalous 15th century warmth. Again, more details are discussed in 'Supplementary Information #3'.

>4) The MM reconstruction presented by MM (Fig 4, bottom) should be in principle very similar to the replicated
>reconstruction MM04c, but it seems to me that they are not. For instance in Fig4, I do not see this excess variance
>compared to the original MBH98 reconstruction. This suggests again that something is not correct in the
>calculation of MM04c by Mann et al.

There appears to have been an important misunderstanding on the part of the reviewer with regard to what MM04 have actually shown. Unlike us in our reply, they did not/have not shown the 'frozen' reconstruction based on the network of indicators available back to AD 1400 continuously through the end of the calibration interval. Instead, they show a stepwise reconstruction, which is based on a completely different predictor network after about AD 1600 (and becomes asymptotically similar to the MBH98 reconstruction over time). This has had the effect of hiding the persistent bias in mean and variance responsible for their anomalous 15th century warm peak by not showing the reconstruction for the same set of predictors, continued beyond AD 1600. We have shown the latter, so the source of the early bias becomes clear. We have tried to clarify this very important distinction in our revised reply.

>In summary, my recommendation is that MM offer validation statistics for their reconstruction and that they make
>their original reconstruction (Fig 4 , bottom) available to MBH, so that these authors can also compute validation
>statistics with the original MM reconstruction. Furthermore, it should be explicitly cleared up if MM are using or
>not the NOAM data set. Should this validation be successful, I would recommend the publication of both
>manuscripts.

We agree entirely with the reviewer, that the essential standard for publication of the MM04 comment is that they had been able to demonstrate that their alternative reconstruction passes verification. Unfortunately, MM did **not** provide their series for verification (i.e., their reconstruction based on the AD 1400 network, through the validation period), so it was impossible for us to estimate verification statistics for their actual reconstruction. However, as discussed above, we have closely reproduced their reconstruction, and have shown that it fails cross-validation (regardless of the issue of precisely how the PCs are scaled, which is dealt with above). Moreover, they have come back with confirmatory evidence of our claim that their reconstruction (that yields anomalous 15th century warmth)_ indeed fails verification, quite dramatically. Faced with this evidence, they have now tried to shift the standard by arguing that (a) *RE* is not an appropriate metric for statistical verification and (b) that the skill diagnostics of our own reconstruction, which do meet this standard, are not as high as claimed.

Both claims are completely false. Regarding point (a), as we note in our revised manuscript, *RE* is indeed the preferred metric for diagnosing predictive/reconstructive skill, originally introduced by Lorenz (1956), and accepted as the appropriate standard widely throughout the fields of atmospheric science and climatology for decades since. MM04 seek to instead promote the use of an R^2 statistic instead, even though it quite well known that this statistic is insufficient as a measure of skill because it *does not assess changes in either mean or variance outside the calibration interval* [see e.g. Wilks, 1995, chapter 7, equations 7.19 and 7.20--note that Wilks defines RE as "skill score ("SS")] .

On point (b), the MM04 NH mean verification statistics are all wrong for a simple reason. They have not taken into account the mask of available 19th century observations to ensure that hemispheric means being compared represent averages over the same region of the Northern Hemisphere. This issue, which requires a 'frozen grid' analysis, was discussed quite clearly in MBH98, and yet MM04 have completely failed to take note of it. MM04 have not taken into account the time-variable nature of the spatial information contributing to the 'variable grid' NH mean series they downloaded from the CRU website. Thus, during the verification interval, the nominal instrumental Northern Hemisphere series represents a mean of only a small number of regions, while the MBH98 reconstruction represents a spatial average over the full region available during the 20th century calibration period. If the mask of 19th century instrumental observations is not applied to the spatial reconstruction, a meaningful comparison of domain-averaged quantities is not possible. The details are discussed further in "Supplementary Information 3". This error by MM04 will lead to an underestimate of all of their verification statistics, and at least partially explains why they underestimate the *RE* score of MBH98 by nearly a factor of two (0.3 vs 0.5). The other reason the *RE* score is too low probably relates to an incorrect 'reproduction' of MBH98 by the authors, due to their incorrect scaling of reconstructed PCs, as detailed above.

>Further evaluation requires months of work and should be left, in my opinion, to the scientific community. This
>should be the normal scientific process. At this stage, I think any Correction or Retraction by MBH98 is
>premature and really not required.

Two papers in press and in review show that the MBH98 reconstruction is robust to the use of an entirely independent statistical methodology, and to the data issues raised by MM. In "Supplementary Information #4" we provide a plot comparing the MBH reconstruction with other published reconstructions, to emphasize that independent already published work broadly reaffirms MBH98, and lies in stark contrast with MM's anomalous early 15th century warmth. Jones et al (1998), Crowley and Lowery (2000) and Jones and Mann (2003) all produce broadly similar reconstructions based on a simple compositing of proxy indicators over the Northern Hemisphere.

REPLIES TO AUTHOR COMMENTS:

General Comments

1. Authors assert that their calculations "included the NOAMER PC1 AND PC2".

The assertion is quite misleading. The authors included the PC1 and PC2 *that result from a completely different centering convention* from that used by MBH98. As we have shown in our revised reply, the PCs switch order when a different centering convention is used, and application of objective selection criteria indicates that 5, rather than 2, PCs should be retained if the MM04 centering convention is used (see our "Supplementary Information #1"). The original PC1 of MBH98 appears, instead, as PC4 using the MMO4 centering convention. Through incorrectly truncating the retained eigenvector basis set, MM04 eliminate the principal pattern of low-frequency variability in this dataset. Correct inclusion of 5 PCs, using MM04's own PCA centering convention, yields essentially the MBH98 reconstruction.

MM Replies to Reviewer #1 Comments:

1. We have definitively demonstrated that their assertion that the standardization/centering procedure used by us in calculating PCs of the North American ITRDB data has no influence at all on the main features of the MBH98 reconstruction.

2. We have never asserted that the Stahle/SWM data are key predictors for the NH mean reconstruction. We had listed them among a larger number of series that had indefensibly been censored from our network by MM03. The Stahle/SWM data are indeed key indicators in the reconstruction of ENSO, as a paper published in *Nature* a few months ago (Adams, J.B., Mann, M.E., Ammann, C.M., *Proxy Evidence for an El Nino-like Response to Volcanic Forcing*, *Nature*, 426, 274-278, 2003), for example, clearly demonstrates. MM misrepresent what we have shown in claiming that elimination of the Gaspé series does not influence the reconstruction. As we showed in our original reply, retaining *either* the Gaspé series OR the correct low-frequency PC pattern of the ITRDB data reproduces the essential features of the MBH98 reconstruction. It is only the elimination of BOTH the ITRDB data AND the Gaspé series that allows MM04 to produce their spurious early 15th century warm peak.

MM then make some completely incorrect statements about verification statistics. They cannot refute our assertion that their reconstruction dramatically fails verification. Instead, they now actually seek to reject the use of the commonly accepted measure of reconstructive skill verification (*RE*). To make matters worse, they attempt to do so based on an incorrect description of the assumptions behind the MBH98 methodology. They claim that the method requires that predictors have "a linear relationship to temperature". The method only assumes that the signal in the predictor (not the predictor itself, which contains both signal and noise), varies linearly with some large-scale pattern of temperature, not with local temperature itself (see top right column of page 780 of MBH98, paragraph beginning with "*Implicit in our approach*…."). For example, a coral indicator in the western tropical Pacific which records precipitation influences due to the El Niño/Southern Oscillation is a suitable proxy for ENSO-related sea surface temperature patterns. This issue is discussed both in MBH98 and numerous follow-up articles by the authors. The demonstration by MBH98 of Gaussian calibration residuals indicates that the linearity assumption of the MBH98 reconstruction is not violated.

Then authors then try to support the use of the R^2 statistic, even though it is well established that this does not provide an adequate measure of predictive reconstructive skill because it does not address changes in mean and variance outside the calibration interval (see e.g. the introductory statistics text by Wilks, 1995 referenced in our reply). Finally, they have calculated all verification statistics incorrectly because they have not applied the appropriate 19th century observational mask to the reconstructed patterns (see our Supplementary Information #3).

Finally, MM continue to claim that the key features of the MBH98 reconstruction are an artifact of the PCA centering convention used by MBH98 to represent the North American ITRDB data. We have demonstrated the falsehood of this claim in two entirely independent ways, (1) by using the same PCA convention as MM, but applying the correct selection rules which in this case indicates the retention of 5PC, producing a similar reconstruction to MBH98, and (2) by not representing the data by PCA at all but instead using all of the individual proxy series with equal weight in the analysis.

In short, there is nothing valid at all in MM04's reply to the reviewer's comment.

MM Replies to Reviewer #2 Comments:

1. As described above, MM04 effectively eliminated the primary pattern of low-frequency variability in the data (MBH98 PC#1) by changing the centering of the PCA, which reorders the basis set, and shifts this pattern to PC#4. Had they correctly applied objective criteria using their revised centering convention, they would have found that 5 PCs should be retained, which indeed retains the original MBH98 PC #1. The fact that we reproduce the key features of MBH98 both by using the correct selection criteria in conjunction w/ the MM04 centering convention and, moreover, without representing the indicators through PCA at all, obliterates their chief criticism.

2. As we have described in more detail above (and in our revised reply), the PCA algorithm has nothing to do with the shape of the reconstruction, because this is reproduced without any PCA representation, using all available proxy records with equal weight. It is only by eliminating both the North American ITRDB data (or, more specifically in the case of MM04, the primary pattern of low-frequency variance in the data) AND the "Gaspé" data, that the spurious features of MM04 are reproduced. Moreover, these features dramatically fail verification (see "3" below).

3. The reviewer appropriately stated that MM04's response should only be published "*should this validation be successful*" and noted that "*the RE statistics seems to me adequate in this context*".

MM04 have not met this standard. Their verification statistics are biased somewhat low for reasons detailed by us above and in "Supplementary Information #3". Nonetheless, the authors confirm that their reconstruction clearly fails verification, while the MBH98 reconstruction passes verification, as measured by the *RE* statistic. Having had their reconstruction fail verification MM04 instead try to promote the flawed use of an R^2 statistic. It is widely known that this is not an appropriate metric, for reasons detailed above and in our revised reply.

4. Our full replies to all of the reviewer's comments are provided above.

MM Replies to MBH Response:

1. Firstly, there are no 'quality defects' in the MBH98 proxy data set. There were some errors in the online supplementary information that have been corrected in a "Corrigendum" we have submitted to *Nature*. The various claims of errors in the dataset used by MM04, were a result of their use of an incorrect data file, a misunderstanding of the stepwise nature of our PCA representation of certain tree-ring networks, and the entirely misguided assumption that data made available to us by trusted colleagues but not posted in the public domain are somehow suspect. Like many of our paleoclimate colleagues, we often make use of high quality unpublished data. We nonetheless made all such data available on our own ftp site. The claim that the full version of the Jacoby Northern Treeline data we used (which are typically longer than the versions obtained on public websites by the authors) is 'obsolete' is completely unsupportable.

2. Once again, this entire line of argumentation by the authors has been nullified, because we have reproduced essentially the same MBH98 result using both the authors' own PCA convention (but retaining the correct number of PCs) and by using the individual proxy series, rather than any PCA representation of them. With regard to the issue of CO_2 fertilization, we have indeed, as MM04 note in their comment, dealt with this issue (Mann et al., 1999)**.** In that publication (five years ago!) we obtained a very similar reconstruction to MBH98, with no warm 15th century, after enforcing strong reduction of the upward growth trend of the late 19th and 20th centuries in the 1st PC of the ITRDB chronologies used there. As described there, that analysis was based on a careful diagnosis of the difference between the behavior exhibited by otherwise similarly behaving tree-ring chronologies in regions both likely and unlikely to exhibit any hypothesized CO_2 fertilization.

3. The claims by MM here have been entirely refuted above (see reply to Reviewer #1 comment #1).

4. The arguments here could not be more specious. As their primary criticism depends on what they claim is a skewed representation of certain ITRDB series by our PCA algorithm, we have demonstrated the complete falsehood of that claim by presenting an analysis in which all indicators are used with equal weight, and the MBH98 calibration approach, as it is designed to do, selects out that information in the network that provides the most skillful reconstruction of the instrumental PCs. So the first part of the argument is complete nonsense. The second

part of their argument is now discredited. We have shown that the same reconstruction is obtained using their PCA centering convention, as long as the correct number of PCs (5 PCs in that case) are retained (see our supplementary information #3).

5. The claims here have already been demonstrated, above, to be spurious and specious.

6. The extension by persistence of the first 3 years of the 'Gaspé' series was indeed performed to allow this indicator to be used in the AD 1400-1450 block, which was simply a matter of convenience since we chose, for simplicity, to perform our stepwise reconstruction in 50 year/100 year blocks rather than 1 year blocks. It would be quite unwise to eliminate an important indicator available over almost the entire 15th century (when the data become increasingly sparse) simply because it starts in AD 1404. Were we instead to perform the stepwise reconstruction in 1 year steps (a quite time consuming process indeed!), this indicator would appear in AD 1404, rather than AD 1400, and would have no influence on the main features of the reconstruction. As described in our reply, we have addressed this both by eliminating the 'Gaspé' series in the reconstruction back to AD 1400, and in performing a reconstruction that goes back only to AD 1404. So first, we have shown that the 'Gaspé'' series is not essential to reproduce the MBH98 reconstruction back to AD 1400. Secondly, we have shown that eliminating it would only slightly change the values of the MBH98 reconstruction for the years AD 1400-1403, a completely negligible effect.