We thank T.M. Melvin and J. Björklund for their thorough reviews. Comments are addressed below. Each comment by the reviewer is first recalled (in italics), then the corresponding authors replies are given.

T.M. Melvin (Referee) *t.m.melvin@uea.ac.uk*

The authors reprocess two MXD data sets derived from northern Fennoscandia which have been used in recently published climate reconstructions. They find distinct low- frequency differences between the two chronologies. They examine the effect of various standardisation options and show that differences in the standardisation methods used do not produce substantially different chronologies and so are not the cause of the differences observed. Generally this part of the work is good, addresses an important aspect of climate reconstruction, and advances the science of dendroclimatology.

Having found that the problem is not with tree-ring standardisation they do not manage to attribute another cause. They combine the two data sets and present a new reconstruction which, because of the unresolved problems that exist with these data, does not add to the current understanding of climate variability in Fennoscandia. The paper could be published as a test of standardisation (without the reconstruction). Alternatively the authors need more work to resolve the "problem" if they wish to produce a reconstruction.

Reply: We think that the new reconstruction is reasonable al least because we updated recent Esper's et al. 2012 reconstruction by:

1) adding more samples from Northern Sweden (and it already had some from the same region). This increased chronology replication;

2) applying multiple RCS (and correction). Previously single RCS was used and here we have showed that multiple RCS (and correction) is more suitable for these data.

Yet many other things could be done in future research (but irresolvable so far, see comments below), we think that it is already a significant step forward into understanding of climate variability of this region in the past. Our new reconstruction must be more precise then the previous one and also present separately calculated confidence limits that are crucial for its utilization in future studies.

Detailed Comments

***Abstract** – contains too many terms like RC1SFC, RC2SF and should be simplified so it can be read by a general audience with the finer details in the discussion and conclusions.*

Reply: The abstract will be rewrited.

P5662 L3 *– whole paragraph – in a paper that is attempting a detailed examination of the problems it is inappropriate to say "there are troubling inconsistencies with temperature amplitudes and timing of events" without explaining which problems have been overcome and which are outstanding.*

Reply: This part of the text will be modified by more explicitly stating which problems have been overcome and which are outstanding.

For Tornetrask there were inconsistencies for which the causes have been identified and corrected and Tornetrask reconstructions now provide roughly consistent long- timescale signals from both MXD and TRW. (Briffa (1992) found chronology values for MXD lower than those for TRW post1700 – the problem was corrected (Briffa 2011) by improvements to the use of RCS. Grudd (2008) found chronology values for MXD lower than those for TRW in the last 2 centuries - the problem was corrected see P5665, lines 24 to 29.

Reply: As mentioned above, this part of the text will be modified by more explicitly stating which problems have been overcome and which are outstanding.

Esper et al. demonstrated that the Finnish MXD and TRW chronologies are inconsistent but did not investigate the cause of the discrepancy.

Reply: This is a good point. We will mention this in the next version of the manuscript.

L8 *– Briffa 1992 reconstruction was >1000 years.*

Reply: We will correct this in the next version of the manuscript.

L14 *– "revealed a previously undiscovered millennial scale cooling trend" – as Esper et al.'s findings are not corroborated by any other tree-ring analyses (both MXD and TRW) this statement is misleading.*

Reply: This is a good point. We will mention this in the next version of the manuscript.

L18 *– Rather than "suggesting an overestimation of the Medieval warmth published by Grudd (2008)" – Melvin et al. clearly showed that the Grudd study contained systematic bias – and should not be considered as suitable to use as a climate reconstruction.*

Reply: We will mention this in the next version of the manuscript.

L20 *– Having drawn attention to the discrepancies among several reconstructions the authors then state the ".. unsatisfactory, if not worrisome .." nature of this situation but then cite Briffa et al. 1992 without acknowledging that this reconstruction is not demonstrably more in error (if indeed is is at all) than any of the earlier cited papers (note that the ad hoc correction applied to the MXD in this paper was substantially found to be reasonable see Briffa 2011).*

Reply: This is a good point. We will modify this part in the next version of the manuscript.

P5663 L1 *"Indeed, the flip side of this..." is unsuitable – I suggest "A problem associated with the ..." .*

Reply: We will follow this suggestion to modify the text.

L14 *"Yet, these biases do not fully explain the obtained differences between the Torneträsk (Melvin et al., 2013) and larger Fennoscandian MXD data (Esper et al., 2012).". This statement is not justified. Melvin et al. investigated and corrected for the bias caused by offset mean values of MXD measurements from different contexts while Esper et al. (2012) did not test for this bias. Despite Esper et al.'s claim that "We carried out a number of tests to the MXD network and noted the robustness of the long-term trends, .." the tests they performed using correlation Table S2 (removes the mean value of sample being compared), using curve-fitting methods Table S2 (removes mean of each series) and creating separate RCS chronologies for each group Figure S4 (sets the mean values of indices of each group to 1.0) did not test the robustness of the long term trend which is obtained from the varying mean value of tree index series over time.*

Reply: This is a good point. We will modify this part in the next version of the manuscript.

L16 *– "all the discussed studies used the same type of standardization method" – they used RCS but selected different options which might explain differences between the resulting chronologies.*

Reply: We will mention this in the next version of the manuscript.

P5664 L4 *"microdensitometry" – the authors have not examined the possibility of systematic bias in MXD measurements.*

Reply: In this study, we simply do not possess such information from experimental wood samples to be examined. It is clear that our study was to examine the existing MXD series.

L6 *– "Tornedalen" - neither Esper et al. nor Melvin et al. use the pre-1860 data because they have not been adjusted for the pre-Stevenson screen period (see Frank 2007 QSR).*

Reply: According to the original publication of this instrumental data set (Klingbjer and Moberg, 2003), the Stevenson screen was set up at Haparanda station no earlier than in 1924. We find no reference suggesting to omit the data prior to that date. Moreover, it is notable that the record we've used was corrected for this and other inhomogeneities (Klingbjer and Moberg, 2003, as cited in the text). In addition, our new study (Helama et al. 2013 [Journal of Geographical Sciences], the reference will be added to the text) did not show increasing inhomogeneities in the 1800s part of the same instrumental record, in comparison to the 1900s part of the record. For these reason, we find it reasonable to use the full record.

P5667 L8 *"In the original RCS technique presented by Briffa et al. (1992), the standardization curve is not fitted individually to each data of tree-ring series." - misleading, perhaps you meant "In the original RCS technique presented by Briffa et al. (1992), the same standardization curve is used to detrend each series of tree-ring measurements."*

Reply: We will modify this in the next version of the manuscript.

P5668 L22 *– Updated reference - Melvin and Briffa 2013 addresses signal-free RCS (2008 is SF curve fitting) see http://www.cru.uea.ac.uk/cru/papers/melvin2013dendrochronologia/.*

Reply: We will update this reference in the next version of the manuscript.

P5669 L21 *– why only the "first 100 years" of each tree. Using all years reduces the risk of inconsistencies due to suppressed early or late growth (e.g. the crossing of RCS curves of Fig 3a).*

Reply: If the hypothesis is (as it should be) that trees with high early growth die young, then this is supposed to be an issue in which the inclusion of old rings cannot improve the calculation as those trees do not even have the old rings (as they died). That's why including all years for the trees of unequal age is not completely correct.

L15-18 (and P5672 L12) *Replication differences – The TORN data are based on "mean tree" series where the FENN data have multiple cores for each tree (which artificially inflates EPS values and reduces error margins).*

Reply: We agree with this comment. New calculations are based on "mean tree" series for FENN data. For the new version of manuscript FENN dataset was processed again, all figures were re-plotted and corresponding statistics re-calculated.

L25 *– combining TORN and FENN MXD to produce the FULL dataset requires careful checking for site or sample type inhomogenuity (specifically offsets of mean values of the data series).*

Reply: In the context of the present data set, such inhomogeneity can be tested for any appearing difference between TORN and FENN data. When combining the datasets, multiple RCs were used, these effectively accounting for any potential difference in the mean of the two types of series to be standardized, in the process of transforming the MXD data into indices.

P5672 L24 *– "eliminate the biases arising from temporal distribution of well and poorly growing trees" – wrong. It will "reduce the 'modern sample bias' created by sampling living trees – which is much larger for TRW than MXD". One idea behind using two RCS curves is that the independent sub-samples can be shown to produce the same common signal – demonstrated clearly for the separate TORN and FENN signals shown in Fig 3b.*

Reply: We will modify this in the next version of the manuscript.

Table S2, S3 and S4 *– two digits is sufficient (3 is OK) for correlation purposes – any more than this should be deleted. Also your discussion of the values should ignore differences that are likely to insignificant.*

Reply: We agree with this comment. For the new version of manuscript new tables were made. We leaved 3 digits and marked significant correlations (based on "effective degrees of freedom" for smoothing splines).

Insignificant differences in Table S1. We calculated significance limits for correlation coefficients (taking into account number of effective degrees of freedom for smoothed data). They are (p=0.05): 0.54 for 300yr splines, 0.44 for 200yr splines, 0.34 for 100yr splines and 0.25 for 50yr splines. Yet, the differences in performance of different methods could be small and insignificant (statistically), they show which method perform better than others for these particular data (in sense of common variations on different timescales). From the Table S1 we can see that:

1) In all cases single-RCS perform better for TORN data and multi-RCS perform better for FENN data. For example, for 200yr and 300yr smoothed data multi-RCS with TORN gives insignificant correlations (p=0.05) and single-RCS with TORN gives significant values.

2) In most cases Correction procedure improves performance of RCS methods. This improvement may not be large but it remains stable for the experiments carried out. For example, for 300yr smoothed data difference between correlation coefficients for TORN-RC1SF and FENN-RC2SF without Correction (r=0.537) and TORN-RC1SFC and FENN-RC2SFC with Correction (r=0.575) is rather small. It can be seen however that the first coefficient will not be significant at p=0.05 but the second will be significant.

P 5674 L4 *Replace "an increasing toward" with "a trend increasing towards".*

Reply: We will modify this in the next version of the manuscript.

P5675 L4-28 *There seems to be little justification for joining measurement series and producing a combined chronology when there are such large unexplained differences outside of the calibration period e.g. the early medieval warmth (8th and 9th centuries) is not consistent.*

Reply: Actually, no proxy is free of non-climatic variations, not even MXD that is considered as one of the most high-quality proxies among them all, as indeed demonstrated through our study, and this is one of the important points we make in the paper. Recently, McCarrol et al. (2013) detailed the benefits of combining proxies that correlate but also show deviations typical to proxy data, in their setting similar to our study. In the manuscript, we will follow the recommendations by McCarrol et al. (2013) in this context. Moreover, the list of benefits as detailed by McCarrol et al. (2013) will be included in the text in this context.

P5677 L6 *– "it may be generally comfortable to rely on massive sample replication in dendrochronology (Büntgen et al., 2012)" – not a useful statement as increasing sample replication does not remove the systematic end effects, caused by tree aging processes, which detrending is designed to remove.*

Reply: This was not our statement but by earlier work (Büntgen et al., 2012). Is the referee suggesting that the earlier study presented misleading statements? If yes, it is possible for us not to state in the manuscript that large sample size may generally thought to produce more robust chronologies.

P5678 L25 *– "Tree-ring standardization is generally understood as an obstacle for deriving the low frequency climate information from tree-ring ..." – wrong. "Tree-ring standardization is necessary for the isolation of low frequency climate information from tree-ring ...".*

Reply: We can modify the text accordingly.

P5679 L23 *"As a consequence, tree-rings have even become notoriously poor indicators of low-frequency climate variability for wider readership (e.g. Broecker, 2001)." – only for those not familiar with background and techniques. "Tree ring chronologies are one of the few proxies for which sensible estimates of their skill at indicating 'longtimescale' variance can be calculated.*

Reply: We can modify the text accordingly.

L28 *- "This is how the correction procedure significantly reduces the "sample error" due to uneven age distribution over the AD 1950–1990 period (see Fig. 7)." – the arguments here are not convincing – "is the error associated with an age distribution problem", "was the reduction in error significant", "why not RC2SFC" and "does Fig 7 show errors". It seems likely that a*

much larger data set is needed to resolve these questions and this statement needs to be qualified with words such as might or likely.

Reply: We can modify the text accordingly.

P5680 L7-17 *Multiple RCS tends to reduce (two-curve RCS roughly halves the amplitude of the variance derived solely from the mean values of series of tree indices) longer timescale variance and improves the shorter timescale variance. This would reduce the effect of any site or sample type variation of the mean value of MXD measurements provided the differences are equally distributed over time. Correlation over the most recent century is not likely to be a useful test of the effect of multiple v single RCS. Two-curve RCS has less error and so reduced need for correction. TORN has ample trees for two-curve RCS – an examination of uncertainty error #1, of supplementary 4 is likely to show this.*

Reply: This comment is not easy to follow as there is no actual recommendation to be followed. We will however consider the issue.

L19-28 *– can be tested by plotting the mean values of young/old trees or young/old rings on the same graph and looking for systematic differences (not necessary for this paper).*

Reply: Thank you for this advice, we will consider the issue.

P5685 L3 *"Thus our new reconstruction can be used as the source of information about year-to-year, as well as centennial and longer variations of summer temperature in Northern Fennoscandia for the Common Era." – need a warning about the unexplained discrepancy.*

Reply: We can modify the text accordingly.

L5 *"Nevertheless, the use of other proxies that can reproduce low-frequency past temperature variations is highly preferable in every paleoclimatic study." – this statement needs a qualification about the error estimates – if there is a proxy that can reproduce low-frequency accurately they should mention it if there is not then this statement should be removed/qualified appropriately.*

Reply: In this context, we will be citing our previous work (Helama et al. 2010; Quaternary Science Reviews) where the tree-ring and pollen proxies were shown to share their low-frequency variations in the study region.

Supplementary *1 Sorry cannot read Russian so this is my only description of the correction procedure. Your definition needs to clearly distinguish between SF-series, SF- curves by saying what has been removed e.g. SF-measurement series contain aging curve and noise but climate*

signal removed or SF-indices where aging curve and climate both removed consist of noise. It seems that the correction is producing an improvement by reducing the low-frequency noise (an informed version of robust mean). Because two-curve RCS removes half the variance of the long-timescale signal the correction will have less effect for multiple RCS.

Reply: Utilization of the terms (SF-series, SF- curves) will be unified. Additional description of the correction procedure (with figures) will be added to supplementary material (because editors advised to move technical details form the main article).

J. Björklund (Referee) *jesper.bjorklund@gvc.gu.se*

The authors reanalyze the two longest existing MXD records from Fennoscandia. They assume that they should have similar signals in all frequencies due to that they have similar forcing agent (summer temperature). In addition to finding that the two records covary most of the time they also find that they diverge significantly from time to time. They address the problem by applying different standardization-techniques but fail to conclude that that is the root of the problem. However, it is very interesting to note how different techniques perform together, but this section could be expanded to include other techniques that are mentioned in the text. Further this analysis would greatly benefit from addition of more chronologies, for instance TRW from the same dataset and MXD and TRW from other datasets, they do not have to be 1500 years. This would add credibility to the otherwise miniscule differences in correlations, with regard to standardization configuration skill. Further it is a little bit confusing to compare how high the correlation is between two different standardization techniques, I do not find that information useful. The attempt of making a new reconstruction of evidently diverging chronologies is advised against since one of the chronologies could hypothetically be more in error than the other. Additional chronologies might give answers to this. The paper should after proper revisions be published, but would benefit a great deal if some of the discussion-points about potential sources were analytically addressed instead of only discussed, see detailed comments.

Reply:

The potential sources of discrepancies between the studied datasets that are discussed are not analytically addressed, because we simply do not possess needed information from experimental wood samples to be examined. Nevertheless our discussion raises the problems that were never addressed before concerning these datasets. Future studies should take these issues into account.

Point by point responses to most of the issues summarized in this general comment are given as replies to the detailed comments below.

Detailed Comments

P5661 L21 *Consider adding Grudd (2008) and McCarrol et al (2013)*

Reply: We will modify the text accordingly.

P5661 L27 *Change "in the most recent" to "In a recent" or specify that you mean MXD from torneträskmaterial*

Reply: We will modify the text accordingly.

P5662 L4 *change "indications" to "records"*

Reply: We will modify the text accordingly.

P5662 L6-L10 *Grudd 2008 has on several occasions been shown to be incorrect (e.g. Melvin et al. 2013, Björklund et al. 2013) and should not be used in this comparison. Again P16-P18.*

Reply: This study should be cited here to emphasize the process that has led to the current situation.

P5662 L2-L24 *This paragraph would benefit from studying McCarrol et al. (2013). The discussion, as it is now, only relates to two different datasets treated in different ways.*

Reply: We will be including the citations to that reference, too.

P5662 L27-L29 *Perhaps also sampling biases? Some trees may be preserved better than others. The preservation conditions are likely different for a dry dead-tree chronology than from a wet subfossil chronology. Also, the living tree material was sampled with different strategies for FENN and TORN.*

Reply: We will modify the text accordingly.

P5663 L14-L16 *Agree with Referee T. M*

Reply: See reply to T.M.Melvin's comment (P5663 L14-L16) above.

P5663 L22-L26 *Since the Matskovsky 2011 paper is in Russian and most of the scientific contribution of this paper is to apply this method, perhaps it deserves a thorough description in the main article.*

Reply: We will add descriptive figures and some more description to supplementary (because editors advised to move technical details form the main article).

P5663 L26-L29 *McCarrol et al. 2013 uses the same RCS method to standardize the MXD chronologies included in that paper. Forfjorddalen Torneträsk and Laanila.*

Reply: We will modify the text accordingly.

P5664 L3-L5 *How did you deal with effects from microdensitometry? You are addressing effects of standardizationmethods in this paper?*

Reply: In this study, we simply do not possess such information from experimental wood samples to be examined. It is clear that our study was to examine the existing MXD series.

P5664 L6 *I do not see a problem using Tornedalen temperatures as long as the analysis is made with high-pass filtered data. But low-frequency correlations, where prestevenson data is included, should be avoided.*

Reply: See reply to T.M.Melvin's comment (P5664 L6) above.

P5664 L7-8 *I agree with T.M that making a reconstruction is premature when the problems identified between the chronologies are not resolved. Instead I suggest to also include TRW and perhaps also more chronologies, does not have to be >1000 years, to have a larger sample when comparing skill of methods.*

Reply: See reply to T.M.Melvin's comments (General Comments and P5675 L4-28) above.

When choosing data for the analysis we selected two longest and best replicated available datasets of the best summer temperature proxy for the studied region (which is MXD). Other chronologies are either unavailable and/or cover shorter period and/or constructed for proxies with weaker correlation with temperature (like TRW) and/or compared in other studies (Björklund et al., 2013; McCarol et al., 2013) .

P5664 Materials and methods*. Previously the authors described that problems do exist between the TORN and FENN. Then they are assuming that signals must be the same or very similar because of the homogenous nature of the dominating forcing agent (summertemperature) upon them. They go on to try to solve these problems with various standardization-techniques, and fail. Could the assumption perhaps be in error considering the heterogeneity of the datasets, it is clear that the elevation, the size, preservation medium, number of subsites, geographical distribution of subsites of the sample-sites vary considerably. The amount of low-frequency variation retrieved in an RCS chronology would arguably be larger in a more homogenous sample. Would there be any way to investigate this possibility by separating the subsites from eachother?*

Reply: As previously alluded to, we have already demonstrated this (that the amount of low-frequency variation retrieved in an RCS chronology would arguably be larger in a more homogenous sample) by showing that in all cases single-RCS perform better for TORN data and multi-RCS perform better for FENN data. We do not however understand the statement that the assumption is in error considering the heterogeneity of the datasets elevation with the size, preservation medium, number of subsites, geographical distribution of subsites affecting MXD, as we did indeed discuss these issues in the text in the detail we were able to. If, instead, the assumption referred to by the referee is that the standardization could cause the differences, then he is actually referring to the hypothesis that was to be tested in this paper, the hypothesis that was justified to be tested, as detailed in the introduction.

P5667 L23 *Why ratios when convention says residuals for MXD data?*

Reply: Ratios were used here as they were used in the original references of MXD data sets.

P5668 Correction procedure. *This is a central part of the paper and should probably be detailed more extensively in the manuscript, maybe even with figures. This because large part of the readership is not Russian-speaking and because it seems as this methodology performs the best in general and thus is the most valuable finding of the paper. Further there are several other methodologies that are left out from comparison that the author mentions, I particularly think of Nicault et al. (2010).*

Reply: We will add descriptive figures and some more description to supplementary (because editor advised to move technical details form the main article).

P5670 Design of experiments. *The main analytic tool used to evaluate standardization performance is pearson correlation. This is done with untreated chronologies as well as for smoothed chronologies. I would like to see p-values or p = 0.05 significance levels for all correlations, especially for the smoothed data, because they are associated with substantial loss of degrees of freedom. In the discussion where, one procedure is deemed to perform better than another, it would also be nice to analyse if the performance is significantly higher/lower than the other. Without this, one procedures performance over another could be purely by chance, because one procedure consequently must yield higher correlations than another if chronologies are slightly changed.*

Reply: See reply to T.M.Melvin's comment (Table S2, S3 and S4) above.

P5670 Reconstruction. *See comment above.*

Reply: See reply to T.M.Melvin's comments (General Comments and P5675 L4-28) above.

P5673 L8-L11 *See comment above about correlations. For example what is the pvalue for the 300 yr smoothing where r=0.44, when adjusted for loss of degrees of freedom, see below. Dawdy, D.R., and Matalas, N.C., 1964, Statistical and probability analysis of hydrologic data, part III: Analysis of variance, covariance and time series, in Ven Te Chow, ed., Handbook of applied hydrology, a compendium of water-resources technology: New York, McGraw-Hill Book Company, p. 8.68-8.90*

Reply: See reply to T.M.Melvin's comment (Table S2, S3 and S4) above.

P5674 L13 *I do not see the need for the quasi-periodicity analyse. It is not associated with the comparison of standardization techniques. But it could perhaps be used early in the paper to establish the assumption that the two dataset are very similar or that they differ in some frequencies.*

Reply: Quasi-periodicity analysis is associated both with comparison of standardization techniques and with datasets themselves. For example if we will use deterministic standardization, we will lose most of low-frequency and it will be seen from quasi-periodicity analysis.

P5676 Comparison of datasets. *The authors discuss potential sources of the differences between TORN and FENN. This should be more the motivation for the paper, and to construct experimental design to try to investigate these, because it is obviously not only standardization that can improve the coherence between the two.*

Reply: While this comment could be acknowledged, the present dataset - that was the premise of the current study - simply do not allow such experiments to be derived. Actually, the need for such experiments becomes obvious after reading our study, as demonstrated by the referee himself by making this comment. And, indeed, this is in line with our arguments and what we have stated in the discussion/conclusions of the paper.

P5694 F3 *Chronologies produced with different standardization-techniques. To me it looks like the different standardization techniques have larger effect on the FENN material and that TORN seem inert to choice. What can be learned from that? Also the change caused by standardization techniques is marginal to the differences between the datasets. Clearly this must be further addressed before a reconstruction can be made, see comment above. Alternatively, more chronologies must be added to arrive in higher certainty around the mean. It can hypothetically be the case that one of them is more correct than the other, and then a new composite reconstruction would be worse than then previous publications of them separately.*

Reply: Actually, assessing this issue analytically and not visually, TORN is changing slightly more. For TORN mean difference between differently standardized chronologies is 0.1076 of STD and for FENN - 0.0918 of STD. So the values are rather similar. In our consideration, the referee statements that "Alternatively, more chronologies must be added to arrive in higher

certainty around the mean." and "It can hypothetically be the case that one of them is more correct than the other, and then a new composite reconstruction would be worse than then previous publications of them separately." are indeed contradicting each others. Please also see our comment to the T.M.Melvin's comment P5675 L4-28.

P5696 F5. *Why are spectrums for differently standardized chronologies shown?*

Reply: We showed the spectra for the "nearest" chronologies, which are TORN-RC1SFC and FENN-RC2SFC.

P5697 F6 *Figure must be redesigned, now it is impossible to see legends and differences between chronologies. Why not use residuals between summer temperatures and chronology-configuration? A trend analysis of the residuals can be performed and see if some technique produce less or more trend?*

Reply: We now see that this figure may confuse the reader and therefore we are ready to omit the figure as its information wasn't crucially important for interpretation of the paper.

P5698 F7 *Tornedalen record seems odd in the beginning and in the context of the discussion with Stevenson-screens, should perhaps be omitted from analysis?*

Reply: See reply to T.M.Melvin's comment (P5664 L6) above.