Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling

Tardif et al.

Authors Responses to Reviewers

In the following, comments from the referees are shown in bold, with our responses shown in plain text. A marked-up version of the revised manuscript, showing the changes from the originally submitted version, is also included. Page and line numbers refer to the revised manuscript.

Anonymous Reviewer 1

We thank the referee for thorough and insightful comments on the manuscript. The comments have challenged us to take a more comprehensive look at the numerous reconstruction experiments we performed, and as a result, we have gained a more complete perspective on the results.

The paleoclimatic discussion needs to go more into depth and the authors need to be more critical about their own results. The authors compare their new data set to the previous version (Hakim et al. 2016) and conclude this new version would be an overall improvement. This conclusion is based on validation statistics in the 20th century. However, the new data set lost all multi-decadal to centennial variability in the global mean temperature and does not show a warmer medieval period nor a cooler "little ice age" anymore. The authors do not discuss this issue at all. The paper suggests that this new reanalysis version would present the more likely global mean temperature evolution of the past 2000 years although it is in contrast to what most other reconstructions and paleodata records suggest.

We agree with the referee's suggestion that results should be framed in the context of other reconstructions, and a discussion focused on the long-term perspective (not limited to the 20th century) of the updated reanalysis be included in the manuscript. To address this issue, a revised section 3 includes a more complete discussion of our reconstruction results, including a comparison of LMR reconstructions of Northern Hemisphere (NH) temperatures with other NH reconstructions found in the IPCC AR4 and AR5 reports (see bottom panel of Figure 2, section 3). A greater perspective on the long-term (i.e. across the Common Era) evolution of temperature in our reconstructions is gained. The main takeaways from this comparison can be found in the fourth paragraph of the revised section 3, page 10 line 23 to page 11 line 7.

We do acknowledge the updated reanalysis in our originally submitted manuscript was characterized by a significant loss of variability compared to our prototype reanalysis. We have revisited some of the choices made in the configuration of our system and conducted additional experiments to identify the source of this loss of variability. We have determined that it is preferable to eliminate the large number of unscreened tree-ring records from the Breitenmoser et al (2014) collection, as included in Anderson et al (2019) from our reanalysis. This is despite an increase in skill in temperature and hydroclimate reconstructions when this dataset is included, as determined from observational data. We believe this underlines issues with our estimates of observation error variance (i.e. the R_k terms in equations 4 in the manuscript) specifically for these records. We have found that the inclusion of these tree-ring chronologies is largely resposible for the warm bias which characterizes reconstructed temperatures in our previous updated reanalysis during the 1600–1700CE and 1810–1920CE periods of the Little Ice Age (LIA). Therefore, we now present a new updated reanalysis using a revised configuration of the assimilated proxies, limited to the proxies from the PAGES 2k Consortium (2017) collection. With this configuration, we find a greater level of temperature variability is recovered, as well as a much improved agreement between the LMR Northern Hemisphere temperature to the other reconstructions during the LIA. We note however that the absence of a notable warm medieval period continues to characterize our reanalysis. An analysis of results from a large number of reconstruction experiments has allowed us to conclude that colder temperatures during the medieval period, compared to the prototype LMR, are related to the change from the proxy dataset of PAGES 2k Consortium (2013) to the more recent PAGES 2k Consortium (2017) collection. The global temperature composites presented in PAGES 2k Consortium (2017) shows that a distinctly warmer medieval period isn't a prominent feature of the new collection, and is not the result of other updates to our data assimilation system. As this dataset reflects the community's most recent and rigorous identification of proxy records suitable for temperature reconstructions, we believe that the lack of a "classic" medieval warm period in our updated reconstructions of global mean surface temperature and NH-mean temperature should not necessarily be considered an outstanding shortcoming of our updated reanalysis. Discussions on the issues outlined above are included in section 3, page 10 lines 3 to 12, and in section 4.3, entirely revised due to the re-focus of the updated reanalysis.

The loss of low-frequency variability is most likely a consequence of the proxy data sets used, because this is the major change in the new version. Many of the tree-ring chronologies in Breitenmoser et al. 2014 are not climate sensitive at all or moisture sensitive. As precipitation does not show any low-frequency variability in contrast to temperature, it is a logical consequence that using covariance information from moisture sensitive trees to correct temperature data leads to a loss of low frequency variability.

The referee raises good points, confirmed following a careful review of our results, including those from additional test reconstructions (see response to previous comment). We note however that the data assimilation (DA) framework is different than other reconstruction methods. Key elements are the forward models (the proxy system models, or PSMs) used to estimate proxy observations from a model prior, and the observation error variance assigned to the proxies, i.e. the R_k terms in equations 4 in the manuscript. Within a DA framework, there are no fundamental reasons why the inclusion of records with sensitivities other than temperature would lead to a deterioration to reconstructions of temperature, provided that PSMs properly account for the proxy sensitivities and the associated observation error variances representative are properly specified. However, it is acknowledged that these conditions may not easily achieved. Initial LMR results (i.e the prototype) suggested that our approach has the ability to delineate proxy records with weaker sensitivity to climate by assigning relatively larger observation error variances (R_k), resulting in such records only weakly influencing the reanalysis results even though they are assimilated. Also, the main motivation for developing the bilinear approach to model tree-ring width (TRW) data as been to gain an ability to seamlessly handle the more complex sensitivities to temperature and/or moisture of these chronologies. With this approach, in principle, TRW records can be assimilated without having to make a binary decision whether each record is dominantly sensitive to temperature or moisture, or having to screen records out a priori. However, we acknowledge that relying on simple regression-based PSMs opens up the process to the influence of spurious correlations between noisy data. With a large number of proxies considered, some records will invariably be characterized by somewhat overestimated confidence, i.e. too-small error variance, and therefore overly weighted in the update. Closer examination of additional test reconstructions strongly suggest that these issues are responsible for the loss of variability observed when the entire set of Breitenmoser TRW records are assimilated. We conclude that the possible inclusion of proxies from this dataset, including an improved characterization of observation errors, requires further attention. Hence, the revised manuscript presents, throughout the paper, an updated LMR using a revised configuration, where assimilated proxies are limited to the PAGES 2k Consortium (2017) collection.

A second reason may be the use of proxy data with dating uncertainties, such as ice cores, in an annual reconstruction. These proxies probably do not have age errors in the 20th century validation period but become just noise if they have an age offset of one or a few years further back in the past. The authors just conclude that using moisture sensitive data leads to improved reconstruction skill, although this is only true in the 20th century validation period but not in the pre-instrumental period, most user of this data set will be interested in.

This is an interesting suggestion. We have tested this hypothesis by performing an additional experiment in which all ice cores records were withheld from assimilation. The results do not support the hypothesis put forward by the referee however. GMST and NH-mean temperatures exhibit similar multi-decadal variability compared to reconstructions which include ice core information. The main difference consists of a modified long-term trend, showing a flatter temporal evolution over most of the Common Era prior to the 20th century warming, worsening the agreement with the other reconstructions. The primary role of ice core proxy data seems to rather anchor the millennial-scale cooling characterizing the pre-industrial era. However, we reserve a discussion on this specific topic for a future publication which is expected to focus on the role of various proxies in the reanalysis.

In the current version, the global mean temperature evolution is the reappearance of the famous “hockey stick” in climate science. After all the discussion, the hockey stick was rising 20 years ago, I would not publish this as a state-of-the-art temperature reconstruction, especially not without a discussion and not if it is an artefact of unscreened input data.

See responses to the comments above. In the light of results outlined above, the revised manuscript presents an alternate updated LMR, showing a greater level of variability, including at lower frequencies, in better agreement with reconstructions from other authors as shown in panel (c) in a revised Figure 2 and discussed in the revised section 3, page 10 lines 3 to 12 and page 10 line 23 t page 11, line 7.

I see two options, the first would involve minor revisions and the second major revisions: 1. It must be stated prominently (already in the abstract) that this reanalysis should not be used or considered to have the correct multi-decadal and centennial variability and that the global mean time series over the last 2000 year potentially has serious issues. The discussion needs to include all problems of data set, too and ideas how to overcome them in the future. In general, the paper should be put more into a context of methodological improvements to achieve better products in the future instead of claiming this would be nearly the prefect reanalysis for the past 2000 years. 2. A proper screening of the data needs to be introduced that prohibits the assimilation of non-climatic information, which has just spurious correction with observations in the short window of overlapping instrumental and proxy data (a minimum of 25 data points has been used in this study, page 5, line 2). These records will have little weight in the assimilation procedure due to large residual variance, but hundreds of little errors probably produce significant noise. Probably, precipitation limited proxies and proxies with age errors have to be removed or treated specifically, too. These are just some ideas and it will need many improvements and new experiments to find and solve the problem.

We appreciate these comments and suggestions. As outlined in our response to previous comments, we have chosen to follow a path inspired by the second suggestion. A complete review of results has been undertaken and new reconstructions experiments were carried out. The newly gained perspective has led us to consider an alternate reanalysis configuration which addresses the issues identified by the referee. Suggestions by the referee were carefully considered and integrated in the design of our latest experiments.

However, we wish to underline our belief that an approach that seeks to simply remove the information from precipitation limited proxies, as suggested by the referee, is not the preferred framework in which to seek improvements in reconstructions. Rather, improved forward models (PSMs), describing more accurately the relationships between climate variables and proxies, in addition to improved characterization of observation errors, are the key aspects where improvements can be achieved. We believe that this is reflected in part in our results showing improved performance with the reanalysis using the seasonal bivariate temperature/precipitation PSMs on the screened PAGES2k TRW records.

A difficulty in the review process is that the input data has not been published, yet. Hence, it is not possible to properly judge the input data base. However, it appears to be basically the Breitenmoser et al. 2014 data set with a few coral and ice core records added. Why do you not simply refer to this first publication and give citations for the additional records or wait until the Anderson et al. paper is published?

The manuscript is now available online at:
https://datascience.codata.org/article/10.5334/dsj-2019-002/

In general, the decrease of skill further back in time is not discussed sufficiently.

This is a great comment. We address this issue using our framework enabling an assessment of reanalysis performance in proxy space. The scope of the verification of reconstruction results presented in the revised paper has been expanded by including results from verification performed in proxy space on independent (withheld) proxies covering both calibration and pre-calibration periods. Results that were originally included in the supplementary material have been revised and moved to the main body of the paper. These results are found in section 4.1, Table 4, to complement the evaluation of the various PSM configurations considered, confirming with increased confidence the choices made in configuring the updated reanalysis system. Results from proxy–space verification are also included in section 4.3, Table 6, to gain an additional perspective on the accuracy of reconstructions performed with the addition of proxy records from the Anderson et al (2019) collection. The discussions on the proxy verification results are found on page 14, lines 1 to 7, and on page 16 lines 3 to 16.

It should also be discussed why forcings are not important and what the consequences of unforced simulations ensembles are for the final product, especially further back in the past when the proxy network is sparse.

We are not sure how the referee has come to believe that forcings are not important. In fact the model simulations from which we draw prior information do include forcings, such as pre-industrial greenhouse gas and aerosol variability, including the effects of volcanic eruptions. This point is mentioned in the manuscript. However, to bring the context into greater focus, we have found that an important characteristic of prior information within our DA framework is the amount of variance characterizing the simulations. We have found that the greater variance generally characterizing the "Last Millennium"-type simulations (which include the forcings listed above) provide for more accurate reconstructions, compared to using simulations performed without the influence of external forcings (as in the "pre-industrial" or piControl CMIP5 protocol). We also have generally refrained from using simulations which cover the 20th century warming to dispel the notion that we are "cooking the books" when reconstructing temperature trends. In our framework, temporal information (trends) come entirely from weighted information from the proxies.

I suggest to evaluate the spatial skill of the reanalysis in the 20th century but with the spatial proxy network at multiple time slices, e.g. 0 AD, 500 AD, 1000 AD, 1500 AD.

This is a great suggestion. We have generated proxy verification results over distinct periods of the Common Era: 1–499, 500–999, 1000–1499, 1500–1879, 1880-2000 (calibration period) in Table 2 and discuss the results from the perspective of reanalysis accuracy across time in a revised section 3. Even though the verification is not performed on an identical set of proxies we also compare the verification statistics from the prototype reanalysis, to further highlight the enhanced performance of the updated LMR. Additions to the text are found on Page 12, lines 16 to 22.

Additionally, not using forced simulation offers the potential to use them in the validation procedure. It could be checked if temporal and spatial patterns of known past events or periods are well represented in the reanalysis, e.g. spatial moisture distribution after eruptions (Iles and Hegerl, 2015).

This is also a good suggestion, however we believe that such efforts are outside the scope of the current work. The suggestion will be considered in future efforts.

Finally, it would be interesting to see a map of the regression residuals to get an idea how many paleodata records have significant influence in the assimilation procedure and which are basically ignored because they have no climate information.

We agree, however a concise presentation of this is challenging, due to the varied nature of the proxy data. We believe this would be addressed in a more informative way by a formal proxy impact study, which is intended to be the subject of another paper.

Additionally, I would like to know how many records in the PAGES2Kv2 data base have expert information on seasonality? I would be interested to read how well the expert-based seasonality in the PAGES data base agrees with the objective assessment in this study. Probably, the experts did a similar search for highest correlation, maybe just including more possible combinations of growing season months.

For the 2017 publication, the PAGES 2k consortium requested that each data certifier assess the seasonality of the temperature response and report its basis. In the LiPD format (McKay and EmileGeay, 2016), this information is encoded in the `climateInterpretation_seasonality` and `climateInterpretation_basis` metadata properties. All records in the database include seasonality information, either as letters (JJA) or numbers ([6 7 8]) indexing calendar months. When the basis is reported, it is either from "first principles" (e.g. trees are known to grow in local summer, which in most cases is synonymous with June July August), or from a search for the highest correlation. When the basis is not reported, the reader is referred to the publication documenting each record, which in most cases uses a mix of first principles and search for highest correlation, similar to the approach used in this paper. A key difference between the expert assessment of seasonality and the one done in our paper lies in the choice of target datasets. Studies focusing on individual series tend to be more careful about selecting an instrumental dataset appropriate for calibration (e.g. local GHCN station, rather than GISTEMP grid box average). These choices of target datasets are likely contributing to differences in the seasonal window determined via this process.

The choice of calibration datasets in LMR is driven by the need to uniformly process a large number of globally distributed records, hence the more general, likely not optimal, selection as is possible when one has to consider a single or few records.

We also wish to point out that the information requested by the referee on the differences between the expert-based seasonality in the PAGES data base and objective assessment in this study is already provided in the supplemental information accompanying the main manuscript (Figure S1).

I was surprised to read that the authors use an extended fall period (JJASON)? Is there any reference for trees which are limited by climatic conditions in these autumn months.

Consideration of this period has initially been motivated following D'Arrigo et al (2005), along with the fact that some seasonal responses found in the expert-derived PAGES2k metadata extend to fall months, as suggested by the data shown in Figure S1a in the supplemental material accompanying our submission. It is interesting to note however that the objectively-derived seasonal responses determined by the approach described in our paper leads to less emphasis on those fall months compared to the PAGES2k expert data.

It would be favorable to store the data at a world data center and not at a personal homepage.

We completely agree with the referee on this point. The LMR team has engaged interactions with the project's sponsor (NOAA) to identify a suitable storage location and access point, but these have yet to be identified, hence the reference to a personal homepage at this point.

Technical corrections

Abstract: skill score increase in percent is misleading. It is easy to have a large relative increase if scores were very low in the comparison data set, e.g. in Z500 where CE improves from very negative to less negative the increase in percent is large but the skill is still negative!

The point made by the referee is well taken. However, our emphasis has been about quantifying differences with respect to our main benchmark, the LMR prototype, to highlight improvements. Therefore we remain convinced that the formulation used is appropriate. The fact that some skill scores remain characterized by negative values is not hidden and becomes quite clear in the core of the text.

Abstract: be more precise what is meant with "ensemble characteristics".

We have eliminated the sentence containing this expression to streamline the abstract. We were referring to the preferred outcome of maintaining good correspondence between ensemble-mean errors and ensemble variance (i.e. uncertainty), a concept referred to as ensemble calibration in the data assimilation literature. We now more clearly define this in the last paragraph of Section 3 in the revised paper, page 12 lines 7 to 16.

Introduction, Line 17: apart from paleoclimate with annual observations, the forecast model is a third important component (this is even written in the Methods section).

We agree that this statement could be interpreted to mean that the prior data has less importance in the DA context. A revised statement, more accurately conveying the importance of the various components, is included at the beginning of the second paragraph of the Introduction, Page 2 lines 12 to 15.

Methods, Page 7, line 6ff: I do not understand why the calibration is done with a different gridded data set than the validation. Both data sets a based on largely overlapping instrumental observations and therefore clearly not independent.

The use of GISTEMP as the calibration dataset has largely been motivated by the fact that a larger number of proxy records could be calibrated (larger number of records with sufficient overlap with valid calibration data) compared to other datasets. Therefore, a larger number of proxies may participate to the reanalysis. We in fact perform validation using all datasets at our disposition, including the calibration dataset. Skill metrics show small differences among the various results, but remain in general agreement. Here we have chosen the Berkeley Earth dataset because it provides a greater spatial coverage, comparable to the calibration dataset, therefore providing a larger sample of spatial verification results.

Methods, Page 4 line 14: How many ensemble members?

The information is already provided in the manuscript, in Section 2.3, where details of the configuration are listed.

Methods, Page 4 line 14 it is written that Hakim et al. 2016 worked "with the same randomly drawn ensemble members used for every year in the reconstruction", whereas on page page 5, line 28 it is written for this study: "each using a different randomly chosen 100member ensemble". Are both studies consistent and is each year build on different randomly chosen ensemble members?

Both statements are consistent, in that the first statement describes how the ensemble members for a given reanalysis realization are used to populate prior information in the time domain, whereas the second statement points out that different sets of states are used to populate the prior ensemble for different Monte-Carlo realizations of reconstructions. A revised statement clarifies this in the last paragraph of section 2.1, page 5, line 10, with the addition of "for every year in the reconstruction of a given reanalysis realization".

Page 5, line 1: "Only records for which a PSM can be established are shown ...". What do you mean by "shown"? There is no reference to a figure. Do you want to say that only records meeting these criteria are assimilated?

We wish to point out that a reference to Figure 1 is included in the previous sentence. The text of the last paragraph of section 2.2 has been modified to ensure the reader is not confused about what the reference is. See page 5, lines 31-32.

Methods, Page 5, line 8: Breitenmoser et al. 2014 is not screened for any climate sensitivity.

This text has been moved to section 4.3, and a clearer statement about the absence of screening for climate sensitivity has been added, page 14, line 30.

Methods, Page 6, Proxy modeling: It should be repeated here over which period the regression coefficients are calculated. As many data points as available in the overlapping period with instrumental data but minimum 25 pairs of x and y?

This information is now briefly clarified in the revised manuscript. We have added the following statement: “a threshold of at least 25 overlapping data is imposed”, page 7 lines 4-5.

Methods, Page 7, line 30: Is there a reference that any tree-ring proxy responds to an extended fall period (JJASON)? The given references point to common growing seasons from May to August in the northern hemisphere. Why are not all combinations of growing season length tested and the optimum is chosen? In the PAGES data base there are also various different length of growing seasons defined.

For the first part of the question, see the reply provided earlier in this document. About the second part, we concede that our objective seasonal PSM calibration methodology is a compromise between comprehensiveness and efficiency in the calculations, performed over more than 2000 proxy records. Nonetheless, we believe that while perhaps non-optimal, the resulting characterization provides realistic results, as suggested by the fact that more accurate reconstructions are obtained when this information is used in forward modeling tree-ring records. This outcome is now further supported in the revised manuscript with verification results performed in proxy space using independent records, with results presented in the new Table 4 and discussed in Page 14 lines 1 to 7.

Methods, Page 8, line 26: “local” should better be “grid box”.

The word “local” has been changed to “grid cell” in the revised text, page 9, line 18.

Updated reanalysis, Page 9, line 5: Can you explain the localization better? Does a cut-off radius of 25000 km mean that each proxy influences basically the entire globe?

A clear reference to section 4.2 has been added in the first paragraph of section 3 , where the influence of covariance localization is characterized in more detail. Also, more details are now provided in section S5 in the supplementary material. See text modifications on Page 9, line 30 and Page 14, line 16.

Updated reanalysis, Page 9, line 7: Mention that the reference for the skill score is climatology.

We believe that the CE skill score has well-known characteristics among the intended readership of this paper. We also provide the reference which describes the metric in detail Therefore we decline to add further details, in order to avoid lengthening the text.

Figures: some text is too small that it cannot be read in a print version.

We have simplified the design of figures in the revised manuscript. Summary information included as text within figures has been minimized for greater clarity. The font size has also been increased when necessary to increase clarity.

Figures: figures should have consistent font types.

We have revised all figures so that a more uniform visual is achieved.

Anonymous Reviewer 2

This reviewer believes the methodology, analysis and the final LMR product presented herein are too premature to be acceptable for formal publication, let alone for its stated purpose to serve as the basis for the first publicly released NOAA last millennium reanalysis.

The data from the first release described in Hakim et al (2016) has been publically available for over two years. The basic method on which Hakim et al (2016) and the present paper are based has been evaluated and tested extensively in the literature (e.g. Bhend et al, 2012; Steiger et al , 2014; Matsikaris et al, 2015; Acevedo et al, 2017; Franke et al, 2017; Okazaki and Yoshimura, 2017; Steiger et al , 2018).

It is misleading for this study (and its prototype in H16) to call the DA method used ... as an ensemble Kalman filter (EnKF). As in Evensen (1994) and subsequent studies, the primary promise of the EnKF is the use of flow dependent background error covariance represented by the forecasting ensemble. The current socalled "offline" DA method has none of that: the ensemble perturbations are randomly sampled from a past-millennium climate simulation that has no relation to the prior estimate, and the same set of sampled perturbations were used at all analysis times.

The reviewer appears to have a view of data assimilation limited to operational weather forecasting. In fact, from a broader perspective, the prior may come from a wide variety of sources, and Monte Carlo sampling of that distribution using ensembles has proved to be a powerful solution method. The "offline" EnKF approach was originally described by Oke et al (2002) and in Evensen's review of the ensemble Kalman filter (Evensen, 2003), as well as subsequently used in ocean data assimilation applications Oke et al (e.g. 2005, 2007), and in various published paleoclimate data assimilation applications (e.g. Steiger et al , 2014; Matsikaris et al, 2015; Steiger et al , 2018). To make that point clearer to readers perhaps less familiar with this literature, we have added a brief discussion in the last paragraph of section 2.1 outlining the aforementioned literature on the origin and applicability of the technique. See Page 5, lines 13 to 20.

This method used in this study is similar to the commonly used 3D-Var method for numerical weather prediction with static background error covariance, and is arguably less advanced than 3D-Var since 3D-Var in NWP used the dynamic model to propagate the previous cycles analysis as the prior before the analysis. The current so-called "offline" DA method neither cycles the analysis nor the ensemble perturbations, with the stated reason that the forecast model is not good enough to do either.

We regret the reviewer's interpretation that the motivation for offline DA is because "the forecast model is not good enough," which is not the case. Peraps the original text was not clear enough. The choice is a result of a cost–benefit analysis: predictive skill of Earth System models on proxy timescales is small, but the cost of ensemble forecasts with these models is high. We now make that point clearer in the revised manuscript, in the last paragraph of section 2.1, Page 5 lines 16 to 20.

If the forecast model is not good enough to cycle the mean analysis or the analysis uncertainties to provide the best estimate of the prior estimate and related prior uncertainties, why would this model(s) be good at all for use as the prior estimate that the LMR reanalysis depends critically on? In this regards, it is premature to state (line 10) that the "LMR employs the ensemble data assimilation to optimally blend the information from the proxies and the climate model data". The current method is more like an objective analysis method.

Yes, the method is a form of "OI", although we believe that using such jargon is not helpful to the readership of this journal.

It is not clear whether the authors are aware that the traditional static 3D-Var methods also derive the background covariance from an ensemble of perturbations, as is traditionally called "the NMC method" using the sampled forecast divergence between different lead times from many realizations. The Kalman filter update in this case is equivalent to the variational update using the 3D-Var algorithm, though again the 3DVar in NWP cycles the analysis and forecast during data assimilation, which is the most basic function in combining the model and data.

We are indeed aware of the NMC method, which samples forecast differences on the timescale of the DA cycle. In our case that is one year, and the random sampling method we employ assumes that forecast differences on that timescale have converged on the climatological distribution; we lack analyses and forecasts over the Common Era to formally apply the NMC method.

The validation performed in this study for the prototype and updated LMR "reanalysis" with several existing 20th-century reanalysis is misleading at best. The quality of the LMR reanalysis for the 20th century is the least issue given the availability of the modern much more advanced reanalysis and given the exponentially increased number of proxies or model instrumental observations. The validation currently focuses exclusively on the 20th century says little on the quality and performance of the LMR products, in particular over the early period when the proxy data are scarce. A more appropriate validation can potentially be done in two objective methods: (1) perform the 20th century "reanalysis" through thinning the observation density and maybe also degrading the observation accuracy to those representation of different periods of the past millennium; and/or (2) performing observing system experiments in which a certain number of observations are not assimilated but reserved for independent validation (or all of them in cross validation).

Our validation efforts focusing on the 20th century were primarily motivated by the desire to validate against the most robust reference datasets, as well as being able to assess the spatial skill of our reconstructions with some confidence. However we agree this is not comprehensive, as information about performance over the pre-industrial period is not provided. Consistent with suggestion (2), part of the DA method we have developed involves withholding 25% of the proxies for independent validation, which is performed randomly for each of the 51 realizations in each experiment, so that all proxies participate in validation. Validation statistics are compiled both before and during the instrumental period. These results are now highlighted more clearly in numerous different parts in the main body of the revised paper. See Page 12 lines 17 to 23, Page 14, lines 2 to 7, and Page 16 lines 3 to 20.

Regarding suggestion (1), we have performed experiments where we vary the percentage of withheld proxies and we have found little sensitivity to the 25% value.

The use of a 2,5000-km covariance localization is highly questionable for the use of a 100 sets of fixed ensemble perturbations. At midlatitudes, this is amount to the observation impacts across the entire global latitude belt. The use of a fixed set of 100 sample perturbations also means a high rank deficiency over such a large area with this large localization distance.

This reasoning is consistent with covariance lengthscales on weather timescales. Covariance lengthscales on annual timescales are much longer, and the effective dimension of the covariance matrix is comparatively smaller.

The current final LMR reanalysis derives from the mean of 51 such 100-member analyses, should it be the same if the 5100 samples of perturbations are used simultaneously in the Kalman filter update given the Kalman filter used is largely a linear operation?

We have tested this and have found that better results are obtained by averaging over Monte Carlo realizations (multiple analyses) as compared to a single large ensemble. The underlying reason is not completely understood and believe that fully exploring this issue is beyond the scope of the present paper. However, we believe that this finding is an artifact of averaging over analysis errors related to poorly estimated observation errors for some proxies. In the revised manuscript, section 2.3, we highlight this issue as an additional motivation for performing multiple Monte-Carlo realizations and outline the hypothesis we have for the behavior. See revised text Page 6, lines 19 to 21.

How much is the result sensitive to the choice of this arbitrary number of sample perturbations? It is also worth noting the the NMC method used for 3D-Var uses singular value decomposition to make it full rank. Such a approach is different from (and likely more advantageous over) the current Kalman filter update using purely non-envolving static ensemble covariances.

We have found little sensitivity to the ensemble size, provided it is at least 100 members and that covariance localization is used (effectively increases the rank of the covariance matrix). We find larger improvements by randomly subsampling the proxies through many Monte Carlo realizations.

It is unclear what is the purpose of such as hastily done LMR reanalysis products with such ad-hoc DA approaches and the not-good-enough forecast models? The so derived climate trend is almost certainly depending too much on the climate models used as a prior and ensemble sampled perturbations (and maybe the assumed climate forcings used in these models), as well as the density of observations over different periods. It could do more harm if such a premature reanalysis product is used or misused and if it

were publicly released through NOAA, unfortunately. A more careful vetting of the products, and a more concerned effort in refined DA methodology are warranted before NOAA sanctioned such a product as reanalysis, in this reviewers opinion.

If the reviewer wishes to see more sensitivity analysis with respect to these issues, please carefully read Hakim et al (2016), where we not only considered the performance statistics of analyses using different priors and calibrations of proxy forward models, but also examples of the differences that result in the spatial fields for an individual year.

Anonymous Reviewer 3

I have two main concerns about the paper, which together require that the manuscript undergo major revisions before it is acceptable for publication. The first is the character of the derived reconstruction and the unsatisfactory verification of the product using only observational data. The second is the use of multiple ad hoc methodological choices, none of which are reasonably justified or widely tested.

The revised manuscript includes a more complete discussion on the character of the temperature reconstruction, compared and contrasted with other reconstructions (see response to next comment).

The scope of the verification of reconstruction results presented in the paper has been expanded by including results from verification performed in proxy space on independent (withheld) proxies. Results that were originally included in the supplementary material have been revised and moved to the main body of the paper. These results are now found in section 4.1, Table 3, to evaluate in a more comprehensive fashion the various PSM configurations considered and to further confirm choices made for the configuration of the updated reanalysis. Results from proxy–space verification are also included in section 4.3, Table 5, to gain an additional perspective on the accuracy of reconstructions performed with the addition of proxy records from the Anderson et al (2019) collection. See revised text page 14, lines 1 to 7, Page 16 lines 3 to 19.

Regarding the referee's concern on the justification of our selection of data assimilation parameters is addressed with the addition of a brief mention on how those choices were made at the end of section 2.3. We believe however that a lengthy discussion on these is not warranted in the present paper, as these choices have been discussed in prior publications. The following statement has been added in the last paragraph os section 2.3: "Little sensitivity to the use of 75% of the proxies for each realization hase been found (not shown), while 100 members have been chosen to maintain consistency with H16. " , Page 6 lines 19 to 23.

I am struck by the comparison in 2a and the little attention the authors give to the differences between the previous LMR product and the newer version (not to mention the complete lack of comparison between either of these results and other temperature reconstructions).

We agree with this comment. As a result, we now compare LMR results of Northern Hemisphere (NH) temperature to other NH reconstructions found in the IPCC AR4 and AR5 reports. This comparison is included in the bottom panel of Figure 2, section 3, to obtain a greater perspective on the long-term (i.e. across the Common Era) evolution of temperature in our reconstructions. A discussion outlining the main takeaways from this comparison can be found in the fourth paragraph of a revised section 3. See revised manuscript, Page 10 line 23 to Page 11 line 7.

The GMT from the newer product looks almost like white noise and has lost not only the multi-decadal to centennial variability in the first product, it is also likely at odds with the now large collection of global and hemispheric temperature reconstructions spanning the last millennium or more. The authors not only need to spend more time discussing

this issue, they also need to compare their results to the collection of large-scale temperature and hydroclimate reconstructions currently available.

We do acknowledge the updated reanalysis in our originally submitted manuscript was characterized by a significant loss of variability compared to our prototype reanalysis. The comparison with other reconstructions (see previous response) has helped gain additional perspective on the long-term evolution of temperature in our reanalyses. As a result, we have revisited some of the choices made in the configuration of our system and conducted additional experiments to identify the source of this loss of variability. We have determined that it is preferable to eliminate from our reanalysis the large number of unscreened tree-ring records from the Breitenmoser et al (2014) collection, as included in Anderson et al (2019). This is despite increased skill in temperature and hydroclimate reconstructions when this dataset is included, as determined from observational data. We believe this underlines some issues with our estimates of observation error variance (i.e. the R_k terms in equations 4 in the manuscript) for these records. We have found that the inclusion of these tree-ring chronologies lead to the warm bias characterizing the reconstructed temperatures in our previous updated reanalysis during the 1600–1700CE and 1810–1920CE periods of the Little Ice Age (LIA).

In the revised manuscript, we present a new updated reanalysis using a revised configuration of the assimilated proxies, limited to the proxies from the PAGES 2k Consortium (2017) collection. With this configuration, we find that greater level of temperature variability is recovered, as well as a much improved agreement between the LMR Northern Hemisphere temperature and other reconstructions during the LIA. We note however that the absence of a notable warm medieval period continues to characterize our reanalysis. An analysis of results from a large number of reconstruction experiments has allowed us to conclude that colder temperatures during the medieval period, compared to the prototype LMR, are related to the change from the proxy dataset of PAGES 2k Consortium (2013) to the more recent PAGES 2k Consortium (2017) collection. The global temperature composites presented in PAGES 2k Consortium (2017) shows that a distinctly warmer medieval period isn't a prominent feature of the new collection, and is not the result of other updates to our data assimilation system. As this dataset reflects the community's most recent and rigorous identification of proxy records suitable for temperature reconstructions, we believe that the lack of a "classic" medieval warm period in our updated reconstructions of global mean surface temperature and NH-mean temperature should not necessarily be considered an outstanding shortcoming of our updated reanalysis. Discussions in the issues outlined above are included in sections 3 and 4.3. See revised mansucript Page 10 lines 3 to 12, and Page 10 line 23 to Page 11 line 7.

...it is essential for them to do more to verify their results beyond the comparisons they make to observational data. While the latter is important and useful, it is not enough. Incidentally, the authors do perform validation exercises on a withheld period of observational data and using withheld proxy series, but that work is buried in the supplemental and not adequately discussed in this context. More should be made of those efforts, which strengthen the authors results with truly out-of-sample validation experiments.

We agree. As described in our response to a previous similar comment, additional verification results are presented in the revised manuscript. More specifically, verification performed in proxy space on independent proxies are used to further evaluate the impact of the various PSM configurations considered and to further confirm choices made for the configuration of the updated reanalysis (section 4.1, Table 4), and to gain additional perspective on the accuracy of test reconstructions using added records from Anderson et al (2019) (section 4.3, Table 6). See revised mansucript, first paragraph on Page 14, and Page 16, lines 3 to 19.

I do not think the use of CE is the same as it is traditionally used in the paleo literature, given that the latter approach requires a true cross-validation period. The authors should clarify this point.

The formulation has been used in numerous other published work, including in a similar manner as we do in this work. We have found that CE is a valuable complementary metric to correlation, due to its sensitivity to mean errors and representation of variance in the evaluated fields. Furthermore, we now use the CE score using independent proxy data as the verification data.

I think it is further important for the authors to derive validation experiments for the sparsely sampled periods early in the proxy network (e.g. deriving reconstructions using only subsets of the proxies that extend back to specific time intervals). This would go a long way toward helping to better understand the loss of proxy information back in time. This is partially addressed by the variance exercise the authors perform, but more can be done. Recons for temporal subsets of the proxy network would in fact be more useful than the MC sampling of the proxy network that the authors perform, given that it would be systematic and inform a direct question about the influence of the declining proxy network.

This is a very good suggestion. Experiments similar to what is suggested here have been conducted in support of another publication, currently under review. This issue can also be investigated through an assessment of reanalysis performance performed in proxy space, with a focus on different time intervals. We have generated proxy verification results over distinct periods of the Common Era: 0–499, 500–999, 1000–1499, 1500–1879, 1880-2000 (calibration period) in Table 2 and discuss the results from the perspective of reanalysis accuracy across time in a revised section 3. Even though the verification is not performed on an identical set of proxies we also compare the verification statistics from the prototype reanalysis, to further highlight the enhanced performance of the updated LMR. See revised manuscript Page 11, line 34 to Page 12 line 22.

Here is a ... list of choices the authors have adopted that are not accompanied by any justifications or sensitivity discussion: 1-Use of 100-member ensemble; 2-Use of the CCSM4 last millennium simulation as the prior; 3-Use of 51 MC realizations; 4-Use of a proxy sampling scheme based on 75% of the proxy records; 5-Degradation of the model resolution to a ∼5x5 grid. All of these choices undoubtedly influence the derived LMR product. Some of them can be justified based on discussions in the literature. Some of them require empirical demonstrations. All of them come across as ad hoc. I would also venture to guess that the LMR results are more dependent on a couple of these choices

than the other dependencies that the authors more systematically test. It is therefore essential that the authors do a better job of justifying these choices and convincing the reader that they are either reasonable choices or chosen based on some methodological/logistical rationale.

Some of these choices have been discussed in prior publications. See Hakim et al (2016) and references therein for example. However we agree that text reminding the reader of how parameters were chosen should be included to convince that careful consideration was involved. To further clarify the context, we should point out that some parameter values used here were chosen to maintain consistency with the configuration used in Hakim et al (2016), in order emphasize the impact of updates specifically addressed in this work.

Some of the parameters were originally chosen based on practical considerations. For example, the ratio of assimilated versus withheld proxies and the number of Monte-Carlo realizations were chosen so that the random sample of proxies set aside for independent validation is representative of the overall dataset, while maintaining minimal sensitivity on the accuracy of the the reanalysis. A revised statement clarifying these points has been added in the last paragraph of section 2.3., page 6, lines 17 to 23.

I should specifically mention the use of the CCSM4 as the prior. The authors say nothing about how their results might depend on the model prior and whether they have tested alternative last-millennium simulations in their analysis. This is an obvious question and the authors need to address it.

This topic is addressed in Hakim et al (2016), with results from reconstructions using a wide range of calibration and prior data (see Fig. 12 of that paper). Work on this topic has since been expanded and explored in more detail. These results are, however, expected to be the subject of a separate publication.

Page 2, lines 5-6: What does ”synthesizing information” mean? This is vague and I am not even sure the statement is true. There are lots of central challenges of paleoclimate science, and it is arguable that what the authors are alluding to is one of them. This strikes me as an unsupported justification for what the authors subsequently say they are attempting to do.

In this part of the text, our goal is to emphasize that data assimilation is a powerful framework in which information from proxies and numerical model simulations is combined for the production of, hopefully, robust climate reconstructions. Therefore, we believe that this goal is appropriately described by the verb ”synthesize”, which may be defined as the act of combining (a number of things) into a coherent whole. We modified the statement as being “one of the challenges of paleoclmate science” in order to more accurately convey our belief that other challenges than this one exist. See revised manuscript Page 2 line 1.

Page 2, lines 30-32: This is a much more mundane objective than the sense given in the abstract. Are the authors attempting to release a shiny new LMR product or should this be seen as an iterative verification step toward some improved effort down the line?

We apologize if the text does not clearly convey the goals pursued here. As with any complex problem to be solved, as is the case here, improvements are generally incremental and modest. The main goal of efforts reported in this paper has been to improve our system over the LMR prototype, and to deliver an incrementally improved product, while being fully cognizant of the fact that room for improvement remains. A revised statement at the end of the third paragraph now includes “compared to the prototype LMR” to better convey the scope of our objective, while the remaining room for improvement is acknowledged in a clearer manner in the next to last paragraph of the conclusion. See Page 2 lines 29-30, and Page 17, lines 12 and 13.

Page 8, line 9: The use of precipitation is not justified and concerning. First, precipitation is almost never the variable associated with moisture sensitivity in trees - some measure of soil moisture is. It is therefore not clear why the authors used precipitation and how it influences their results. Why not use a more conventional variable like PDSI? Secondly, how do the characteristics of precipitation influence the results? Does it matter that precip is likely not Gaussian and that it has limited spatial and temporal covariance structure? Is the use of precip perhaps adding to the loss of low-frequency variance in this new LMR product? My guess is that this specific choice has a large impact on the derived reconstruction and the use of precip is not justified in any way.

The reviewer is raising fair and important points here. We acknowledge that soil moisture is the preferred response variable for the modeling of tree-ring widths. However, given our approach, which relies on the availability of calibrated forward models, the absence, to our knowledge, of a reliable century-long soil moisture dataset is an important limiting factor. An alternative, as pointed out by the referee, would be PDSI (or other drought indices such as SPEI) instead of soil moisture. This option, is enabled within our framework with the use of the Dai et al (2004) dataset for forward model calibration (see below). To provide additional context about our efforts, more particularly with the development of bivariate “temperature and moisture” PSMs, our intent is to replicate VSlite’s (Tolwinski-Ward et al, 2011) general approach of combining the influences of temperature and moisture in modeling tree-ring width variability, albeit in a simpler fashion, while avoiding the issues associated with VSlite’s formulation involving thresholds (see Dee et al, 2016). We use precipitation as the moisture source variable in bivariate models rather than PDSI, since the latter is itself a function of temperature through the potential evapotranspiration term involved in its calculation.

In response to the reviewer’s concern, we compare two reconstruction experiments in which the univariate “temperature or moisture” (“TorM”) models are used to forward model tree-ring width proxies. One experiment is performed with PSMs calibrated on PDSI using the dataset from Dai et al (2004), while the other uses models calibrated on precipitation with the GPCC dataset (Schneider et al, 2014). In each case, as described in section 2.4.2 in the main text, decisions are made whether each tree-ring record is moisture sensitive or temperature sensitive by comparing moisture calibrations (with PDSI or precipitation) and temperature calibrations (using GISTEMP as the calibration data). The regression providing the better fit is used to forward model the proxy record. In all experiments. all other proxies are modeled with univariate temperature PSMs. Other reconstruction parameters are the same as with experiments described in the paper (100 ensemble members, CCSM4 as the prior model, 51 Monte-Carlo realizations with 75% of proxies assimilated). Covariance localization is not applied here. Temperature reconstructions from both experiments are verified against instrumental-era analyses and against independent (withheld from assimilation) proxies and results are compared in Tables AR.1 and AR.2. Results obtained from a reconstruction using bivariate PSMs calibrated on temperature and precipitation are included for comparison. The skill of reconstructions using the univariate models are similar, whether PSMs are calibrated on PDSI or precipitation. But more importantly, the skill scores are generally inferior to those obtained with the use of bivariate models. This is particularly true with the CE score. Proxy-based verification results are less definitive due to the noisy nature of the proxies, but generally support to the conclusion drawn from instrumental-era verification. We have also calculated the spectra of Northern Hemisphere mean temperature from both reconstructions, shown in Fig. AR.1. The results are nearly identical (indistinguishable at the 95% confidence level) , suggesting that the use of precipitation-calibrated PSMs does not lead to a loss of variability in temperature reconstructions. Furthermore, the exercise has shown that bivariate temperature/precipitation PSMs provide superior results compared to univatiate "TorM" models, even when the latter are calibrated on PDSI. As part of the revision of the manuscript, the results discussed herein are now included in the supplementary material, while a statement referring to section S4 of the supplementary material is included in the last sentence of section 2.4. See revised manuscript Page 7, lines 33-34.

Figures: In general, there is a lot of small text in the figures that is hard to read and also rather confusing and messy.

We have simplified the design of figures in the revised manuscript. The summary information included as text on the figures has been minimized for greater clarity.

Figures: The many colorbars are also unnecessary in many plots when one would do.

We thank you for your recommendation. We acknowledge the presence of multiple colorbars, however we point out that the ranges shown in the various frames are not all identical. Therefore, we elect to keep showing the colorbars. However we made efforts in cleaning up and streamlining the design of the figures for greater clarity.

References

Acevedo, W., Fallah, B., Reich, S. and Cubasch, U., Assimilation of pseudo-tree-ring-width observations into an atmospheric general circulation model, *Climate of the Past*, 13, 545–557, doi:10.5194/cp-13-545-2017, 2017

Anderson, D. M. and Tardif, R. and Horlick, K. and Erb, M. P. and Hakim, G. J. and Emile-Geay, J. and Noone, D. and Perkins, W. A. and Steig, E. J., Additions to the Last Millennium Reanalysis multi-proxy database, *Data Science Journal*, 18, doi:110.5334/dsj-2019-002, 2019

Bhend, J., Franke, J., Folini, D., Wild, M. and Brönnimann, S., An ensemble-based approach to climate reconstruction, *Climate of the Past*, 8, 963–976, 2012.

Breitenmoser, P., Brönnimann, S. and Frank, D, Forward modelling of tree-ring width and comparison with a global network of tree-ring chronologies, *Climate of the Past*, 10, 437–449, doi:10.5194/cp-10-437-2014, 2014

Dai, A., Trenberth, K. E. and Qian, T., A global data set of Palmer Drought Severity Index for 1870-2002: Relationship with soil moisture and effects of surface warming, *J. Hydrometeorology*, 5, 1117–130, doi:10.1175/JHM-386.1, 2004

D'Arrigo, R., Mashig, E., Frank, D., Wilson, R. and Jacoby, G., Temperature variability over the past millennium inferred from Northwestern Alaska tree rings, *Clim. Dyn.*, 44, 227–236, doi:10.1007/s00382-004-0502-1, 2005

Dee, S., Steiger, N. J., Emile-Geay, J. and Hakim, G. J., The utility of proxy system modeling in estimating climate states over the Common Era, *J. Adv. Model. Earth Syst.*, 8, 1164–1179, doi: 10.1002/2016MS000677, 2016

Diaz, H. F., Trigo, R., Hughes, M. K., Mann, M. E., Xoplaki, E. and Barriopedro, D., Spatial and temporal characteristics of the climate in medieval times revisited, *Bull. Amer. Meteorol. Soc.*,99(11), 1487–1500, doi:10.1175/BAMS-D-10-05003.1, 2011

Evensen, G., The ensemble Kalman filter: Theoretical formulation and practical implementation, *Ocean Dyn.*, 53, 343–367, 2003

Franke, J., Brönnimann, S., Bhend, J. and Brugnara, Y., A monthly global paleo-reanalysis of the atmosphere from 1600 to 2005 for studying past climatic variations, *Sci. Data*, 4, doi: 10.1038/sdata.2017.76, 2017

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., Steiger, N. J. and Perkins, W. A., (2016) The Last Millennium Climate Reanalysis Project: Framework and First Results, *JGR: Atmospheres*, 121, 6745–6764, doi:10.1002/2016JD024751, 2016

Matsikaris, A., Widmann, M. and Jungclaus, J., On-line and off-line data assimilation in palaeoclimatology: a case study, *Climate of the Past*, 11, 81–93, doi:10.5194/cp-11-81-2015, 2015

McKay, N. P. and Emile-Geay, J., Technical note: The Linked Paleo Data framework a common tongue for paleoclimatology, *Clim. Past*, 12, 1093–1100, doi:10.5194/cp-12-1093-2016, 2016.

Okazaki, A. and Yoshimura, K., Development and evaluation of a system of proxy data assimilation for paleoclimate reconstruction, *Climate of the Past*, 13, 379–393,doi:10.5194/cp-13-379-2017, 2017

Oke, P. R., Allen, J. S., Miller, R. N., Egbert, G. D. and Kosro, P. M., Assimilation of surface velocity data into a primitive equation coastal ocean model, *JGR:Oceans*, 107, doi:10.1029/2000JC000511, 2002

Oke, P.R., Schiller, A. Griffin, D. A. and Brassington, G. B., Ensemble data assimilation for an eddy-resolving ocean model of the Australian region, *Quart. J. of the Royal Meteor Soc.*, 131, 3301–3311, 2005

Oke, P. R., Sakov, P. and Corney, S. P., Impacts of localisation in the EnKF and EnOI: experiments with a small model, *Ocean Dyn.*, 57, 32–45, 2007

PAGES 2k Consortium, Continental-scale temperature variability during the past two millennia, *Nature Geoscience*, 6(5), doi:10.1038/ngeo1797, 2013

PAGES 2k Consortium, A global multiproxy database for temperature reconstructions of the Common Era, *Scientific Data*, 4(170088), 1–33, doi:10.1038/sdata.2017.88, 2017

Perkins, W. A. and Hakim, G. J., Reconstructing paleoclimate fields using online data assimilation with a linear inverse model, *Climate of the Past*, 13, 421–436, 2017

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B., GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle, *Theoretical and Applied Climatology*, 115, 15–49, doi:10.1007/s00704-013-0860-x, 2014

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S. and Roe, G. H., Assimilation of time-averaged pseudoproxies for climate reconstruction, *J. of Climate*, 27, 426–441, doi: 10.1175/JCLI-D-12-00693.1, 2014

Steiger, N. J., Smerdon, J. E., Cook, E. R. and Cook, B. I., A reconstruction of global hydroclimate and dynamical variables over the Common Era, *Sci. Data*, 5, doi: 10.1086/sdata.2018.86, 2018

Tolwinski-Ward, S. E., Evans, M. N., Hughes, M. K. and Anchukaitis, K. J., An efficient forward model of the climate controls on interannnual variation in tree-ring width, *Climate Dyn.*, 36, 2419–2439, doi: 10.1007/s00382-011-1062-9, 2011

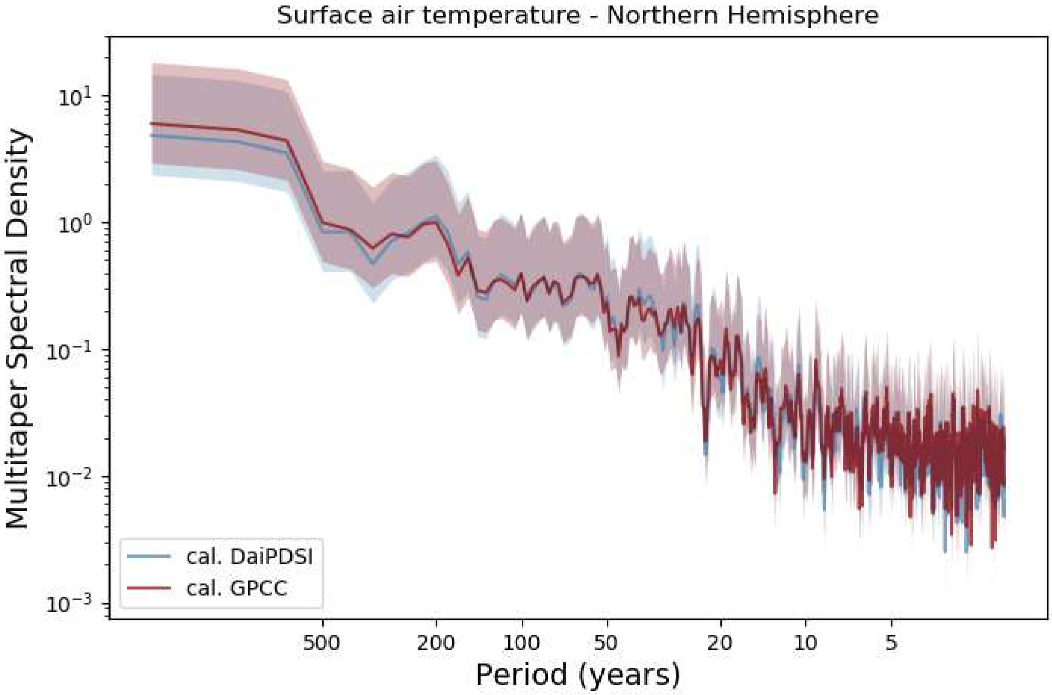


Figure AR.1: Comparison of spectra of the Northern Hemisphere mean temperature from two Common Era reconstructions using the Dai et al (2004) PDSI dataset versus the GPCC precipitation dataset (Schneider et al, 2014) to calibrate univariate “TorM” PSMs used to model tree-ring width proxies. Shading indicate the χ^2 95% highest density regions (HDR).

Table AR.1: Summary of instrumental–era verification results for reconstructions performed with various PSM configurations to model tree-ring width proxies (see main text, section 2.4.2 for PSM details). Verification scores shown are correlation (r) and coefficient of efficiency (CE) for the annual global mean temperature (GMT) and detrended GMT verified against the consensus of instrumental–era analyses, the global mean of gridpoint r and CE characterizing the spatially reconstructed temperature, verified against the Berkeley Earth analysis.

PSM configuration	Annual GMT		Detrended GMT		Spatial temperature	
	r	CE	r	CE	r	CE
Univariate “TorM” (PDSI)	0.92	0.79	0.72	0.46	0.52	0.17
Univariate “TorM” (precip.)	0.93	0.77	0.74	0.48	0.53	0.17
Bivariate	0.93	0.86	0.77	0.54	0.53	0.20

Table AR.2: Verification of LMR reconstructions against independent (withheld from assimilation) proxies, for experiments using various PSM configurations to model tree-ring width proxies. Skill scores shown are the median of distributions for correlation (r), the fraction of proxy records characterized by a positive ΔCE (%+CE), and the median of the ΔCE distribution. Statistics are compiled over 51 Monte-Carlo realizations, for two distinct periods: 1880–2000 (PSM calibration period) and 0–1879 (pre-calibration period).

PSM configuration	1880–2000			0–1879		
	r	%+CE	ΔCE	r	%+CE	ΔCE
Univariate “TorM” (PDSI)	0.34	80.0	0.09	0.21	70.4	0.05
Univariate “TorM” (precip.)	0.33	77.6	0.08	0.19	66.3	0.04
Bivariate	0.36	78.9	0.11	0.22	66.0	0.06

Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling

Robert Tardif[1], Gregory J. Hakim[1], Walter A. Perkins[1], Kaleb A. Horlick[2], Michael P. Erb[3], Julien Emile-Geay[4], David M. Anderson[5], Eric J. Steig[6,1], and David Noone[2]

[1]Department of Atmospheric Sciences, University of Washington, Seattle, Washington, USA
[2]College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, Oregon, USA
[3]School of Earth Sciences and Environmental Sustainability, Northern Arizona University, Flagstaff, AZ, USA
[4]Department of Earth Sciences, University of Southern California, Los Angeles, California, USA
[5]Retired, NOAA Paleoclimatology Program, Boulder, CO, USA
[6]Department of Earth and Space Sciences, University of Washington, Seattle, Washington, USA

Correspondence: R. Tardif (rtardif@atmos.washington.edu)

Abstract.

The Last Millennium Reanalysis utilizes an ensemble methodology to assimilate paleoclimate data for the production of annually resolved climate field reconstructions of the Common Era. Two key elements are the focus of this work: the set of assimilated proxy records, and the forward models that map climate variables to proxy measurements. Results based on an ~~extensive~~ updated proxy database and seasonal regression-based forward models are compared to the prototype reanalysis of Hakim et al. (2016), which was based on a smaller set of proxy records and simpler proxy models formulated as univariate linear regressions against annual temperature. Validation against various instrumental–era gridded analyses shows that the new reconstructions of surface air temperature, 500 hPa geopotential height ~~and the Palmer Drought Severity Index~~ are significantly improved ~~, with skill scores increasing~~ (from 10% to more than ~~200%, depending on the variable and verification measure~~100%), while improvements in reconstruction of the Palmer Drought Severity Index are more modest. Additional experiments designed to isolate the sources of improvement reveal the importance of ~~additional~~ the updated proxy records, including coral records for improving tropical reconstructions; ~~tree-ring-width chronologies, including moisture-sensitive trees, for thermodynamic and hydroclimate variables in mid-latitudes; and~~ and tree-ring density records for temperature reconstructions, particularly in high northern latitudes. Proxy forward models that account for seasonal responses, and ~~the dual sensitivity to~~ dependence on both temperature and moisture ~~characterizing tree-ring-width proxies, are also found to be particularly important. Other experiments highlight the beneficial role of covariance localization on reanalysis ensemble characteristics. This improved paleoclimate data assimilation system served as the basis for the production of the first publicly released NOAA Last Millennium Reanalysis.~~

for tree-ring width, also contribute to improvements in reconstructed thermodynamic and hydroclimate variables in mid-latitudes.

1 Introduction

Reconstructions of Earth's past climate, particularly covering periods prior to instrumental data sets, are key to understanding the causes of natural climate variability. For example, understanding natural variability provides the basis for improving predictions of climate variability in the coming decades. Information on past climates has traditionally been derived either from climate proxy data (e.g., tree rings, ice cores, etc.) or from Earth system model simulations~~. Synthesizing~~ , and synthesizing information from these two sources is ~~a central challenge of~~ one of the challenges of paleoclimate science. Paleoclimate data assimilation (PDA) has emerged as a powerful framework for such synthesis because it provides the optimal combination of climate signals recorded by proxies as constrained by the dynamics of Earth system models. PDA-generated climate field reconstructions have been used to investigate climate variability prior to the instrumental era (Goosse et al., 2006, 2010; Widmann et al., 2010; Bhend et al., 2012; Steiger et al., 2014; Matsikaris et al., 2015; Franke et al., 2017; Okazaki and Yoshimura, 2017; Steiger et al., 2018). Within this general PDA framework, a flexible PDA system is being developed for the Last Millennium climate Reanalysis (LMR) project for the production of annually resolved reconstructions of the Common Era. Hakim et al. (2016) describe a prototype configuration of the LMR and show results in good agreement with previous reconstructions of Northern Hemisphere mean near-surface air temperature. Detailed comparisons with several gridded instrumental temperature data products revealed significant skill over tropical regions but less skillful reconstructions over Northern Hemisphere continental areas, where a large proportion of proxy data are located.

As with any data assimilation system, two of three important components impacting the quality of the resulting analyses are the set of assimilated observations (here, proxy records) and the forward models that map variables from climate model output to proxy measurements ("proxy system models"; hereafter, PSMs). The third component is the model providing the prior state, although it is not the focus of this work. Hakim et al. (2016) assimilated proxy records from the first compilation of the PAGES 2k Consortium (PAGES 2k Consortium, 2013), and modeled the proxies through univariate linear regressions calibrated against annually-averaged instrumental temperature data. Here we examine the impact on LMR reconstructions of improvements to these two key components: (1) ~~a much larger proxy data base, with~~ an updated and expanded proxy database, primarily composed of records from PAGES 2k Consortium (2017), ~~Breitenmoser et al. (2014) and~~ and assessment of the additional records described in ~~?~~Anderson et al. (2019); (2) more realistic PSMs in which seasonality and, for tree-ring-width proxies, temperature and moisture sensitivity are taken into account. Motivation for expanding the proxy database derives from evidence that climate reconstructions are generally sensitive to the set of proxy records used as input (e.g., Wang et al., 2015), while the introduction of more sophisticated PSMs is ~~in part motivated~~ motivated in part by the fact that comprehensive reconstructions of temperature and hydroclimate variables depend on properly treating temperature-sensitive and moisture-sensitive tree ring proxies (e.g., Steiger and Smerdon, 2017).

The focus of improvements in PSMs here is on regression-based (i.e. statistical) forward models, in contrast to recent efforts focusing on process-based PSMs (see e.g., Breitenmoser et al., 2014; Dee et al., 2016; Acevedo et al., 2017). Our objective is to establish baseline skill of PDA reconstructions using statistical PSMs, to serve as a benchmark for evaluating possible improvements associated with process-based PSMs (e.g., Dee et al., 2016). Here we develop a hierarchy of statistical PSMs to identify aspects that contribute increased skill to reconstructed temperature and hydroclimate states compared to the prototype LMR.

The remainder of the paper is organized as follows. Section 2 outlines the LMR PDA-based framework and describes the proxy database and PSMs. Reconstructions based on this configuration are presented in section 3, with comparisons to the prototype described in Hakim et al. (2016). Section 4 explores the contributions to improvements in the new reconstructions. A concluding summary is given in section 5.

2 Methods

Paleoclimate data assimilation has three main components: proxy records, providing an indirect record of past climatic conditions; climate models, providing prior estimates of the climate; and proxy system models, providing the connection between the model prior and the proxy values. The method for each component is now described.

2.1 Data assimilation framework

LMR employs ensemble data assimilation (DA) to ~~optimally~~ blend information from proxies and climate model data. DA is performed using a variant of the ensemble Kalman filter, which for our application appears to perform well compared to alternative PDA methods such as particle filters (Liu et al., 2017). The update equation is given by

$$\mathbf{x}_a = \mathbf{x}_b + \mathbf{K}[\mathbf{y} - \mathbf{y}_e]. \tag{1}$$

Here, $\mathbf{x}_b$ is the prior state vector, which contains the climate variables to be reconstructed, averaged over an appropriate timescale (here, annual), ~~while~~ and $\mathbf{x}_a$ is the posterior state vector (i.e. the reanalysis, or reconstruction). The state vector may include scalars, such as climate indices, and/or grid-point data for spatial fields. Vector $\mathbf{y}$ contains the assimilated proxy data (i.e. observations) and $\mathbf{y}_e$ is a vector containing estimates of the proxies derived from the prior by

$$\mathbf{y}_e = \mathcal{H}(\mathbf{x}_b), \tag{2}$$

where $\mathcal{H}$ is the forward model mapping the prior $\mathbf{x}_b$ to proxy space (i.e. the PSM, see section 2.4). The innovation, $[\mathbf{y} - \mathbf{y}_e]$, is the new information from the proxies not already contained in the prior. This new information is weighted against the prior through the Kalman gain matrix

$$\mathbf{K} = \mathbf{B}\mathbf{H}^{\mathrm{T}}\left[\mathbf{H}\mathbf{B}\mathbf{H}^{\mathrm{T}} + \mathbf{R}\right]^{-1} \tag{3}$$

where $\mathbf{B}$ is the prior covariance matrix, $\mathbf{R}$ is the error covariance matrix for the proxy data, and $\mathbf{H}$ is the linearization of $\mathcal{H}$ about the mean value of the prior. Here, Eq. (1) is solved using the ensemble square-root filter (EnSRF) approach of Whitaker and Hamill (2002), in which the ensemble-mean and perturbations about the ensemble mean are solved separately. Moreover, $\mathbf{R}$ is taken as a diagonal matrix (uncorrelated observation errors) where the diagonal elements represent the error variance for each assimilated proxy record; details on how these are estimated are provided in Section 2.4. This allows for serial processing of observations, in which observations are assimilated one at a time. This greatly simplifies the implementation of covariance localization, which is used to control sampling error in the prior covariance. Solutions for the ensemble mean, $\overline{\mathbf{x}}_a$, and perturbations, $\mathbf{x}'_a$, for the single k^{th} proxy y_k, are obtained from:

$$\overline{\mathbf{x}}_a = \overline{\mathbf{x}}_b + \frac{w_{loc} \circ cov(\mathbf{x}_b, y_{e,k})}{var(y_{e,k}) + R_k} \left(y_k - \overline{y}_{e,k}\right) \tag{4a}$$

$$\mathbf{x}'_a = \mathbf{x}'_b - \left[1 + \sqrt{\frac{R_k}{var(y_{e,k}) + R_k}}\right]^{-1} \frac{w_{loc} \circ cov(\mathbf{x}_b, y_{e,k})}{var(y_{e,k}) + R_k} \left(y'_{e,k}\right) \tag{4b}$$

where $y_{e,k}$ is the prior estimate of the proxy from (2) and R_k is the diagonal element of $\mathbf{R}$ corresponding to proxy y_k. The ensemble of updated states is then recovered by combining the posterior ensemble mean and perturbations

$$\mathbf{x}_a = \overline{\mathbf{x}}_a + \mathbf{x}'_a. \tag{5}$$

Covariance localization, given by a Schur product denoted by $\circ$ in Eqs. 4 (i.e. element-wise multiplication), is a distance-weighted filter w_{loc} on the prior covariance matrix (see e.g. Hamill et al., 2001). ~~Section 4.2provides for~~ Sections 4.2, and S5 in the supplementary material, provide more details on localization.

We also use an "appended state vector" approach ~~for serially processing proxy measurements~~ that avoids the need to re-compute (2) after each proxy is assimilated. ~~This approach also simplifies the~~ The $\mathbf{y}_e$ proxy estimates from each record are appended to the state vector $\mathbf{x}_b$:

$$\mathbf{x}_b = \begin{bmatrix} x_1 \\ \vdots \\ x_N \\ y_e^1 \\ \vdots \\ y_e^P \end{bmatrix}, \tag{6}$$

where the $x_1 \ldots x_N$ elements contain the ensemble grid point data from model variables included in the state (e.g. temperature, precipitation etc.), with N the sum of the number of variables times the number of grid points, and the $y_e^1 \ldots y_e^P$ are the ensemble proxy estimates for each of the P proxy records considered. Each of the $x_1 \ldots x_N$ and $y_e^1 \ldots y_e^P$ elements are of dimensions $1 \times N_{ens}$, where N_{ens} is the specified size of the ensemble. Hence, $\mathbf{x}_b$ is a matrix of dimension $(N+P) \times N_{ens}$. With such an appended state, the y_e elements in Eq. 6 are updated through Eqs. 4 as any other state variables, eliminating the need to re-evaluate $\mathbf{y}_e$ with Eq. 2 once the state has been updated. This simplification is particularly attractive in the context of LMR updates discussed herein as it enables a straightforward implementation of seasonal ~~forward models as described in more detail in Appendix ??.~~ PSMs (i.e. forward models more accurately representing the seasonal responses of individual proxy records) as discussed in section 2.4. In our implementation for the reconstruction of annually-averaged states, the data assimilation procedure follows this general algorithm.

1. The proxy estimates ($\mathbf{y}_e$) are pre-calculated using Eq. 2 with either annually- or seasonally-averaged model data as input (i.e. the $\mathbf{x}_b$ in Eq. 2).

2. A sample ensemble of annually-averaged model states is randomly drawn from a pre-existing simulation to form the main part of the prior state vector (i.e. the $x_1 \ldots x_N$ elements in Eq. 6).

3. The pre-calculated $y_e^1 \ldots y_e^P$ proxy estimates are added on to form the appended state as shown in Eq. 6. This appended state becomes the $\mathbf{x}_b$ in Eq. 1, which is decomposed in an ensemble-mean ($\overline{\mathbf{x}}_b$) and perturbations about the mean ($\mathbf{x}'_b$) as shown in Eqs. 4.

4. Proxies forming the $\mathbf{y}$ vector are then serially processed, with the updated state, including the proxy estimates, obtained from Eqs. 4. The reanalysis is completed for one year once all proxies have been assimilated.

We note here that with a configuration involving seasonal PSMs without the use of an appended state, the vector $\mathbf{x}_b$ has to include states with sufficient temporal resolution to allow the calculation of the updated seasonal $y_e^1 \ldots y_e^P$ proxy estimates. In this scenario, an additional step to the ones listed above is required, involving Eq. 2 using the appropriate seasonally-averaged updated states as input. With proxies characterized by a wide range of seasonal responses, this requirement would impose an $\mathbf{x}_b$ composed of monthly data which would greatly increase the computational cost of the reanalysis. Reanalysis results would also likely be adversely affected by the larger noise level characterizing data at shorter (i.e. monthly) timescales through its impact on ensemble estimates of prior covariances (see, e.g. Tardif et al., 2016).

As in Hakim et al. (2016), an "offline" DA approach is used, where the prior ensemble is formed by random draws of time–averaged states from a pre-existing millennial-long model simulation, with the same randomly drawn ensemble members used for every year in the reconstruction of a given reanalysis realization (see Section 2.3 below). This is in contrast to online DA (e.g., Matsikaris et al., 2015; Perkins and Hakim, 2017), where a numerical model is used to dynamically forecast the evolution of climate states from the latest proxy-informed analysis to the following year, when new proxy observations are assimilated. The "offline" approach, introduced by Oke et al. (2002) and Evensen (2003) and used in an ocean DA system by Oke et al. (2005), offers several practical advantages, particularly from a computational cost perspective ~~. Furthermore, it is~~ (Oke et al., 2007). Its use is further justified when model forecasts have ~~little~~ limited skill over timescales corresponding to the time interval between updates, as is the case here with global climate models and proxies assimilated on an annual basis. This scenario is further supported by the PDA results of Matsikaris et al. (2015) who show similar performance is achieved with online and offline approaches. From a cost-benefit perspective, the high cost of running ensembles of comprehensive global climate model simulations does not appear justified. However, ongoing research suggests cost-effective online PDA may be achieved by using simplified climate models (Perkins and Hakim, 2017).

2.2 Climate proxies

~~Here we use a combination of proxy records from~~ Our proxy database is updated to the latest PAGES 2k collection ~~(PAGES 2k Consortium, and the additional records assembled by ?. As in the LMR prototype (Hakim et al., 2016, hereafter H16), only records with sub-annual to annual resolutions are considered; sub-annual records are averaged to annual. The PAGES 2k Consortium (2017) dataset (hereafter PAGES2k-2017) is a~~ (PAGES 2k Consortium, 2017, hereafter PAGES2k-2017). This dataset represents the community standard in global proxy observations covering the Common Era (CE) ~~,~~ and serves as the core source of proxy information used ~~here. Proxies included in~~ in our updated reanalysis. PAGES2k-2017 proxies were screened to retain temperature-sensitive records~~only~~, extensively quality controlled, and described by more metadata compared to previous collections. ~~Figure 1 shows the various proxy networks considered for assimilation, comparing~~ The additional records assembled by Anderson et al. (2019)[1], consisting in large part of the tree ring width records from Breitenmoser et al. (2014) (hereafter B14), are considered as potential enhancement to proxy information used in our paleo-reanalyses (see section 4.3).

As in the LMR prototype (Hakim et al., 2016, hereafter H16), only records with sub-annual to annual resolutions are considered; sub-annual records are averaged to annual. Figure 1 compares the PAGES 2k Consortium (2013) (hereafter PAGES2k-2013) dataset used in H16, ~~to~~ and the PAGES2k-2017 update~~, and to the entire updated LMR database including the records of ?~~ . Only records for which a PSM can be established are shown in Fig. 1, defined by proxy records with at least 25 years of (non-contiguous) overlap with calibration data (see section 2.4).

Compared to the proxies assimilated in H16, PAGES2k-2017 data provide enhanced spatial coverage in the tropics with additional coral δ^{18}O and Sr/Ca records. Additional tree ring wood density records from Europe and western North America are also included. The temporal distribution of the total number of records remains similar, except for significant increases in the number of tree ring width and coral proxies during 1800–2000CE, and tree ring wood density records during 1500–2000CE. ~~The additional records from ? include the large number of tree ring width records from Breitenmoser et al. (2014) (hereafter B14), not strictly screened for temperature sensitivity in contrast to the PAGES 2k collection. These records provide enhanced coverage over eastern North America, southern Europe, boreal Eurasia and southern South America. Two tree ring records are also added in the data-void region of southern Africa. Other additions, totaling 94 records, provide additional records in the Tropics (23 coral records), and a greatly enhanced number of ice core records concentrated over Greenland and eastern Canadian Arctic (37 records) and Antarctica (26 records in West Antarctica and Drönning Maud Land). A few lower latitude ice core records (6 records) are also added in the Peruvian Andes and Tibetan Plateau, along with two higher latitude lake core records. From a temporal perspective relative to proxies used in H16, the addition of the tree ring width records from B14 contributes a notable number of additional proxies back to 1000CE, more than double the number of records available for assimilation from 1500CE onward, up to a fourfold increase during the 19th and 20th centuries.~~

[0] ~~An exception is the use of the Palmyra coral record from Cobb et al. (2003) rather than the Emile-Geay et al. (2013) update, as described in ?.~~

[1] An exception is the use of the Palmyra coral record from Cobb et al. (2003) rather than the Emile-Geay et al. (2013) update, as described in Anderson et al. (2019).

2.3 Climate model prior information

For all reconstruction experiments reported in this paper, the prior state vector is formed with data from the CMIP5 (Taylor et al., 2012) Last Millennium simulation from the Community Climate System Model version 4 (CCSM4) coupled atmosphere-ocean-sea ice model. The simulation covers years 850 to 1850 CE and includes incoming solar variability, variable greenhouse gases as well as stratospheric aerosols from volcanic eruptions known to have occurred during the simulation period (see Landrum et al., 2013). The same "offline" DA methodology as in H16 is used, where the prior ensemble is a random sample of annual averages, with the same sample used for all years of the reconstruction. The sampled states are deviations (i.e. anomalies) from the temporal mean taken over the entire length of the simulation. Therefore, the prior ensemble–mean does not contain time-specific information about climate events (e.g. a volcanic eruption) or trends characterizing specific periods (e.g. twentieth century warming). Finally, the spatial resolution of prior state variables is reduced from $0.95^o \times 1.25^o$ of the Last Millennium simulation to a of $4.3^o \times 5.7^o$ Gaussian grid as in H16.

All reconstruction experiments are composed of 51 Monte-Carlo assimilation realizations, each using a different randomly chosen 100–member ensemble and 75% of available proxy records for assimilation. This Monte-Carlo sampling over subsets of prior states and proxy records is designed to incorporate uncertainties in covariance estimates derived from model states, and uncertainties associated with proxy error estimates. Moreover, we have found that averaging over ensembles from Monte-Carlo realizations leads to more accurate results. This is likely the result of averaging over random errors introduced into the reanalysis from few randomly chosen proxy records with underestimated observation errors. Little sensitivity to the use of 75% of the proxies for each realization has been found (not shown), while 100 members have been chosen to maintain consistency with H16. In the following, climate reanalyses are taken as the mean over the 100–member DA ensembles and 51 Monte-Carlo realizations (i.e., a 5100–member "grand ensemble").

2.4 Proxy modeling

A critical component of PDA is the mapping of prior climate state variables (e.g., temperature, precipitation from a climate model) to the assimilated proxies (e.g., tree ring width). This is expressed mathematically by Eq. 2, section 2.1, where the operator $\mathcal{H}$ (i.e. the forward model) ideally represents the complete set of processes associated with proxy values, i.e. a comprehensive physically-based PSM. This remains a major challenge as the information archive is often complex, involving physical, biological and chemical processes (Evans et al., 2013). Despite recent progress in the development and use of process-based PSMs (e.g., Dee et al., 2015, 2016; Goosse, 2016; Steiger et al., 2017; Acevedo et al., 2017), the focus here is on statistical PSMs, which offer distinct advantages: 1) ease of implementation and flexibility with respect to forward modeling of multiple proxies, regardless of archive types, measurements, units etc.; 2) observation error statistics for each assimilated record are well-defined from the regression (see below); and 3) regressions are formulated on the basis of deviations from the mean over a reference period (e.g. 1951–1980) of the driving climate variable(s), therefore avoiding issues with absolute calibration where climate model bias is problematic, particularly for PSMs having threshold transitions (see e.g., Dee et al., 2016). Statistical PSMs also have distinct disadvantages: 1) PSMs cannot be calibrated without sufficient overlap with calibration data (a threshold of at least 25 overlapping data is imposed); 2) the accuracy of the models depends on the limitations of the calibration datasets (e.g. less reliable analysis over the Southern Ocean and over high latitude continental areas due to a lack of observations); and 3) possible lack of stationarity of the derived relationships established with instrumental–era data. Despite these limitations, ~~we believe~~ statistical PSMs provide advantageous capabilities within the context of the LMR and, moreover, define a baseline to measure future progress with the development of process-based PSMs.

Here, univariate and bivariate statistical PSMs are considered,

$$y_k = \beta_{0k} + \beta_{1k}\overline{X_1'} + \epsilon_k \tag{7}$$

and

$$y_k = \beta_{0k} + \beta_{1k}\overline{X_1'} + \beta_{2k}\overline{X_2'} + \epsilon_k \tag{8}$$

where y_k are annualized observations from the k^{th} proxy time series, X_1', X_2' are anomalies of key climate variables (e.g. near-surface air temperature and precipitation) from calibration instrumental–era datasets, β_0 is the intercept and β_1, β_2 are the slopes with respect to the X_1' and X_2' independent variables respectively, and ϵ is a Gaussian random variable with zero mean and variance σ^2. The overbar in Eqs. 7 and 8 denotes time averages over annual periods as in H16, or over appropriate seasonal intervals for the seasonal PSMs. Calibration data concurrent with available proxy observations are taken at the grid point nearest the proxy location and the appropriate least-squares solution determines regression parameters $(\beta_0, \beta_1, \beta_2, \sigma)$. In this version of LMR, PSM configuration is the same for each proxy category (e.g., univariate for all coral $\delta^{18}O$, bivariate for all tree ring widths records, etc.).

With the framework described above, the regression–based approach measures the diagonal elements in matrix $\mathbf{R}$ through the variance of regression residuals, i.e. $R_k = \sigma^2$. This is a key parameter in PDA as it determines the extent to which the information provided by the proxy is weighted against prior information in the resulting reanalysis. This method provides a sound basis through which assimilated proxy records influence the reanalysis depending on the strength of their relationship to the dependent climate variables. For example, a record with a poor fit to calibration data will be characterized by larger residuals, hence larger observation error variance, and less weight in the reanalysis relative to a record that has a stronger correlation with climate variables. We note that modestly different results are obtained with different observational calibration datasets (see H16).

The calibration datasets used in this study are the NASA Goddard Institute for Space Studies (GISS) Surface Temperature Analysis (GISTEMP) (Hansen et al., 2010) version 4 for temperature, and the gridded precipitation dataset from the Global Precipitation Climatology Centre (GPCC) (Schneider et al., 2014) version 6 as the source of monthly information on moisture input over land surfaces. The use of precipitation instead of the more traditional Palmer Drought Severity Index (PDSI) to account for moisture is described in more detail in section S4 of the supplementary material.

2.4.1 Seasonality

Here we take advantage of the availability of expert information about the seasonal response to temperature for each proxy record included in the PAGES2k-2017 metadata. This information is not available in PAGES2k-2013, hence the use of PSMs calibrated on annual averages for all records in H16. Seasonality information is provided for each record as a numerical representation of a sequence of consecutive months (e.g. JJA as [6,7,8]). Seasonal PSMs are derived by using this sequence as the averaging period defining $\overline{X_1'}$ and $\overline{X_2'}$ in Eqs. 7 and 8.

Precise information on proxy seasonality is however not available for all records in the updated LMR proxy database. The proxies from ~~? (section 2.2)~~ Anderson et al. (2019), considered as a possible additional expansion of the proxy database, have not been subjected to extensive community–wide screening and vetting as with the PAGES2k-2017 proxies. In particular, seasonality information for the large number of additional tree ring records from B14 has been encoded using a simple latitudinal dependence which does not attempt to represent possible record-by-record diversity ~~(see ?)~~(see Anderson et al., 2019). This lack of expert-informed seasonality motivates an objective alternative to the metadata seasonality information for calibrating tree ring width (TRW) forward models. We consider several potential seasonal periods, perform a regression over each possible season, and identify the linear relationship providing the best fit to proxy values, as defined by the maximum value of the adjusted R^2, a goodness-of-fit measure defined as (Goldberger, 1964, p. 217):

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(N-1)}{N-M-1}\right]. \tag{9}$$

Here, R^2 is the variance explained by the linear model, N is the sample size and M is the number of predictors in the model. The adjusted R^2 penalizes complexity (i.e. the number of predictors) of the model in such a way that values characterizing a more complex model will increase only if the additional predictors improve the fit more than would be expected by chance. Test periods considered include, in addition to the seasonal response in the proxy metadata (if available), the calendar year, boreal summer (JJA) and boreal winter (DJF), and extended Spring and Fall growing seasons (MAMJJA, JJASON for NH trees, SONDJF, DJFMAM for SH trees) to account for ecosystem–dependent variations in tree growth shifted toward the earlier or later parts of the warm season (see, e.g., Sano et al., 2009; D'Arrigo et al., 2005). With this test set of seasonal responses, the dominant sensitivity of some TRW chronologies to winter temperature (D'Arrigo et al., 2012) is included, as well as the winter and spring precipitation sensitivities characterizing some tree species (see, e.g., Stahle et al., 2009; Touchan et al., 2003). The latter point is germane to the calibration of seasonal TRW models using precipitation as a predictor (see next section).

2.4.2 Tree ring width sensitivity to temperature and moisture

Proxy number is strongly dominated by TRW records in the LMR proxy database, particularly with the addition of chronologies from B14. Furthermore, these records have not been screened on temperature, which opens the opportunity to measure moisture sensitivity through the regression framework. The addition of an explanatory variable increases the potential for overfitting, and our framework is designed to measure that using the 25% of proxies withheld from assimilation, for which we can measure reconstruction errors and compare results with proxies that were assimilated.

Two methods are considered, both adding a dependence to moisture input (as represented here with precipitation). The first maintains the univariate approach (Eq. 7) but considers linear PSMs calibrated against either temperature or precipitation. For each TRW record, distinct regressions with either ~~variables~~ variable are established and the model providing the best fit to proxy data is selected. Following a common practice in dendroclimatology, this approach determines whether the record is predominantly temperature or moisture limited (see, e.g., St. George, 2014). Similar univariate "temperature or moisture" models (abbreviated as "TorM" hereafter) are successfully used in Steiger et al. (2018). The second method consists of simultaneously factoring both temperature and moisture sensitivities through the bivariate relationship expressed in Eq. 8.

Seasonal univariate TorM and bivariate TRW models are considered, with distinct sets of models calibrated using proxy seasonality either from the proxy metadata or objectively-derived during calibration. This selection has important implications for the representation of the proxy seasonal response to moisture in particular. For the proxy metadata, seasonality for moisture is assumed to be identical to temperature as this is the only information available, whereas the objective approach allows for independent encoding of seasonal responses to temperature and moisture. For TorM models, the objective seasonality for univariate moisture models is independent of temperature as it is determined solely from the fit to precipitation data. For bivariate PSMs, all possible combinations of seasonal responses specified independently for temperature and moisture are considered and the combination providing the best fit is selected. With such flexibility, TRW models with objectively-derived seasonality are expected to provide a more realistic representation of the significant variability in seasonal responses to moisture characterizing TRW records (see, e.g., St. George et al., 2010). We note that this approach is similar to the methodology used to calibrate the VS-Lite model (Tolwinski-Ward et al., 2011) in that ~~local~~ grid cell temperature and precipitation data are used to determine site-specific growth seasons and seasonally-dependent temperature and moisture growth parameters.

An examination of PSM characteristics, summarized here, with more detail provided in appendix A, confirms that proxies are represented more accurately by seasonal models, particularly for tree-ring wood density and width records (see Table A1). Moreover, more-accurate fits to TRW data are obtained when proxy seasonal responses are determined objectively during model calibration. Finally, the addition of moisture input as a climate driver in TRW modeling proves most beneficial when implemented in bivariate models (see Table A2). These findings serve as the basis for defining a PDA configuration used for the reconstruction described in the next section.

3 The updated reanalysis

We present a comparison between the updated reanalysis described by the method in the previous section with the LMR prototype described in H16[2]. Specifically, the updated reanalysis consists of all proxy records in the expanded database, using objectively-derived seasonal PSMs, with a bivariate formulation for all TRW proxies and univariate for all other proxy types. Covariance localization is applied with a 25000 km cut-off radius (see section 4.2 for more details). In the next section we identify the sources of improvement that contribute to the increase in skill of the updated reconstruction. Results are evaluated against various twentieth century instrumental data and reanalyses, as well as verification performed in proxy space, using the Pearson correlation coefficient and the coefficient of efficiency (CE) (Nash and Sutcliffe, 1970).

Figure ~~3~~2a shows a comparison of reconstructed ~~global–mean~~ global-mean temperature (GMT) between the prototype and updated reanalyses over the entire Common Era. Similar features are observed in the ensemble mean from both reanalyses,

[2] We use the experiment included in Figure 12 of H16, with PSMs calibrated using GISTEMP. Moreover, we use this configuration to generate a reconstruction of the Palmer Drought Severity Index (PDSI), which was not included in H16.

namely the cooling trend over most of the Common Era, followed by the ~~industrial–era~~ industrial-era warming. Superimposed on these main trends, significant multidecadal to multicentennial variability characterize both reanalyses, including a cool period prior to the industrial warming, consistent with the Little Ice Age (LIA). Noticeable differences also exist between the reanalyses, ~~such as the absence~~ most noticeably the absence in the updated LMR of the relatively warm period during 870–1000CE, representing the Medieval Climate anomaly (MCA)~~, in the updated LMR. Weaker decadal-scale temperature variations largely characterize the updated reconstruction, particularly after 1000CE when more proxies are assimilated. Cooler conditions also prevail during the LIA~~. Also, warmer conditions prevail in the prototype ~~compared to the updated reanalysis, although verification against instrumental–era temperature analyses provides evidence that the prototype reanalysis is too cold~~ during the second half of the fifteenth century, i.e. Spörer Minimum, while cooler conditions occur during the early part of the instrumental period ~~.~~

~~GMT verification results of the LMR ensemble mean against various instrumental temperature products are shown in Figs. 3c and d for the prototype and updated reanalyses respectively. Noticeably higher verification scores characterize the updated LMR, including a 10% increase in CE relative to the average of observations-based~~ in the prototype compared to the updated reanalysis. We note however that verification against instrumental–era temperature analyses (~~“consensus”), and an increase in CE in the verification of the detrended GMT (over 1880-2000CE) from 0.32 in the prototype to 0.60 in the updated reanalysis (see Table 1). A narrower ensemble is also a characteristic of the new reanalysis, indicating a decrease in reconstructed GMT uncertainty. This represents an improvement over the prototype from the point of view of ensemble calibration, defined as the relationship between the mean squared error in the ensemble-mean and ensemble variance. The two should closely match for a well-calibrated ensemble. The GMT ensemble from the prototype is found to be overdispersive, meaning that the ensemble variance is much larger than the error in the ensemble mean, whereas the GMT ensemble in the updated LMR is found to be well-calibrated (see section 4.2 for more details)~~discussed later in the section) provides evidence that the prototype reanalysis is too cold during that period.

~~Having access to ensembles can provide further insight into uncertaintyin the reanalyses~~Ensembles provide access to useful diagnostics regarding reconstruction uncertainty. It can be shown mathematically that the assimilation of observations ~~invariably leads to a reduction in the variance characterizing~~ monotonically reduces the variance of the posterior ensemble compared to the prior. The ratio of ensemble ~~variances~~ variance of the posterior (reanalysis) to the prior is a measure of the information provided by the assimilated proxies. Figure ~~3~~2b shows the temporal evolution of $1 - \mathrm{Var}[\mathbf{x}_a]/\mathrm{Var}[\mathbf{x}_b]$, so that a value of zero indicates no influence from proxies and one implies that all error has been removed. In the early part of the Common Era, when few proxy data are available, a variance ~~decrease of only 10% occurs~~ decreases of only 10-15% occur in the prototype compared to ~~20~~15-20% for the updated reanalysis. The influence of proxies gradually increases after 450CE, at similar rates in both reanalyses. The ~~reduction in variance is more pronounced in the updated LMR beginning at 1500CE~~reductions in variance are roughly similar in both reanalyses until 1700CE, corresponding to the period with a significantly larger number of proxies in the updated database (see Fig. 1). The largest reduction, ~~60~~68% in the prototype compared to ~~80% to 90%~~ 78% in the updated reanalysis, is found during the 20th century when the most proxies are available~~. The uncertainty characterizing~~

~~the updated reanalysis is therefore reduced overall compared to~~ , which underscores the importance of the expanded proxy database in LMR.

To gain further perspective on our results, we compare the reconstructed Northern Hemispheric average 2m air temperature from the prototype and updated reanalyses with other reconstructions quoted in the Intergovernmental Panel on Climate Change Fourth and Fifth Assessment Reports (IPCC AR4 and AR5) (Fig. 2c). Here we restrict the comparison to reconstructions covering the entire hemisphere and with a temporal coverage extending to at least 1980. A 30-year low-pass Butterworth filter is applied on all results to highlight variability at the lower frequencies. The comparison shows that most reconstructions from other studies are within the bounds of the LMR ensemble most of the time, indicating a general agreement between the different products, at least within the bounds of uncertainty as defined from LMR. As with GMT, periods with the largest differences correspond to the MCA (870–1000CE), the Spörer Minimum (1450–1550CE) and the latter part of the nineteenth century. First, the reconstructed colder temperatures during the medieval period are in contrast with the prototype LMR and other reconstructions. However, this period is one where the various reconstructions exhibit significant disagreement. This sensitivity to the proxy network and reconstruction method underscores the inherent ambiguities in defining this feature, as discussed in Diaz et al. (2011). With respect to LMR, differences between the update and prototype are primarily rooted in the change from PAGES 2k Consortium (2013) to the more recent PAGES 2k Consortium (2017) proxy data. A distinctly warmer medieval period isn't a prominent feature of the new collection, as indicated by the global temperature composites presented in PAGES 2k Consortium (2017). Second, the colder temperatures in the updated reanalysis during the Spörer Minimum is in better agreement with the majority of reconstructions in other studies, with respect to both the magnitude and trend of temperature anomalies. The LMR prototype appears as a warm outlier for this 100-year period. In contrast, the ~~prototype, which underscores the importance of the expanded proxy database in LMR~~prototype LMR appears as a cold outlier during the latter part of the nineteenth and early twentieth centuries. During that period, the updated reanalysis is in better agreement with results from other authors, in particular with the borehole temperature reconstruction by Pollack and Smerdon (2004).

GMT verification results of the LMR ensemble mean against various instrumental temperature products are shown in Figs. 3a and b for the prototype and updated reanalyses respectively. Noticeably higher verification scores characterize the updated LMR, including a 9% increase in CE relative to the average of observations-based temperature analyses ("consensus"), and an increase in CE in the verification of the detrended GMT (over 1880-2000CE) from 0.32 in the prototype to 0.59 in the updated reanalysis (see Table 1). Spatial verification is provided by comparing the LMR gridded 2m air temperature field against the Berkeley Earth (BE) instrumental–era temperature analysis (Rohde et al., 2013) (Fig. 4). BE is chosen as the verification reference as it ~~has not been~~ is not used to calibrate the PSMs, and provides the most complete spatial coverage compared to other instrumental products. The updated temperature reconstruction is largely improved compared to the prototype over large areas, including the tropical Pacific, northern Atlantic, western North America, northern Europe, ~~eastern and~~ central Asia, Oceania, and over portions of the Pacific sector of the Southern Ocean. The improvement is reflected in both correlation and CE scores, indicating improved timing and amplitude in reconstructed temperature variability. Exceptions are found over parts of the southern Atlantic and Indian oceans, although the decrease in skill is generally more modest compared to the magnitude of improvements elsewhere.

~~Hydroclimate verification is defined by a comparison of the reconstructed Palmer Drought Severity Index (PDSI) with the Dai (2011) product. We note here that this verification is entirely independent as TRW forward models were calibrated on precipitation and not on PDSI as in Steiger et al. (2018). PDSI is also improved in the updated reanalysis compared to the prototype (Fig. ??). Enhanced skill is particularly noticeable over North America, Europe and Asia, where most of the additional TRW records are located. An exception is the decreased skill found over a narrow band along the Siberian Taiga.~~

~~Finally~~ Next we verify a climate variable away from the surface, the 500 hPa geopotential height field, against the corresponding field from NOAA's twentieth century reanalysis (20CR-v2, Compo et al., 2011) (Fig. 5). Once again we find the largest improvements over extratropical continental locations, and over the Arctic. We note similar improvements are found ~~in the tropics and~~ over the Northern Hemisphere mid-latitudes when verified against the ERA-20C reanalysis (Poli et al., 2016) (not shown)~~. However~~; however, over the Northern ~~and Southern~~ Hemisphere high latitudes verification against ERA-20C is worse, which underscores significant differences between twentieth-century reanalyses in these data-sparse regions.

Table 1 summarizes the verification results discussed above through globally averaged verification scores. The table also includes verification results of reconstructed PDSI, not discussed above. A more detailed analysis for this variable is reserved for section 4.3, where the role of additional proxy records is discussed. Improvements in the updated reanalyses are evident for all reconstructed variables, particularly with respect to the CE score, which is sensitive to bias and amplitude in interannual variability. These skill improvements suggest significant positive impact from the updated tropical coral proxies and ~~the addition of a large number of~~ tree-ring proxies at higher latitudes. Furthermore, we anticipate that generalizing PSMs to accounting for seasonality and moisture sensitivity for TRW proxies also contribute to the improvements.

We consider now an independent evaluation of the reconstructions in proxy-space using proxies withheld from assimilation. Proxy time series estimated (forward-modeled) from the posterior (i.e. the reconstructions) are compared to the actual proxy observations and various skill metrics are evaluated. Verification of proxy estimates obtained from the uninformed climate-model prior serve as a reference for comparison. Specifically, we use the change in CE between the posterior proxy estimates and estimates obtained from the prior, ΔCE = (CE$_{posterior}$ - CE$_{prior}$). Values are compiled from all proxy records withheld from assimilation, and the following summary scores are considered: the fraction of all proxy records which are characterized by a positive ΔCE (i.e. proxy records more accurately represented in the posterior than in the prior), and the median of the ΔCE distribution compiled over all proxy time series. These provide global summary measures of how reanalyses skill differs from the prior. An additional discriminating factor on the quality of the reanalysis is "ensemble calibration" as defined by (Murphy, 1988),

$$ECR = \left[\frac{1}{N-1}\sum_{n=1}^{N}(v_n - \overline{x_n})^2\right]\left[\frac{1}{N-1}\sum_{n=1}^{N}(\sigma_{x,n}^2 + \sigma_{v,n}^2)\right]^{-1}, \tag{10}$$

where the numerator is the mean square error (MSE) of the analysis ensemble mean with respect to verification data v (i.e. the proxies), and the denominator is the innovation variance: the sum of the analysis ensemble variance σ_x^2 and the error variance σ_v^2 characterizing the verification data. Here we apply Eq. 10 to proxy time series so the error variance σ_v^2 corresponds to the R_k terms in Eqs. 4. The ECR ratio expresses the degree to which the ensemble predicts the distribution of observations. A well-calibrated ensemble exhibits an approximate agreement between the ensemble variance and the ensemble–mean MSE, i.e. ECR $\approx$ 1.0, while an overdispersive ensemble has variance larger than the ensemble-mean MSE (ECR $<$ 1.0), and an underdispersive ensemble is diagnosed when its variance is smaller than the ensemble–mean MSE (ECR $>$1.0). Proxy verification results are shown in Table 2, over different periods of the Common Era. Significantly reduced skill characterizes the earliest period of the Common Era, followed by a continuous increase over time in all verification metrics considered, for both LMR reanalyses. We also note that reanalysis ensembles are generally well-calibrated throughout the Common Era, indicating that respective uncertainties remain consistent with mean errors (i.e. reliable ensembles). Although verification data is not identical between prototype and updated reanalyses, we also note that the increase in skill is more pronounced in the updated reanalysis, particularly from 1000CE onward. These results provide further evidence of a more skillful updated LMR. In the following section we systematically evaluate improvements from ~~these~~ various sources.

4 Sources of improvement

In this section, we ~~turn our attention to the identification of the~~ identify the sources of reanalysis improvement. Results from multiple reconstruction experiments are presented, designed to quantify the impact of PSM formulation, the ~~assimilation of various proxy data sets and the~~ role of covariance localization and the assimilation of additional proxies.

4.1 Proxy system models

The different PSM configurations described in section 2.4 are used in a series of reconstruction experiments ~~. In order to isolate PSM improvements, we first use~~ using PAGES2k-2017 proxies exclusively~~, as they represent the community standard in Common Era proxy information, and~~ . We note that these records have well-defined seasonal metadata.

The impact of seasonal PSMs is first considered with three experiments performed using univariate temperature regression models for: (1) annual-mean calibration; (2) seasonality defined by expert metadata; and (3) objectively determined seasonality. Performance is again measured by correlation and CE scores with verification against the Berkeley Earth analysis. Relative to reconstructions with annual-mean PSMs (Figs. 6a and b), the reconstructions with seasonal PSMs (Figs. 6c–f) show improvements in both measures over nearly the entire globe (Figs. 6g–j). Results show a larger improvement for CE (Figs. 6h and j) compared to correlation (Figs. 6g and i), reflecting improvement in both the amplitude of temperature variability and bias. Noteworthy improvements are found in regions with large numbers of tree-ring proxies, such as the western United States, the region around and including Alaska, Northern Canada and western Arctic ocean, over Scandinavia and Norwegian Sea, central Asia and over the Southern Pacific west of the Antarctic Peninsula (see Fig. 6h). Comparing the differences of correlations and CE in Figs. 6i and j to those shown in Figs. 6g and h reveals that PSMs with objectively-derived seasonality contribute positively to skill for the aforementioned regions, especially where tree ring width records are most abundant (e.g. North America and Asia).

We turn now to the impact of moisture on seasonal TRW PSMs on the reconstructions. Since objectively defined seasonality performs best (i.e., Figs. 6e and f), reconstructions generated with univariate PSMs are used as the reference for measuring skill improvements for modeling TRW records as univariate in either temperature or moisture (abbreviated as “TorM”) (Figs. 7c and d) and for bivariate “temperature and moisture” PSMs (Figs. 7e and f). Improvement over univariate PSMs is apparent for the bivariate approach compared with the univariate “TorM” approach (cf. Fig. 7 panels g,h with i,j, respectively). In the bivariate approach regions such as western North America and central Asia, where most of the TRW records are found, improve the most in CE, but also over Australia, likely in response to the improved modeling of TRW records in New Zealand and Tasmania. Improvements are also noticeable, through teleconnections with proxy locations in the central Atlantic and southern India Oceans, and over the eastern North Pacific Ocean. A decrease in skill is present over the mid-latitude Pacific ocean, but this is smaller in magnitude compared with skill enhancements elsewhere.

Verification of GMT for reconstructions using seasonal PSMs (~~Fig. ??~~Table 3) yields a similar interpretation to the spatial verification results. Compared to the consensus of ~~instrumental–era~~ instrumental-era products, we find that the 20th century trend in GMT is overestimated with the PAGES2k-2017 proxy data set if univariate PSMs are used. This is particularly the case with annual PSMs. Better agreement is obtained when seasonal bivariate PSMs are used to model TRW proxies. The representation of GMT interannual variability as measured by verification of the detrended GMT is also improved with seasonal PSMs, particularly for the CE metric. Similar to spatial verification results, PSMs with objectively-derived seasonality and bivariate TRW modeling have GMT reconstructions with consistently higher skill scores.

We recognize that the previous evaluation relies on comparisons with observation-based products covering the same time period as the data used to calibrate the statistical PSMs. To test the sensitivity of the results to the calibration period, we conduct additional independent instrumental–era calibration–validation experiments where PSMs are calibrated over a subset of the instrumental-era period and reconstructions are evaluated with data not used in calibration. ~~Moreover, we perform an independent evaluation of reconstructions in proxy-space using proxies withheld from assimilation, for both the calibration and pre-calibration periods. These results, described in sections~~ Results from these experiments, described in section S3 ~~and S4~~ in the supplementary material, confirm the main results and conclusions drawn here on the superiority of seasonal PSMs relative to those calibrated with annual averages, and the use of bivariate models for TRW proxies. ~~Building upon this improved proxy modeling, we turn now to~~

We now examine results from an evaluation performed in proxy-space using proxies withheld from assimilation as in section 3. Results for both the PSM calibration and pre-calibration periods are shown in Table 4. Differences among the various experiments suggest the superiority of the seasonal (with objective seasonality) PSMs as skill scores consistently rank among the highest among all experiments, for both calibration and pre-calibration periods. The reconstruction using univariate annual PSMs show the weakest verification statistics, confirming the verification based on instrumental-era analyses. Finally, use of bivariate seasonal PSMs for TRW records is also suggested from proxy validation results, as larger correlations and ΔCE are obtained with this configuration.

4.2 Covariance localization

One approach to managing sampling error in ensemble data assimilation is through spatial covariance localization. Localization is applied to minimize the adverse impact of spurious covariances at large distances from a proxy location, which results from sample error in finite ensembles (Hamill et al., 2001). If localization is not applied, spurious covariances allow proxies to affect remote locations, which adversely affects the quality of the ~~role of increasing the number of proxy measurements on reconstruction skill .~~ analysis. On the other hand, too-short localization length scales reduces the useful information that can be derived from the proxies. Therefore a balance is sought between minimizing sampling noise versus retaining useful proxy information.

We use the Gaspari-Cohn (Gaspari and Cohn, 1999) fifth-order polynomial with a specified cut-off radius for the localization function (w_{loc} in (4)). See section S5 in the supplementary material for information on the characteristics of w_{loc}. A series of reconstructions are performed with a wide range of localization length scales. As with previous experiments, 51 Monte-Carlo realizations are carried out, each with 100 ensemble members assimilating 75% of proxy records. Results from the instrumental-era verification scores previously described are summarized in Table 5. We observe that the GMT trend is underestimated and verification scores are significantly reduced when "too-small" localization radii are used, indicating the information on temperature provided by some proxy records is not properly incorporated in the reanalysis. In contrast, the trend is overestimated and verification cores are generally reduced without covariance localization. This is particularly the case for the CE score for the detrended GMT, sensitive to the amplitude in interannual variability. This skill measure is maximized for a localization radii within the 15000 to 25000 km range. A localization radius at the upper end of this range (25000 km) is preferable, as results from the other verification scores suggest that a skillful reconstruction is obtained with this covariance localization configuration. We note that the optimal localization radius depends on a number of factors, such as ensemble size, the observation network and observation error characteristics.

4.3 Proxy data sets

Here we explore the ~~role of the number of assimilated proxies on reanalysis verification.~~ impact of adding the large number of proxies from Anderson et al. (2019) (hereafter A19), which include the tree ring width chronologies from Breitenmoser et al. (2014) (hereafter B14), not strictly screened for climate sensitivity in contrast to the PAGES2k collection. Figure 8 shows the spatial and temporal distributions of the B14 records, which reveals enhanced coverage over eastern North America, southern Europe, boreal Eurasia and southern South America. Other additions, totaling 94 records, provide additional records in the Tropics (23 coral records), and an enhanced number of ice core records concentrated over Greenland and eastern Canadian Arctic (37 records) and Antarctica (26 records in West Antarctica and Drönning Maud Land). A few lower latitude ice core records (6 records) are also added in the Peruvian Andes and Tibetan Plateau, along with two higher latitude lake core records. From a temporal perspective, the addition of the B14 tree ring width records contributes a notable number of additional proxies back to 1000CE, more than double the number of records available for assimilation from 1500CE onward, up to a fourfold increase during the nineteenth and twentieth centuries.

In order to measure ~~this~~ the impact with the best configuration, the reconstruction experiments reported in this section are carried out using seasonal PSMs with ~~objectively-derived~~ objectively derived seasonality for all records, with a bivariate formulation on temperature and precipitation for all TRW proxies, and univariate on temperature for all other proxies. The baseline reconstruction uses the PAGES2k-2017 proxies (as in ~~the previous section~~ section 3), which we compare to ~~experiments~~ results first obtained with the addition of the B14 TRW records, and ~~then~~ finally with the further addition of the coral, ice and lake core records from ~~?.~~A19 (i.e. the full proxy database). Other trial reconstructions performed with the vastly expanded proxy network, not reported here, have shown that a well-calibrated GMT ensemble is obtained with a covariance localization cut-off radius of 25000 km. Next, we compare reconstruction results from this configuration to the baseline reanalysis.

Differences in correlation and CE associated with the addition of the B14 collection over the PAGES2k-2017 proxies show skill improvements in temperature reconstructions over the continental United States and Mexico, Europe, and the southern edge of the Tibetan Plateau ~~, where the additional records provide enhanced coverage~~ (cf. Figs. 9g and h). Through the influence of significant spatial covariances with the added records, assimilation of the additional TRW records also leads to improved temperature skill over remote areas of the mid-latitude Pacific ~~, northern Atlantic and Indian oceans.~~

and northern Atlantic oceans. The addition of records described in ~~?~~ A19 has minimal additional impact overall, with the exception of modest increases in correlation and CE over Greenland (see see Figs. 9i and j). ~~Results for GMT show that the 20th century trends from the prototype and the three experiments described here are within the uncertainty of the instrumental products (Fig. ??) . However, the consensus trend (defined as the mean of trends from all instrumental analyses) is best reproduced with the most comprehensive proxy network. Skill metrics for the detrended GMT are higher for all of the updates to the proxy database compared to the prototype configuration, with a slight improvement for the reconstruction using the PAGES2k-2017 and B14 records; however, differences between results with the various subsets of the expanded proxy database are not statistically significant at the 95% confidence level.~~

Hydroclimate verification is defined by a comparison of the reconstructed Palmer Drought Severity Index (PDSI) with the Dai (2011) product. We note here that this verification is entirely independent as TRW forward models were calibrated on precipitation and not on PDSI as in Steiger et al. (2018). A comparison of the reconstructed PDSI between the prototype~~and experiments discussed in this section (Fig.10) indicate that the increase in skill shown in Fig. ?? derives from the combined influence of the updated PAGES2k-2017 records and the~~ , the updated reanalysis of section 3 and a reconstruction carried out with the B14 ~~TRW records.With bivariate TRW forward models, enhanced skill is found predominantly over the western United States and Asia, as well as~~ TRW records and the additional coral, ice and lake core records (i.e. the full database) is shown in Figure 10. The PDSI is slightly improved in the updated reanalysis compared to the prototype (Figs.10g and h). Enhanced skill is noticeable over western North America, and over eastern Europe ~~(Figs. 10g and h)~~and Asia to a lesser degree. Decreased skill is found over the central plains of North America and along a narrow band along the Siberian Taiga. The impact of adding the ~~B14 TRW~~ Anderson et al. (2019) records is mostly found over the eastern part of the United states and over western Europe (Figs.10i and j). Finally, we note that ~~the~~ this impact is due entirely to the B14 TRW records, as the additional coral, ice and lake core records from ~~?~~ A19 do not significantly affect the PDSI reconstruction skill (~~Figs.10e and f are nearly identical to Figs. ??c and d~~from results of additional reconstruction experiments carried out to isolate this impact, not shown).

4.4 ~~Covariance localization~~

~~A key question with ensemble data assimilation is whether spatial covariance localization should be applied, and if so, with what length scale (i. e. cut-off distance). Localization is applied to minimize the adverse impact of spurious covariances at large distances from a proxy location, which results from sample error in finite ensembles (Hamill et al., 2001). If localization is not applied, spurious covariances allow proxies to affect remote locations, which adversely affects the quality of the analysis. On the other hand, too-short localization length scales reduces the useful information that can be derived from the proxies. Therefore a balance is sought between minimizing sampling noise versus retaining useful proxy information.~~

~~We use the Gaspari-Cohn fifth-order polynomial with a specified cut-off radius (Gaspari and Cohn, 1999) for the localization function w_{loc} in (4). A series of reconstructions are performed with localization radii ranging from 5000 km to 25000 km. As with previous experiments~~ Examining the differences between reconstructions over the entire Common Era (Fig. 11), ~~51 Monte-Carlo realizations are carried out, each with 100 ensemble members assimilating 75% of proxy records in the expanded LMR database. In addition to the verification scores previously described, an additional discriminating factor on the quality of the reanalysis is "ensemble calibration" as defined by (Murphy, 1988),~~

$$ECR = \left[\frac{1}{N-1}\sum_{n=1}^{N}(v_n - \overline{x_n})^2\right]\left[\frac{1}{N-1}\sum_{n=1}^{N}(\sigma_{x,n}^2 + \sigma_{v,n}^2)\right]^{-1},$$

~~where the numerator is the mean square error (MSE) of the analysis ensemble mean with respect to a verification data set v, and the denominator is~~ we see a significantly modified Northern Hemisphere temperature (NHMT) resulting from the ~~innovation variance: the sum of the analysis ensemble variance σ_x^2 and the error variance σ_v^2 characterizing the verification data. Here we apply Eq. 10 to the GMT ensembles from each Monte-Carlo realization, where the verification data corresponds to the consensus of instrumental–era data sets as before. The error variance σ_v^2 is estimated from the deviations of the constituent sample data (the instrumental datasets) from their mean (i. e. the consensus), and the N "observations" are taken from the time series of annual analysis and verification data points over the 20th century. The ECR ratio expresses the degree to which the ensemble predicts the distribution of observations. A well-calibrated ensemble exhibits an approximate agreement between the ensemble variance and the ensemble–mean MSE, i. e. ECR ≈ 1.0, while an overdispersive ensemble has variance larger than the ensemble-mean MSE (ECR $<$ 1.0), and an underdispersive ensemble is diagnosed when its variance is smaller than the ensemble–mean MSE (ECR $>$1.0).~~ assimilation of the additional proxies. A generally warmer NHMT is obtained throughout the Common Era, but most significantly during the LIA, worsening the agreement with reconstructions from other studies shown in Figure 2. A noticeable loss of variability is observed, confirmed by comparing spectra from both experiments (Fig. 11c). This loss of variability in the reconstruction using all proxies occurs at nearly all scales, underlining an adverse impact from assimilating B14 tree ring width proxies.

~~An ECR value is obtained for each Monte-Carlo reconstruction, and the average is taken over the 51 values. Results indicate that GMT ensembles tend to be underdispersive without covariance localization (Table 5). An excessive reduction in posterior ensemble spread, compared to the mean error, can arise from two factors: too-small observation error variance for some records, leading to overestimated weight of these records in the reanalysis, combined with sample error in the ensemble-estimated covariances, which artificially reduces ensemble variance over the entire state vector. In contrast, ensembles are overdispersive~~

~~when "too-small" localization radii are used, indicating the information on global-mean temperature provided by some proxy records is not properly incorporated in the reanalysis. The ECR closest to one (i.e. well-calibrated ensemble) is obtained when covariance localization is applied with a cut-off radius of about 25000 km. This result, along with those from the common verification scores (see Fig. S4) , suggest that a skillful and reliable reconstruction is obtained with this covariance localization configuration.~~

~~We note that the optimal localization radius depends on a number of factors, such as ensemble size, the observation network and observation error characteristics. For example, as seen in Table 5, an overdispersive GMT ensemble is obtained when proxies from PAGES2k-2013 are assimilated in the absence of covariance localization, due to the much smaller number of proxies. This is in contrast with the underdispersive ensemble resulting from the~~ We now turn to verification in proxy space, which is the only source available prior to the instrumental period. Proxy estimates from reanalyses (estimated using the appropriate PSM) are compared directly to proxy observations. Here, reanalysis skill is assessed using independent (the 25% withheld from assimilation) proxies. We further restrict our analysis to verification against tree ring wood density proxies, as they are among the most reliable recorders of temperature in our database, as evidenced by the generally better fits to calibration temperature data obtained when calibrating the univariate PSMs. Also, these proxies provide good temporal coverage of the latter portion of the LIA into the industrial period as shown in Figure 1. The results, presented in Table 6, show distinctly larger skill scores for the experiment using PAGES2k-2017 proxies only compared to when all proxies are assimilated. Improved skill is observed for both periods of interest. Results from a third reconstruction experiment are also presented, where only a small fraction of B14 records are assimilated (B14 subset experiment in Table 6). A total of 188 records (out of the 2156 available) have been selected on the basis of their strong relationship to calibration temperature and precipitation data as determined from the correlation coefficient characterizing bivariate PSMs. Records with a calibration correlation above 0.6 are found to be located for the most part over the United States. Proxy verification results indicate an increase in skill in the representation of tree ring wood density proxies, as indicated by skill metric values only slightly lower than in the PAGES2k-2017 experiment. Spatial verification of temperature and PDSI (not shown) also suggest that some of the skill enhancements shown in Figure 10i and j are retained even when this small fraction of the B14 records is considered. This suggests that the issues with the assimilation of the ~~significantly larger number of proxies from the updated proxy database used here.~~ B14 records identified above can possibly be mitigated while maintaining some of the skill they provide toward enhanced temperature and hydroclimate reconstructions in local regions. Optimal selection of these records requires further careful attention, and could serve as the basis for future efforts.

5 Concluding summary

A paleoclimate reanalysis of the Common Era has been developed using an updated data assimilation framework. Results show significant improvement over the prototype Last Millennium Reanalysis presented in Hakim et al. (2016). ~~The development of a vastly expanded~~ An updated proxy database and implementation of proxy system models (PSMs) with improved realism are shown to be key contributors to the enhanced reanalysis. ~~Upgrades~~ The main upgrade to the proxy database consist of a change from the community-standard of PAGES 2k Consortium (2013) to the more recent PAGES 2k Consortium (2017) data set, ~~complemented by~~ while the records described in ~~?, bringing a fivefold increase in the number of proxy records available for assimilation~~Anderson et al. (2019) remain available for possible future enhancements to the proxy information used in the reanalysis. Moreover, new methods to map state variables to observations extend the prototype's linear univariate models calibrated on annual-mean temperature in two key aspects: accounting for seasonal dependencies of individual proxy records, and the modeling of tree-ring-width proxies using temperature and moisture as predictors. The encoding of proxy seasonality information within PSMs has also been refined by objectively determining the characteristic seasonal response of individual records, and by decoupling the seasonality for temperature and precipitation sensitivity for tree-ring-width.

Climate field reconstructions from a series of assimilation experiments carried out with various proxy and PSM configurations have been compared to available instrumental–era observation-based analyses, revealing notable improvements not only in the reconstructed global mean temperature in general, but also in reconstructed spatial fields. More skillful tropical Pacific temperatures are obtained primarily due to the updated set of coral records in the PAGES 2k Consortium (2017) collection. Improved temperature reconstructions over continental extratropical regions are the result of the newly implemented seasonal PSMs, combined with the ~~assimilation of the large number of Breitenmoser et al. (2014)~~ forward modeling of tree-ring-width chronologies ~~, forward-modeled~~ using a bivariate temperature-moisture formulation. Improvements are reflected not only in temperature reconstructions, but also in ~~dynamical variables (~~500 hPa geopotential ~~heights) and~~ height and to some extent in hydroclimate variables such as the PDSI. ~~The introduction within LMR's proxy database~~ Lastly, the introduction of the large collection of Breitenmoser et al. (2014) tree-ring-width ~~records~~ chronologies, not screened for temperature sensitivity~~appears to be a significant factor in enabling this capability. Lastly, covariance localization, applied with a relatively large cut-off length scale, has been shown to have a positive impact , particularly in maintaining an appropriate relationship between mean error and variance in the reanalysis ensemble~~, appears to provide local skill enhancements in hydroclimate variables (e. g. PDSI over the eastern United States). However this is achieved at the expense of accuracy in the reconstruction of important features of pre-industrial climate such as the colder temperatures during the Little Ice Age. However, the generally positive impact of a simple *ad hoc* screening of the Breitenmoser et al. (2014) suggest that further improvements may be possible with a careful selection of tree-ring chronologies.

Results presented here, based upon regression PSMs, may serve as a reference for future efforts designed to assess the value of more comprehensive process-based PSMs in paleoclimate data assimilation research. Finally, we note that the version of the PDA system described ~~herein~~ here corresponds to the configuration used in the production ~~of the first publicly released~~ release of the NOAA Last Millennium Reanalysis, available at https://atmos.washington.edu/~hakim/LMR/.

Code and data availability. The code used in the production of the reanalysis is publicly available at https://github.com/modons/LMR, and data are available from https://atmos.washington.edu/~hakim/LMR/.

Appendix A: ~~DA with an appended state~~

~~For reasons of computational efficiency and flexibility we perform data assimilation with an “appended state”, where the $\mathbf{y}_e$ proxy estimates from each record are appended to the state vector $\mathbf{x}_b$:~~

$$\mathbf{x}_b = \begin{bmatrix} x_1 \\ \vdots \\ x_N \\ y_e^1 \\ \vdots \\ y_e^P \end{bmatrix},$$

~~where the $x_1 \dots x_N$ elements contain the ensemble grid point data from model variables included in the state (e.g. temperature, precipitation etc.), with N the sum of the number of variables times the number of grid points, and the $y_e^1 \dots y_e^P$ are the ensemble proxy estimates for each of the P proxy records considered. Each of the $x_1 \dots x_N$ and $y_e^1 \dots y_e^P$ elements are of dimensions $1 \times N_{ens}$, where N_{ens} is the specified size of the ensemble. Hence, $\mathbf{x}_b$ is a matrix of dimension $(N+P) \times N_{ens}$. With such an appended state, the y_e elements in Eq. 6 are updated through Eqs. 4 as any other state variables, eliminating the need to re-evaluate $\mathbf{y}_e$ with Eq. 2 once the state has been updated. This simplification is particularly attractive in the context of LMR updates discussed herein as it enables a straightforward implementation of seasonal PSMs (i.e. forward models more accurately representing the seasonal responses of individual proxy records) as discussed in section 2.4. In our implementation with an appended state and serial processing of observations, along with the reconstruction of annually-averaged states, the data assimilation procedure follows this general algorithm.~~

1. ~~The proxy estimates ($\mathbf{y}_e$) are pre-calculated using Eq. 2 with either annually- or seasonally-averaged model data as input (i.e. the $\mathbf{x}_b$ in Eq. 2).~~

2. ~~A sample ensemble of annually-averaged model states is randomly drawn from a pre-existing simulation to form the main part of the prior state vector (i.e. the $x_1 \dots x_N$ elements in Eq. 6).~~

3. ~~The pre-calculated $y_e^1 \dots y_e^P$ proxy estimates are added on to form the appended state as shown in Eq. 6. This appended state becomes the $\mathbf{x}_b$ in Eq. 1, which is decomposed in an ensemble-mean ($\overline{\mathbf{x}}_b$) and perturbations about the mean ($\mathbf{x}_b'$) as shown in Eqs. 4.~~

4. ~~Proxies forming the $\mathbf{y}$ vector are then serially processed, with the updated state, including the proxy estimates, obtained from Eqs. 4. The complete reanalysis is completed once all proxies have been assimilated.~~

~~We note here that with a configuration involving seasonal PSMs without the use of an appended state, the vector $\mathbf{x}_b$ has to include states with sufficient temporal resolution to allow the calculation of the updated seasonal $y_e^1 \dots y_e^P$ proxy estimates. In~~

~~this scenario, an additional step to the ones listed above is required, involving Eq. 2 using the appropriate seasonally-averaged updated states as input. With proxies characterized by a wide range of seasonal responses, this requirement would impose an $\mathbf{x}_b$ composed of monthly data which would greatly increase the computational cost of the reanalysis. Reanalysis results would also likely be adversely affected by the larger noise level characterizing data at shorter (i.e. monthly) timescales through its impact on ensemble estimates of prior covariances (see, e.g. Tardif et al., 2016).~~

Appendix A: Proxy system model characteristics

Features introduced in the updated LMR proxy modeling capabilities include a representation of the seasonal response to climate drivers characterizing individual proxy records (i.e. proxy seasonality), as well as proxy system models (PSMs) that include ~~moisture input~~ precipitation and temperature as driving variables for modeling tree-ring-width (TRW) records.

The first approach is to use univariate PSMs calibrated against temperature data, with proxy seasonality either defined from the available proxy metadata or derived objectively using the method described in section 2.4.1. PSM performance is compared using the Bayesian Information Criterion (BIC), defined as (Schwarz, 1978),

$$BIC = -2\ln(\hat{L}) + k\ln(n) \tag{A1}$$

where $\hat{L}$ is the maximized value of the likelihood function of the model, n is the sample size and k is the number of estimated parameters in the model. We note that the second term in Eq. A1 represents a penalty for models with a larger number of explanatory variables, i.e. a more complex model. This feature is particularly useful when comparing univariate and bivariate models. Here we use the difference in BIC values between two models $\Delta BIC = (BIC_M - BIC_{ref})$, to determine the relative accuracy of model M over a reference. The model with the lowest BIC is preferred (i.e. a better fit to the data), hence a negative ΔBIC indicates the superiority of the test model over its reference. Here, the seasonal PSMs are tested against the univariate PSMs calibrated with annually-averaged temperatures as the reference. Significant evidence of the superiority of the test model over its reference is obtained when ΔBIC < -2.0.

Table A1 presents a summary of ΔBIC results for records in each proxy category considered in LMR. The advantage of seasonal PSMs is particularly significant for tree-ring wood-density chronologies, a proxy known for its strong seasonal response (Briffa et al., 2004). Seasonal PSMs also provide improved fits to tree ring width data, although to a lesser extent compared to density records. As indicated by the larger negative ΔBIC values, models based on objectively-derived seasonal responses lead to more accurate descriptions of proxy data compared to those calibrated using metadata seasonality, even for tree ring chronologies within the community-curated PAGES2k-2017 data set. These results suggest that the objectively-derived seasonality information is noticeably different than in the metadata, particularly for tree ring records in the Breitenmoser et al. (2014) (i.e. B14) data set, but also for those in PAGES 2k Consortium (2017) (i.e. PAGES2k-2017). More details on this aspect are provided in the supplementary material. The use of objectively-defined seasonality improves upon the simple latitude-dependent relationship described in ~~?~~Anderson et al. (2019), more consistent with records from the PAGES2k-2017 data set. Apart from lake sediment records, which are also more accurately modeled with seasonal PSMs, Table A1 shows that PSMs for other proxy types are not as sensitive to seasonality. In fact, the majority of the (tropical) coral records included in the current database have metadata seasonality defined as annual already, as do the high-latitude ice core records. Note that some of these records originate from the collection described by ~~?, where seasonality~~ Anderson et al. (2019), where seasonal metadata information is generally not available. As a result, these records are assumed to be annual.

In addition to seasonal models, other improvements involve the development of PSMs ~~which~~ that add precipitation as an input variable for the modeling of TRW proxies as outlined in section 2.4.2. One approach consists of selecting the univariate models, either calibrated on temperature or moisture input, which best describe the proxy data. This "temperature or moisture" selection (abbreviated as "TorM") is performed on individual TRW records, and the resulting proportion of TRW proxies identified as temperature-sensitive is 56.4% versus 43.6% for moisture when metadata seasonality information is considered. This is compared to 36.8% temperature-sensitive versus 63.2% moisture-sensitive trees when seasonal responses are determined objectively. The latter option, leading to a larger proportion of moisture-sensitive records, is in better agreement with a comparable characterization performed by Steiger et al. (2018) on a similar set of TRW records.

A second approach consists of bivariate PSM formulation, where TRW depends on both temperature and precipitation (see Eq. 8. The ΔBIC results characterizing the univariate "TorM" and bivariate PSMs against their univariate temperature-only counterparts (as the reference) are summarized in Table A2. The negative mean ΔBIC values confirm the advantage of including moisture in TRW linear models. The evidence is more pronounced for the B14 records, perhaps not surprisingly given the larger proportion of moisture-sensitive records included in this data set. Nonetheless, the prevalent reduction in BIC for models of PAGES2k-2017 trees suggests a non-negligible response to moisture despite the screening of records for temperature. The mean positive ΔBIC characterizing the bivariate models calibrated using metadata seasonality confirm that the assumption of identical seasonal responses for temperature and moisture is problematic for modeling tree ring growth, at least with these more complex models. On the other hand, allowing distinct representations of temperature and moisture seasonal responses in bivariate PSMs, as enabled by the goodness-of-fit objective determination of these responses, leads to significantly more accurate TRW modeling compared to univariate temperature PSMs.

Author contributions. R.T. developed the code related to all the updates discussed in the paper (generation of expanded proxy database, calibration and application of the new proxy forward models), configured and performed all the reanalysis experiments, analyzed the experimental results and wrote the manuscript with G.J.H. G.J.H. proposed the general design of the updated tree-ring proxy models, contributed the Kalman update and original versions of the verification codes. W.A.P. developed several key parts of the reanalysis code. K.A.H. collected and formatted the proxy records taken from sources other than the PAGES 2k collection, re-processed the tree ring data taken from Breitenmoser et al. M.P.E. contributed to the identification of proxy records to be included/excluded from the updated proxy database and suggested key code modifications. J.E.G. provided the PAGES 2k proxy data and advice on proxy system modeling. D.M.A. supervised the collection of proxy records and helped define the content of the proxy database. E.J.S. contributed advice on proxy system modeling, statistical measures, and helped define the content of the proxy database. D.N. contributed advice on proxy system modeling. All authors contributed to the final editing of the manuscript.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors thank Nathan Steiger of the Lamont-Doherty Earth Observatory, Columbia University, for providing the PDSI data from the CCSM4 Last Millennium simulation, and Jörg Franke of the University of Bern, Switzerland, for providing the tree-ring data from the Breitenmoser et al. (2014) study. This research was supported by grants from the National Oceanic and Atmospheric Administration (grant NA14OAR4310176) and the National Science Foundation (~~grant~~ grants AGS-1304263 and AGS-1702423 to the University of Washington)~~and~~. We acknowledge the World Climate Research Programme's Working Group on Coupled ~~Modelling~~Modeling, which is responsible for CMIP. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals. CMIP5 data used in this paper may be obtained from the Earth System Grid Federation at http://esgf.llnl.gov/. Some calibration and verification data sets were provided by the NOAA/OAR/ESRL PSD, Boulder, Colorado, USA, from their website at https://www.esrl.noaa.gov/psd/.

References

Acevedo, W., Fallah, B., Reich, S., and Cubasch, U.: Assimilation of pseudo-tree-ring-width observations into an atmospheric general circulation model, Clim. Past, 13, 545–557, https://doi.org/10.5194/cp-13-545-2017, 2017.

Anderson, D. M., Tardif, R., Horlick, K., Erb, M. P., Hakim, G. J., Emile-Geay, J., Noone, D., Perkins, W. A., and Steig, E. J.: Additions to the Last Millennium Reanalysis multi-proxy database, Data Science Journal, 18, https://doi.org/10.5334/dsj-2019-002, 2019.

Bhend, J., Franke, J., Folini, D., Wild, M., and Brönnimann, S.: An ensemble-based approach to climate reconstructions, Clim. Past, 8, 963–976, https://doi.org/10.5194/cp-8-963-2012, 2012.

Breitenmoser, P., Brönnimann, S., and Frank, D.: Forward modelling of tree-ring width and comparison with a global network of tree-ring chronologies, Clim. Past, 10, 437–449, https://doi.org/10.5194/cp-10-437-2014, 2014.

Briffa, K. R., Osborn, T. J., and Schweingruber, F. H.: Large-scale temperature inferences from tree rings: a review, Glob. Planet. Change, 40, 11–26, https://doi.org/10.1016/S0921-8181(03)00095-X, 2004.

Cobb, K. M., Charles, C. D., Cheng, H., and Edwards, R. L.: El Niño/Southern Oscillation and tropical Pacific climate during the last millennium, Nature, 424, 271–276, https://doi.org/10.1038/nature01779, 2003.

Compo, G., Whitaker, J., Sardeshmukh, P., Matsui, N., Allan, R., Yin, A., Gleason, B., Vose, R., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R., Grant, A., Groisman, P., Jones, P., Kruk, M., Kruger, A., Marshall, G., Maugeri, M., Mok, H., Nordli, Ø., Ross, T., Trigo, R., Wang, X., Woodruff, S., and Worley, S.: The twentieth century reanalysis project, Quarterly Journal of the Royal Meteorological Society, 137, 1–28, https://doi.org/10.1002/qj.776, 2011.

Dai, A.: Characteristics and trends in various forms of the palmer drought severity index during 1900–2008, Journal of Geophysical Research: Atmospheres, 116, 1–26, https://doi.org/10.1029/2010JD015541, 2011.

D'Arrigo, R., Mashig, E., Frank, D., Wilson, R., and Jacoby, G.: Temperature variability over the past millennium inferred from Northwestern Alaska tree rings, Clim. Dyn., 44, 227–236, https://doi.org/10.1007/s00382-004-0502-1, 2005.

D'Arrigo, R., Anchukaitis, K. J., Buckley, B., Cook, E., and Wilson, R.: Regional climatic and North Atlantic Oscillation signatures in West Virginia red cedar over the past millennium, Global and Planetary Change, 84-85, 8–13, https://doi.org/10.1016/j.gloplacha.2011.07.003, 2012.

Dee, S., Emile-Geay, J., Evans, M. N., Allam, A., Steig, E. J., and Thompson, D. M.: PRYSM: An open-source framework for PRoxY System Modeling, with applications to oxygen-isotope systems, Journal of Advances in Modeling Earth Systems, 7, 1220–1247, https://doi.org/10.1002/2015MS000447, 2015.

Dee, S., Steiger, N. J., Emile-Geay, J., and Hakim, Gregory J.: The utility of proxy system modeling in estimating climate states over the Common Era, Journal of Advances in Modeling Earth Systems, 8, 1164–1179, https://doi.org/10.1002/2016MS000677, 2016.

Diaz, H. F., Trigo, R., Hughes, M. K., Mann, M. E., Xoplaki, E., and Barriopedro, D.: Spatial and temporal characteristics of the climate in medieval times revisited, Bulletin of the American Meteorological Society, 99, 1487–1500, https://doi.org/10.1175/BAMS-D-10-05003.1, 2011.

Emile-Geay, J., Cobb, K. M., Mann, M. E., and Wittenberg, A. T.: Estimating central equatorial Pacific SST variability over the past millennium. Part I: Methodology and validation, Journal of Climate, 26, 2302–2328, 2013.

Evans, M. N., Tolwinski-Ward, S. E., Thompson, D. M., and Anchukaitis, K. J.: Applications of proxy system modeling in high resolution paleoclimatology, Quaternary Science Reviews, 76, 16–28, 2013.

Evensen, G.: The ensemble Kalman filter: Theoretical formulation and practical implementation, Ocean Dynamics, 53, 343–367, 2003.

Franke, J., Brönnimann, S., Bhend, J., and Brugnara, Y.: A monthly global paleo-reanalysis of the atmosphere from 1600 to 2005 for studying past climatic variations, Sci. Data, 4, https://doi.org/10.1038/sdata.2017.76, 2017.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quart. J. Roy. Meteor. Soc., 125, 723–757, https://doi.org/10.1002/qj.49712555417, 1999.

Goldberger, A. S.: Econometric Theory, Wiley, New York, NY, 1964.

Goosse, H.: An additional step toward comprehensive paleoclimate reanalyses, Journal of Advances in Modeling Earth Systems, 8, 1501–1503, https://doi.org/10.1002/2016MS000739, 2016.

Goosse, H., Renssen, H., Timmermann, A., Bradley, R. S., and Mann, M. E.: Using paleoclimate proxy-data to select optimal realisations in an ensemble of simulations of the climate of the past millennium, Climate Dynamics, 27, 165–184, https://doi.org/10.1007/s00382-006-0128-6, 2006.

Goosse, H., E., C., de Montety, A., Mann, M. E., Renssen, H., and Timmermann, A.: Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation, J. Geophys. Res.-Atmos., 115, https://doi.org/10.1029/2009jd012737, 2010.

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., Steiger, N. J., and Perkins, W. A.: The Last Millennium Climate Reanalysis Project: Framework and First Results, Journal of Geophysical Research: Atmospheres, 121, 6745–6764, https://doi.org/10.1002/2016JD024751, 2016.

Hamill, T. M., Whitaker, J. S., and Snyder, C.: Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter, Monthly Weather Review, 129, 2776–2790, https://doi.org/10.1175/1520-0493(2001)129<2776:DDFOBE>2.0.CO;2, 2001.

Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, Rev. Geophys., 48, https://doi.org/10.1029/2010RG000345, 2010.

Juckes, M. N., Allen, M. R., Briffa, K. R., Esper, J., Hegerl, G. C., Moberg, A., Osborn, T. J., and Weber, S. L.: Millennial temperature reconstruction intercomparison and evaluation, Clim. Past, 3, 591–609, https://doi.org/10.5194/cp-3-591-2007, 2007.

Landrum, L., Otto-Bliesner, B. L., Wahl, E. R., Conley, A., Lawrence, P. J., Rosenbloom, N., and Teng, H.: Last millennium climate and its variability in CCSM4, Journal of Climate, 26, 1085–1111, https://doi.org/10.1175/JCLI-D-11-00326.1, 2013.

Liu, H., Liu, Z., and Lu, F.: A systematic comparison of particle filter and EnKF in assimilating time–averaged observations, Journal of Geophysical Research: Atmospheres, 122, https://doi.org/10.1002/2017JD026798, 2017.

Mann, M. E. and Jones, P. D.: Global surface temperatures over the past two millennia, Geophysical Research Letters, 30, https://doi.org/10.1029/2003GL017814, 2003.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Northern hemisphere temperatures during the past millennium: Inferences, uncertainties, and limitations, Geophysical Research Letters, 26, 759–762, https://doi.org/10.1029/1999GL900070, 1999.

Mann, M. E., Zhang, Z., Hughes, M. K., Bradley, R. S., Miller, S. K., Rutherford, S., and Ni, F.: Proxy-based reconstructions of hemispheric and global surface temperature variations over the past two millennia, Proceedings of the National Academy of Sciences, 105, 13 252–13 257, https://doi.org/10.1073/pnas.0805721105, 2008.

Mann, M. E., Zhang, Z., Rutherford, S., Bradley, R. S., Hughes, M. K., Shindell, D., Ammann, C., Faluvegi, G., and F., N.: Global signatures and dynamical origins of the Little Ice Age and Medieval Climate Anomaly, Science, 326, 1256–1260, https://doi.org/10.1126/science.1177303, 2009.

Matsikaris, A., Widmann, M., and Jungclaus, J. H.: On-line and off-line data assimilation in palaeoclimatology: a case study, Clim. Past, 11, 81–93, https://doi.org/10.5194/cp-11-81-2015, 2015.

Moberg, A., Sonechkin, D. M., Holmgren, K., Datsenko, N. M., and Karlén, W.: Highly variable Northern Hemisphere temperatures reconstructed from low- and high-resolution proxy data, Nature, 433, 613–617, 2005.

Morice, C. P., Kennedy, J. J., Rayner, N. A., and Jones, P. D.: Quantifying uncertainties in global and regional temperature change using an ensemble of observational estimates: the HadCRUT4 dataset, Journal of Geophysical Research: Atmospheres, 117, https://doi.org/10.1029/2011JD017187, 2012.

Murphy, J. M.: The impact of ensemble forecasts on predictability, Quart. J. Roy. Meteor. Soc., 114, 463–493, https://doi.org/10.1002/qj.49711448010, 1988.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, Journal of Hydrology, 10, 282–290, 1970.

Okazaki, A. and Yoshimura, K.: Development and evaluation of a system of proxy data assimilation for paleoclimate reconstruction, Clim. Past, 13, 379–393, https://doi.org/10.5194/cp-13-379-2017, 2017.

Oke, P., Schiller, A., Griffin, D. A., and Brassington, G. B.: Ensemble data assimilation for an eddy-resolving ocean model of the Australian region, Quarterly Journal of the Royal Meteorological Society, 131, 3301–3311, https://doi.org/10.1256/qj.05.95, 2005.

Oke, P. R., Allen, J. S., Miller, R. N., Egbert, G. D., and Kosro, P. M.: Assimilation of surface velocity data into a primitive equation coastal ocean model, Journal of Geophysical Research: Oceans, 107, https://doi.org/10.1029/2000JC000511, 2002.

Oke, P. R., Sakov, P., and Corney, S. P.: Impacts of localisation in the EnKF and EnOI: experiments with a small model, Ocean Dynamics, 57, 32–45, https://doi.org/10.1007/s10236-006-0088-8, 2007.

PAGES 2k Consortium: Continental-scale temperature variability during the past two millennia, Nature Geoscience, 6, 339–346, https://doi.org/10.1038/ngeo1834, 2013.

PAGES 2k Consortium: A global multiproxy database for temperature reconstructions of the Common Era, Sci. Data, 4, https://doi.org/10.1038/sdata.2017.88, 2017.

Perkins, W. A. and Hakim, G. J.: Reconstructing paleoclimate fields using online data assimilation with a linear inverse model, Clim. Past, 13, 421–436, https://doi.org/10.5194/cp-13-421-2017, 2017.

Poli, P., Hersbach, H., and Dee, D. P.: ERA-20C: An atmospheric reanalysis of the twentieth century, J. Climate, 29, 4083–4097, https://doi.org/10.1175/JCLI-D-15-0556.1, 2016.

Pollack, H. N. and Smerdon, J. E.: Borehole climate reconstructions: Spatial structure and hemispheric averages, Journal of Geophysical Research, 109, https://doi.org/10.1029/2003JD004163, 2004.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickman, C., and Mosher, S.: Berkeley Earth temperature averaging process, Geoinformatics and Geostatistics: An Overview, 1:2, 1–13, https://doi.org/10.4172/2327-4581.1000103, 2013.

Rutherford, S., M. E. Mann, M. E., Osborn, T. J., Bradley, R. S., Briffa, K. R., Hughes, M. K., and Jones, P. D.: Proxy-based Northern Hemisphere surface temperature reconstructions: Sensitivity to method, predictor network, target season, and target domain, Journal of Climate, 18, 2308–2329, https://doi.org/10.1175/JCLI3351.1, 2005.

Sano, M., Furuta, F., and Sweda, T.: Tree-ring-width chronology of *Larix gmelinii* as an indicator of changes in early summer temperature in east-central Kamchatka, J. Forest Research, 14, 147–154, https://doi.org/10.1007/s10310-009-0123-y, 2009.

Schneider, U., Becker, A., Finger, P., Meyer-Christoffer, A., Ziese, M., and Rudolf, B.: GPCC's new land surface precipitation climatology based on quality-controlled in situ data and its role in quantifying the global water cycle, Theoretical and Applied Climatology, 115, 15–40, https://doi.org/10.1007/s00704-013-0860-x, 2014.

Schwarz, G. E.: Estimating the dimension of a model, Annals of Statistics, 6, 461–464, https://doi.org/10.1214/aos/1176344136, 1978.

Smith, T. M., Reynolds, R. W., Peterson, T. C., and Lawrimore, J.: Improvements to NOAA's historical merged land–ocean surface temperature analysis (1880-2006), Journal of Climate, 21, 2283–2296, https://doi.org/10.1175/2007JCLI2100.1, 2008.

St. George, S.: An overview of tree-ring width records across the Northern Hemisphere, Quaternary Science Reviews, 95, 132–150, https://doi.org/10.1016/j.quascirev.2014.04.029, 2014.

St. George, S., Meko, D. M., and Cook, E. R.: The seasonality of precipitation signals embedded within the North American Drought Atlas, The Holocene, 20, 983–988, https://doi.org/10.1177/0959683610365937, 2010.

Stahle, D. W., Cleaveland, M. K., Grissino-Mayer, H. D., Griffin, R. D., Fye, F. K., Therell, M. D., Burnette, D. J., Meko, D. M., and Villanueva Diaz, J.: Cool- and warm-season precipitation reconstructions over western New Mexico, J. Climate, 22, 3729–3750, https://doi.org/10.1175/2008JCLI2752.1, 2009.

Steiger, N. J. and Smerdon, J. E.: A pseudoproxy assessment of data assimilation for reconstructing the atmosphere–ocean dynamics of hydroclimate extremes, Clim. Past, 13, 1435–1449, https://doi.org/10.5194/cp-13-1435-2017, 2017.

Steiger, N. J., Hakim, G. J., Steig, E. J., Battisti, D. S., and Roe, G. H.: Assimilation of time-averaged pseudoproxies for climate reconstruction, Journal of Climate, 27, 426–441, https://doi.org/10.1175/JCLI-D-12-00693.1, 2014.

Steiger, N. J., Steig, E. J., Dee, S. G., Roe, G. H., and Hakim, G. J.: Climate reconstruction using data assimilation of water isotope ratios from ice cores, Journal of Geophysical Research: Atmospheres, 122, 1545–1568, https://doi.org/10.1002/2016JD026011, 2017.

Steiger, N. J., Smerdon, J. E., Cook, E. R., and Cook, B. I.: A reconstruction of global hydroclimate and dynamical variables over the Common Era, Sci. Data, 5, https://doi.org/10.1086/sdata.2018.86, 2018.

Tardif, R., Hakim, G. J., and Snyder, C.: Coupled atmosphere–ocean data assimilation experiments with a low-order model and CMIP5 model data, Climate Dynamics, 45, 1415–1427, 2016.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American Meteorological Society, 93, 485–498, https://doi.org/10.1175/BAMS-D-11-00094.1, 2012.

Tolwinski-Ward, S. E., Evans, M. N., Hughes, M. K., and Anchukaitis, K. J.: An efficient forward model of the climate controls on interannnual variation in tree-ring width, Climate Dynamics, 36, 2419–2439, https://doi.org/10.1007/s00382-011-1062-9, 2011.

Touchan, R., Garfin, G., Meko, D., Funkhouser, G., Erkan, N., Hughes, M., and Wallin, B.: Preliminary reconstructions of spring precipitation in southwestern Turkey from tree-ring width, Int. J. Climatol., 23, 157–171, https://doi.org/10.1002/joc.850, 2003.

Wang, J., Emile-Geay, J., McKay, N. P., and Rajaratnam, B.: Fragility of reconstructed temperature patterns over the common era: Implications for model evaluation, Geophysical Research Letters, 42, 7162–7170, https://doi.org/10.1002/2015GL065265, 2015.

Whitaker, J. S. and Hamill, T. M.: Ensemble data assimilation without perturbed observations, Monthly Weather Review, 130, 1913–1924, https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2, 2002.

Widmann, M., Goosse, H., Schrier, G., Schnur, R., and Barkmeijer, J.: Using data assimilation to study extratropical Northern Hemisphere climate over the last millennium, Clim. Past, 6, 627–644, https://doi.org/10.5194/cp-6-627-2010, 2010.

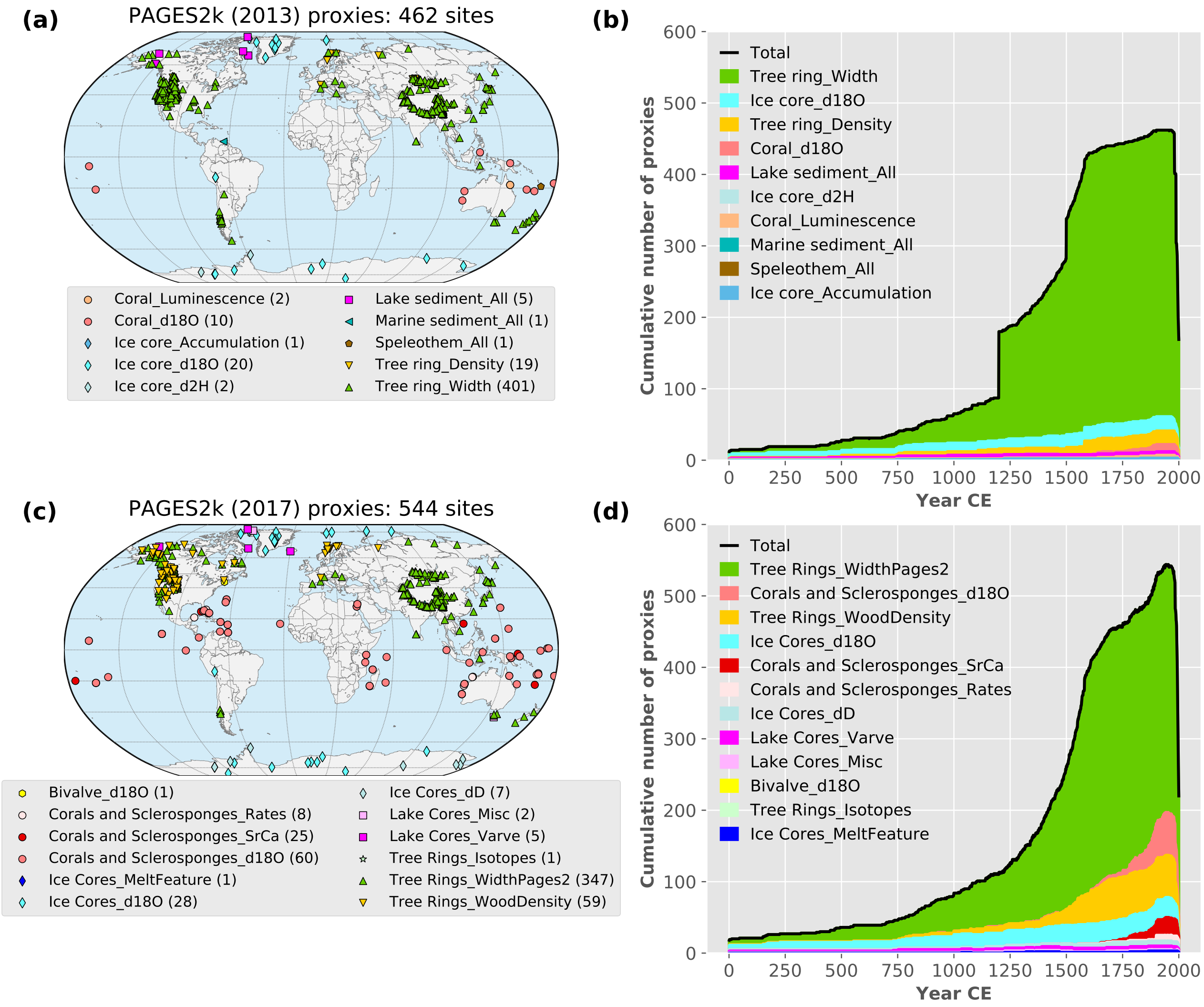


Figure 1. Locations (left column) and temporal (right column) distributions of proxy records available for assimilation (proxies for which linear PSMs calibrated with GISTEMP version 4 are available), (a) and (b) used in the prototype version, (c) and (d) LMR proxy database updated to PAGES 2k Consortium (2017) proxies~~, and finally (e) and (f) with the further addition of proxies from ?~~.

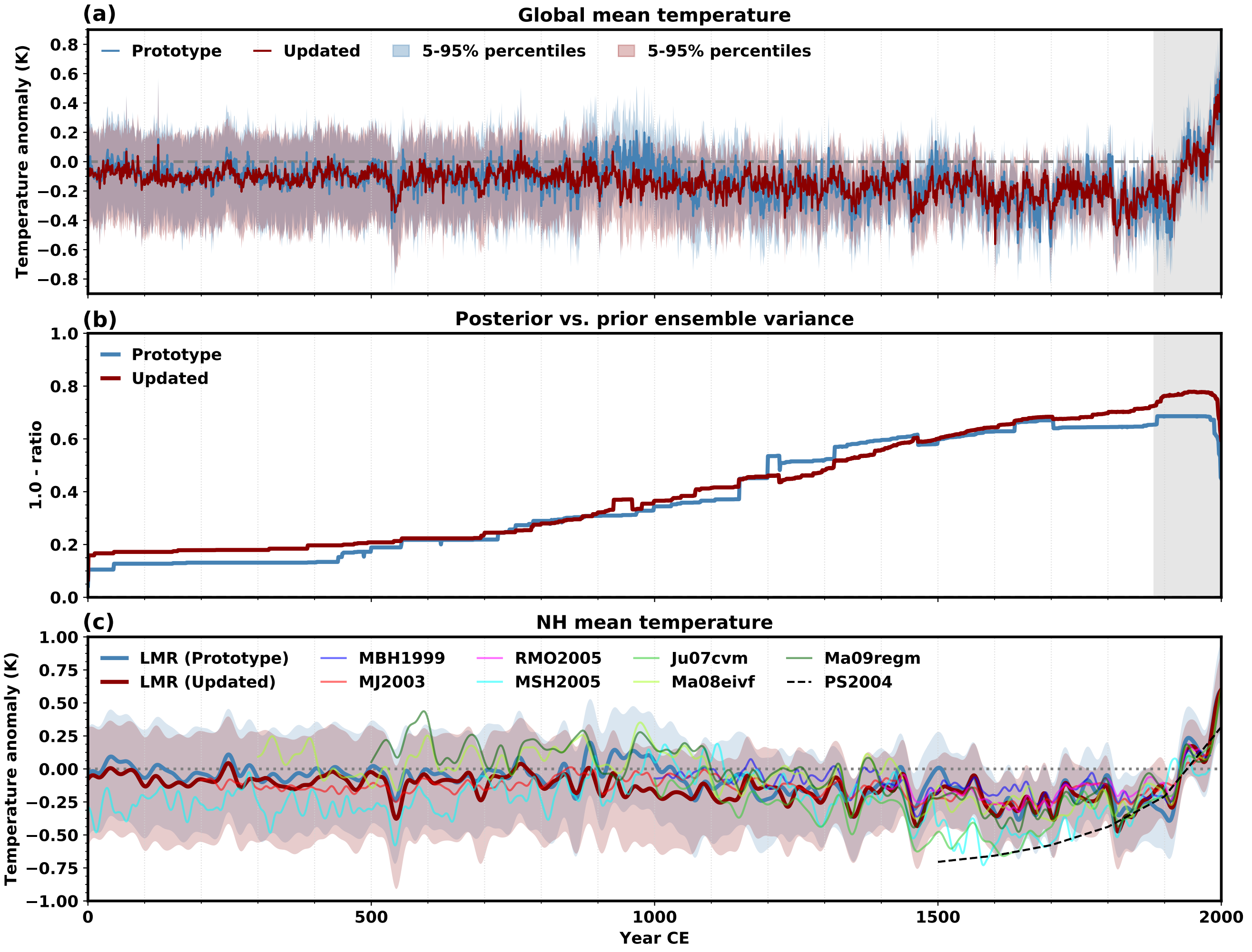


Figure 2. Comparison of the LMR global-mean 2 m air temperature (GMT) (a) grand ensemble mean (solid lines) and 5–95% percentile range (shading) from the prototype (blue) and updated (red) reanalyses over the Common Era, and (b) one minus the mean (across Monte-Carlo realizations) ratio of the posterior and prior GMT ensemble variance. (c) Comparison of the LMR Northern Hemisphere 2 m air temperature grand ensemble mean (solid lines) and 5–95% percentile range (shading) from the prototype and updated reanalyses with reconstructions from other authors: MBH1999: Mann et al. (1999), ~~over~~ MJ2003: Mann and Jones (2003) , RMO2005: Rutherford et al. (2005), MSH2005: Moberg et al. (2005), Ju07cvm: Juckes et al. (2007), Ma08eivf: Mann et al. (2008), Ma09regm: Mann et al. (2009), PS2004: Pollack and Smerdon (2004). All series in (c) represent anomalies (K) from the ~~Common Era~~1900–1980 mean, and have been smoothed with a 30-year low-pass Butterworth filter. The light gray shading in (a) and (b) indicate the verification period discussed in Fig. 3.

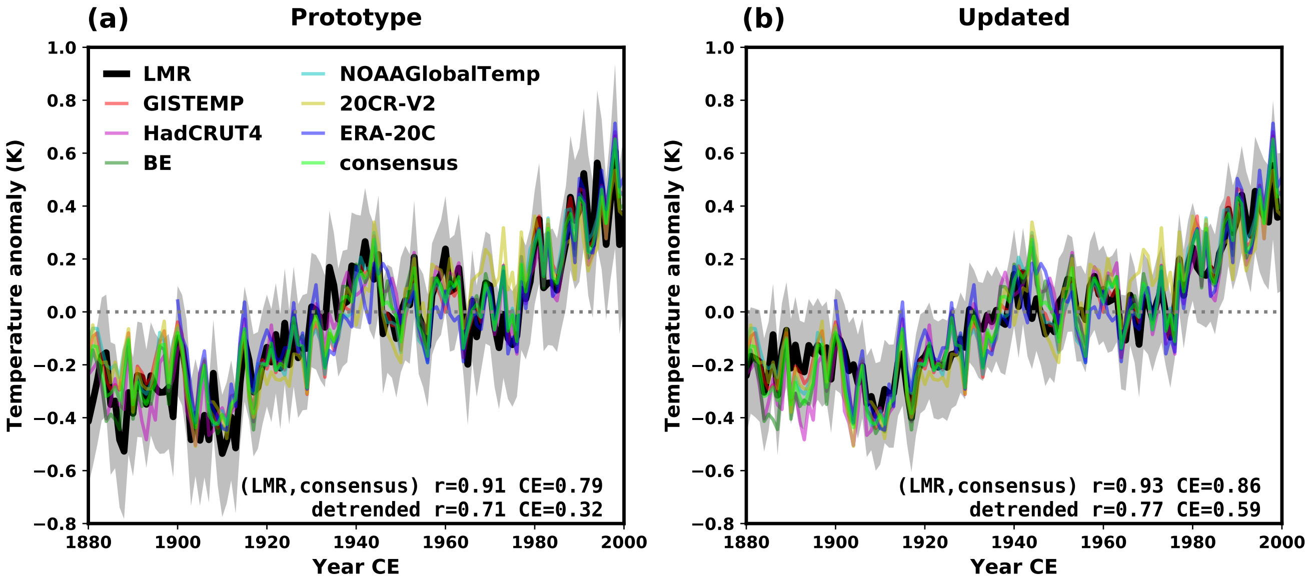


Figure 3. Comparison of LMR global-mean 2 m air temperature (~~e~~GMT) (a) prototype and (~~d~~b) updated reanalyses, against instrumental–era analyses (GISTEMP: NASA GISS surface temperature (Hansen et al., 2010); HadCRUT4: Hadley Center/Climate Research Unit at the University of East Anglia temperature data set version 4 (Morice et al., 2012); BE: Berkeley Earth surface temperature (Rohde et al., 2013); NOAAGlobalTemp:NOAA merged land-ocean surface temperature version 3.5.4 (Smith et al., 2008); 20CR-V2: NOAA twentieth century reanalysis version 2 (Compo et al., 2011); ERA-20C: ECMWF reanalysis of the twentieth century (Poli et al., 2016); Consensus: average of all but LMR). The gray bands show the LMR 5–95% percentile range. Verification correlation (r) and coefficient of efficiency (CE) values are shown ~~below~~ at the ~~figures. The light gray shading in (a) and (b) indicate~~ bottom of each panel, for the ~~verification period shown in (c)~~ original and ~~(d)~~detrended time series.

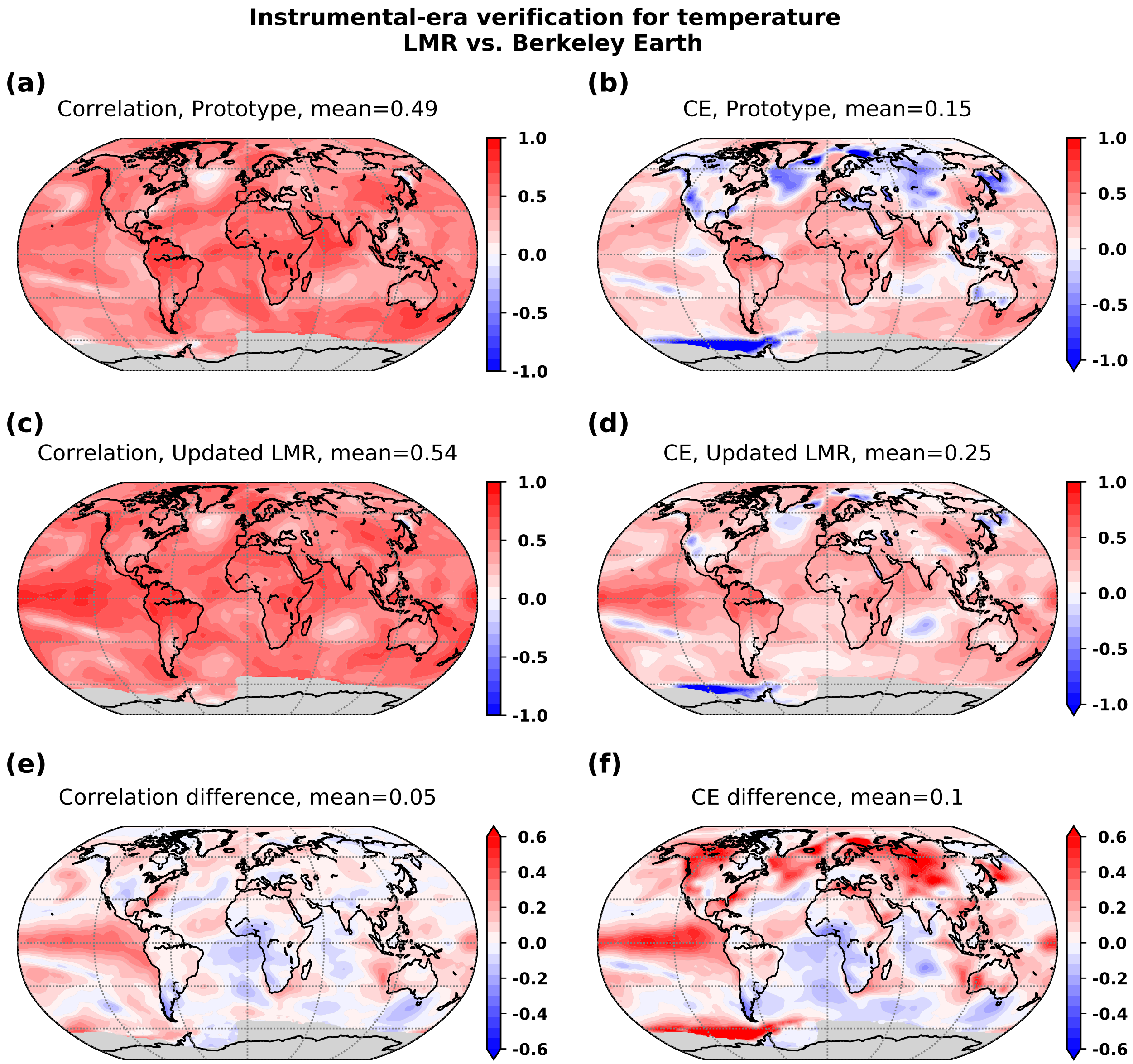
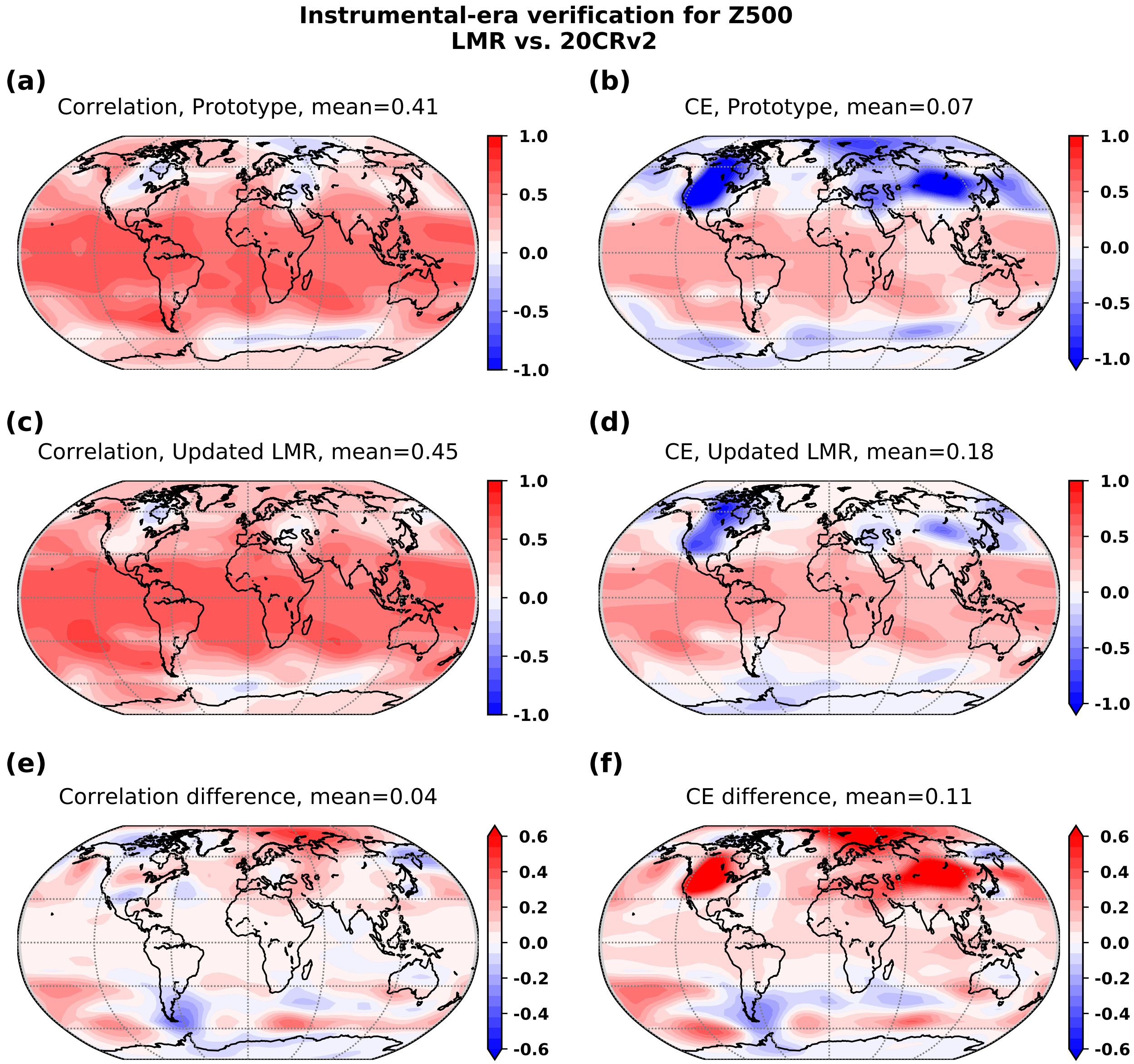


Figure 4. Verification of LMR 2m air temperature against the Berkeley Earth instrumental–era analysis over the 1880–2000 period. Shown are time series correlation (left column) and coefficient of efficiency (CE, right column), for (a) and (b) the prototype and, (c) and (d) the updated reanalysis. Differences in correlations and CE between the two experiments are shown in (e) and (f) respectively. Gray shading indicate regions with insufficient valid data for meaningful verification statistics.



Instrumental-era verification for Z500
LMR vs. 20CRv2
(a)
Correlation, Prototype, mean=0.41
(b)
CE, Prototype, mean=0.07
(c)
Correlation, Updated LMR, mean=0.45
(d)
CE, Updated LMR, mean=0.18
(e)
Correlation difference, mean=0.04
(f)
CE difference, mean=0.11
1.0
0.5
0.0
-0.5
-1.0
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6

Figure 5. As in Fig. 4 except for the verification of LMR 500 hPa geopotential height anomalies against the 20CR-v2 reanalysis.

~~As in Fig. 4, except for verification of LMR Palmer Drought Severity Index (PDSI) against the Dai (2011) instrumental–era PDSI analysis.~~

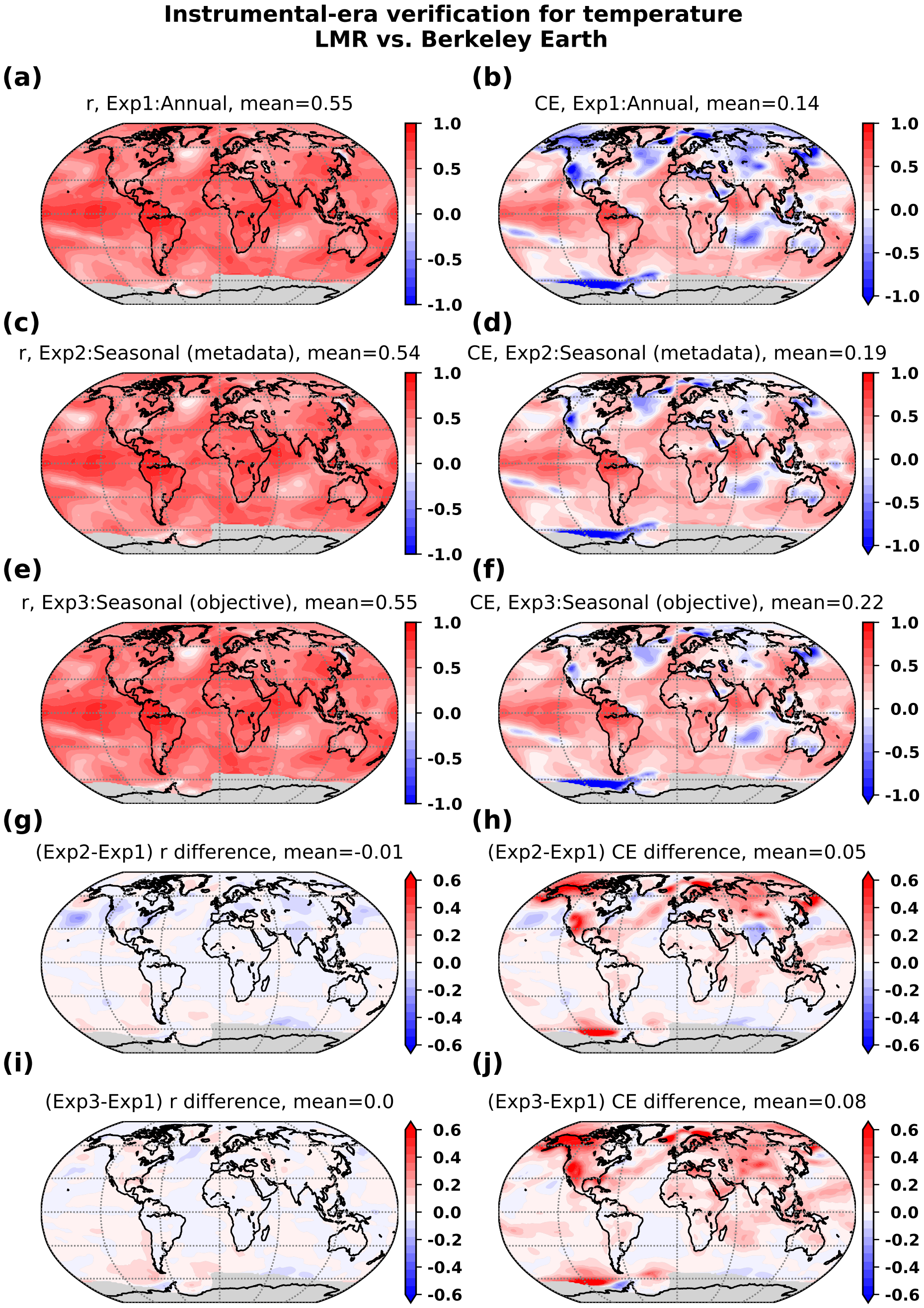

Instrumental-era verification for temperature
LMR vs. Berkeley Earth
(a)
r, Exp1:Annual, mean=0.55
(b)
CE, Exp1:Annual, mean=0.14
(c)
r, Exp2:Seasonal (metadata), mean=0.54
(d)
CE, Exp2:Seasonal (metadata), mean=0.19
(e)
r, Exp3:Seasonal (objective), mean=0.55
(f)
CE, Exp3:Seasonal (objective), mean=0.22
(g)
(Exp2-Exp1) r difference, mean=-0.01
(h)
(Exp2-Exp1) CE difference, mean=0.05
(i)
(Exp3-Exp1) r difference, mean=0.0
(j)
(Exp3-Exp1) CE difference, mean=0.08
1.0
0.5
0.0
-0.5
-1.0
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6

Figure 6. Verification of LMR temperature anomalies against the Berkeley Earth instrumental–era analysis, for experiments using PAGES2k-2017 proxies and univariate PSMs, with contrasting seasonalities. Shown are time series correlation (r) and coefficient of efficiency (CE), for (a) and (b) experiment 1: annual, (c) and (d) experiment 2: seasonality from the proxy metadata, and (e) and (f) experiment 3: objectively-derived seasonality. Differences in skill metrics are also shown, (g) and (h) between experiments 2 and 1, (i) and (j) between experiments 3 and 1.

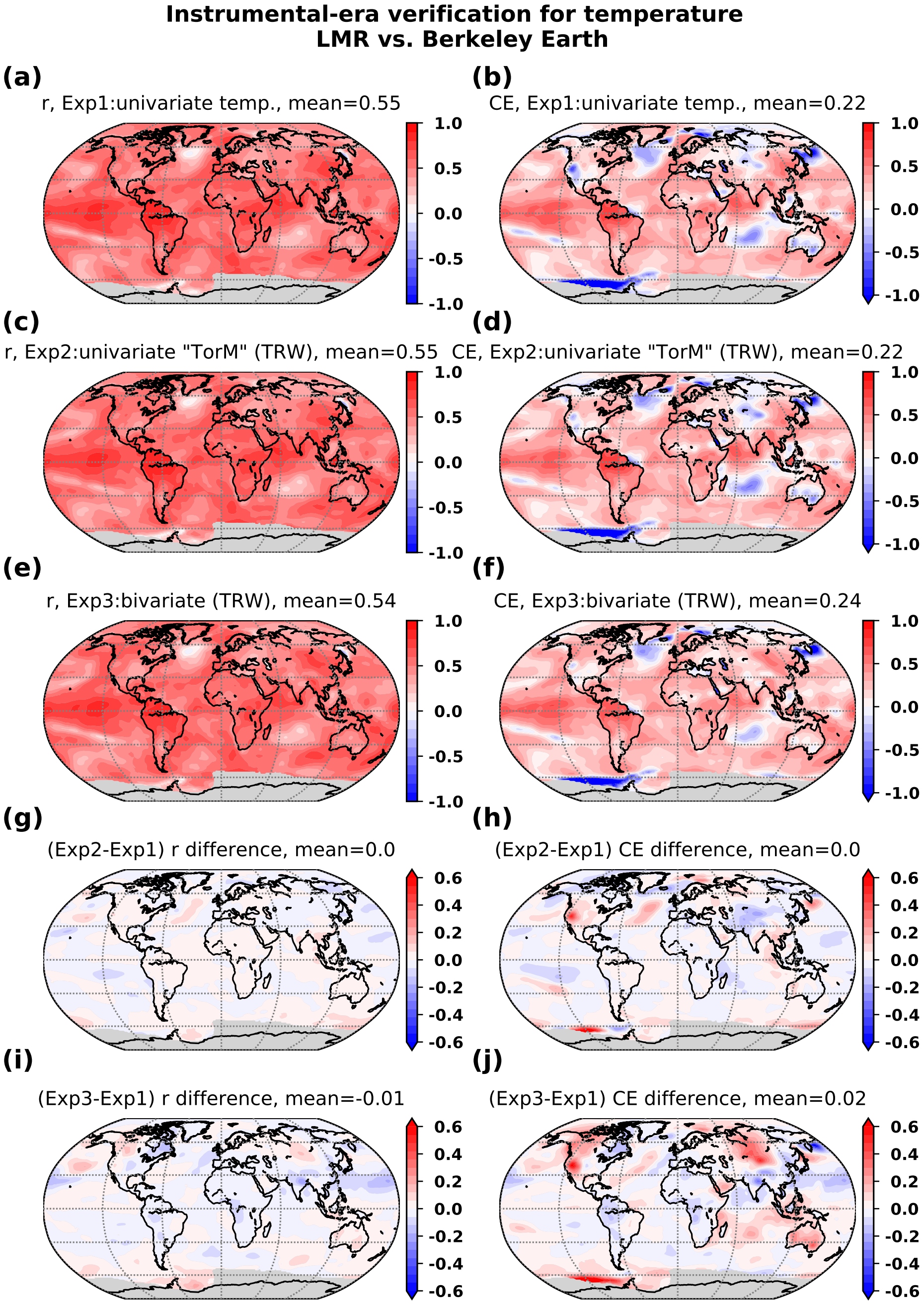

Instrumental-era verification for temperature
LMR vs. Berkeley Earth
(a)
r, Exp1:univariate temp., mean=0.55
(b)
CE, Exp1:univariate temp., mean=0.22
(c)
r, Exp2:univariate "TorM" (TRW), mean=0.55
(d)
CE, Exp2:univariate "TorM" (TRW), mean=0.22
(e)
r, Exp3:bivariate (TRW), mean=0.54
(f)
CE, Exp3:bivariate (TRW), mean=0.24
(g)
(Exp2-Exp1) r difference, mean=0.0
(h)
(Exp2-Exp1) CE difference, mean=0.0
(i)
(Exp3-Exp1) r difference, mean=-0.01
(j)
(Exp3-Exp1) CE difference, mean=0.02
1.0
0.5
0.0
-0.5
-1.0
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6

Figure 7. As in Figure 6, but comparing experiments performed using PAGES2k-2017 proxies with different PSM configurations for tree ring width proxies. (a) and (b) experiment 1: univariate on temperature for all proxies, (c) and (d) experiment 2: univariate with respect to temperature or moisture for TRWs, and (e) and (f) experiment 3: bivariate on temperature and moisture for tree ring widths. Differences in skill metrics are shown, (g) and (h) between experiments 2 and 1, and (i) and (j) between experiments 3 and 1. All reconstructions are based on objectively-derived seasonal PSMs.

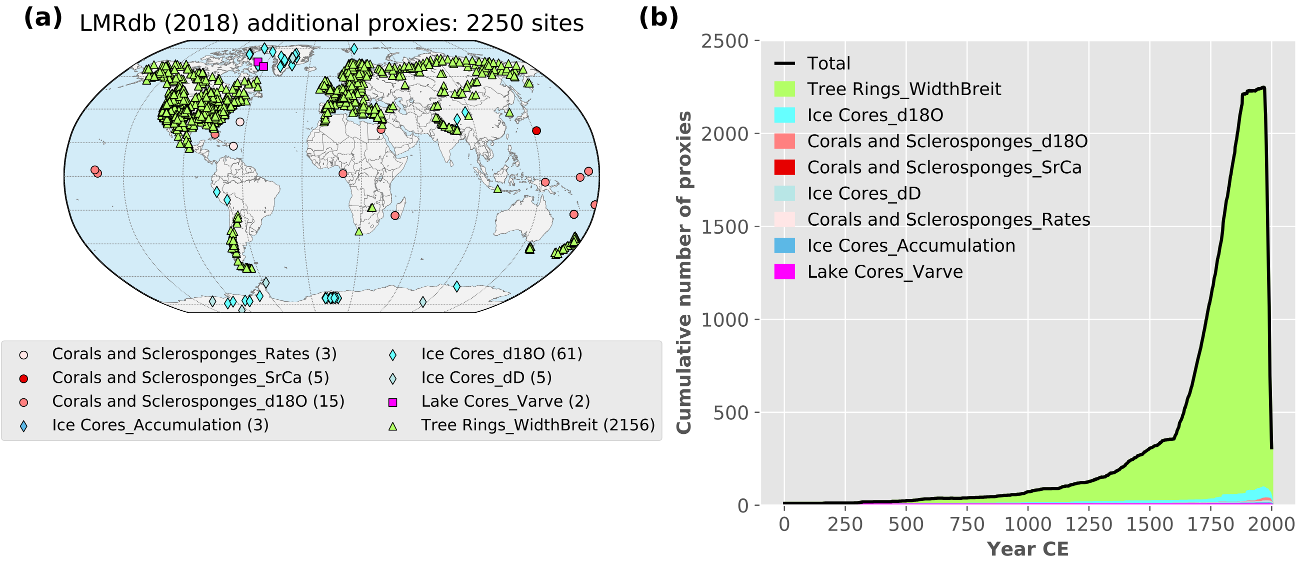


Figure 8. ~~Summary skill metrics from verification of LMR temperature reconstruction experiments, all using PAGES2k-2017 proxies only, against the consensus of instrumental–era products. Experiments are a combination of those shown in Figs. 6 and 7. Metrics shown are the 20th century trend of global mean temperature~~ Locations (~~GMT~~a) ~~, correlation (r)~~ and ~~coefficient of efficiency~~ temporal distributions (~~CE~~b) ~~for the detrended GMT. The GMT trend from consensus~~ of ~~instrumental–era products is shown by~~ the ~~arrow and dashed black line~~additional proxies from Anderson et al. (2019) considered for assimilation, ~~along with~~ including the ~~range defined by the individual instrumental–era products shown by the gray-shaded area~~tree ring chronologies from Breitenmoser et al. (2014). ~~Error bars~~ As in Fig. 1 only records available for assimilation (proxies for which regression based PSMs can be calibrated) are ~~the 5-95% bootstrap confidence intervals on the corresponding skill metric~~shown.

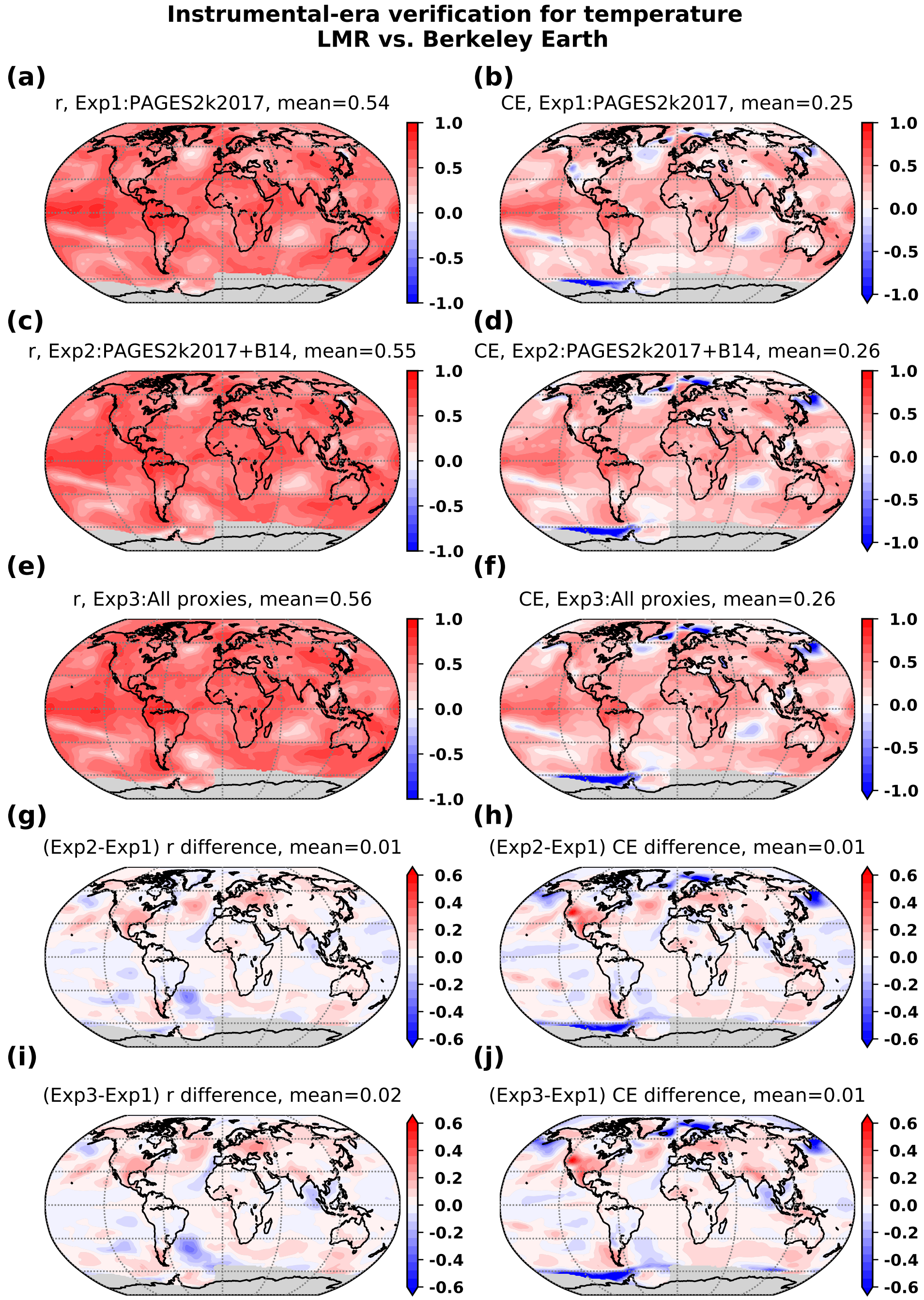

Instrumental-era verification for temperature
LMR vs. Berkeley Earth
(a)
r, Exp1:PAGES2k2017, mean=0.54
(b)
CE, Exp1:PAGES2k2017, mean=0.25
(c)
r, Exp2:PAGES2k2017+B14, mean=0.55
(d)
CE, Exp2:PAGES2k2017+B14, mean=0.26
(e)
r, Exp3:All proxies, mean=0.56
(f)
CE, Exp3:All proxies, mean=0.26
(g)
(Exp2-Exp1) r difference, mean=0.01
(h)
(Exp2-Exp1) CE difference, mean=0.01
(i)
(Exp3-Exp1) r difference, mean=0.02
(j)
(Exp3-Exp1) CE difference, mean=0.01
1.0
0.5
0.0
-0.5
-1.0
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6

Figure 9. As in Figure 6, but comparing experiments performed with different proxy networks: (a) correlation (r) and (b) CE for experiment 1: PAGES 2k Consortium (2017) proxies only, (c) and (d) experiment 2: with the addition of tree ring chronologies from Breitenmoser et al. (2014), and (e) and (f) experiment 3: with all proxies in the updated LMR database. The differences in correlation and CE between experiments 2 and 1 are shown in (g) and (h) respectively, and between experiments 3 and 2 in (i) and (j). Notice the latter is different to Figure 6, where differences between experiments 3 and 1 are shown.

~~As in Fig. **??** for reconstruction experiments performed with different proxy networks: PAGES 2k Consortium (2017) proxies only, with the addition of tree ring chronologies from Breitenmoser et al. (2014), and with all proxies in the updated LMR database.~~

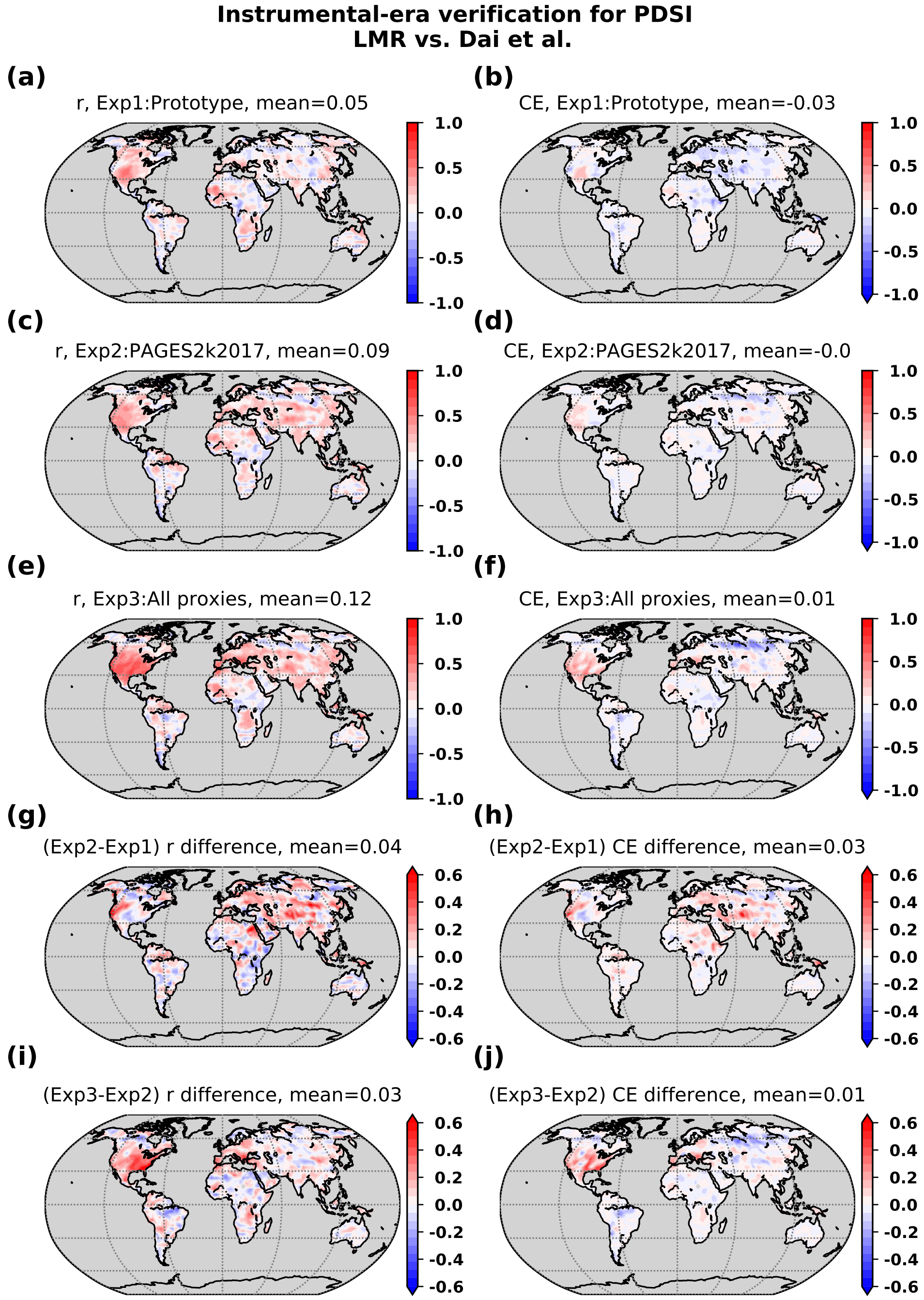

Instrumental-era verification for PDSI
LMR vs. Dai et al.
(a)
r, Exp1:Prototype, mean=0.05
(b)
CE, Exp1:Prototype, mean=-0.03
(c)
r, Exp2:PAGES2k2017, mean=0.09
(d)
CE, Exp2:PAGES2k2017, mean=-0.0
(e)
r, Exp3:All proxies, mean=0.12
(f)
CE, Exp3:All proxies, mean=0.01
(g)
(Exp2-Exp1) r difference, mean=0.04
(h)
(Exp2-Exp1) CE difference, mean=0.03
(i)
(Exp3-Exp2) r difference, mean=0.03
(j)
(Exp3-Exp2) CE difference, mean=0.01
1.0
0.5
0.0
-0.5
-1.0
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6

Figure 10. ~~As in~~ Similar to Figure 9, but comparing PDSI reconstructions against the Dai (2011) analysis for experiments performed with different proxy networks: (a) correlation and (b) CE for experiment 1: prototype reanalysis ~~already presented in Fig. ??~~from H16, experiment 2: PAGES 2k Consortium (2017) proxies~~only~~, (e) and (f) experiment 3: with further the addition of tree ring chronologies from Breitenmoser et al. (2014) and the coral, ice and lake core records from Anderson et al. (2019) (i.e. the full proxy database). The differences in correlation and CE between experiments 2 and 1 are shown in (~~e~~g) and (~~f~~h) respectively, and between experiments 3 and 2 in (i) and (j).

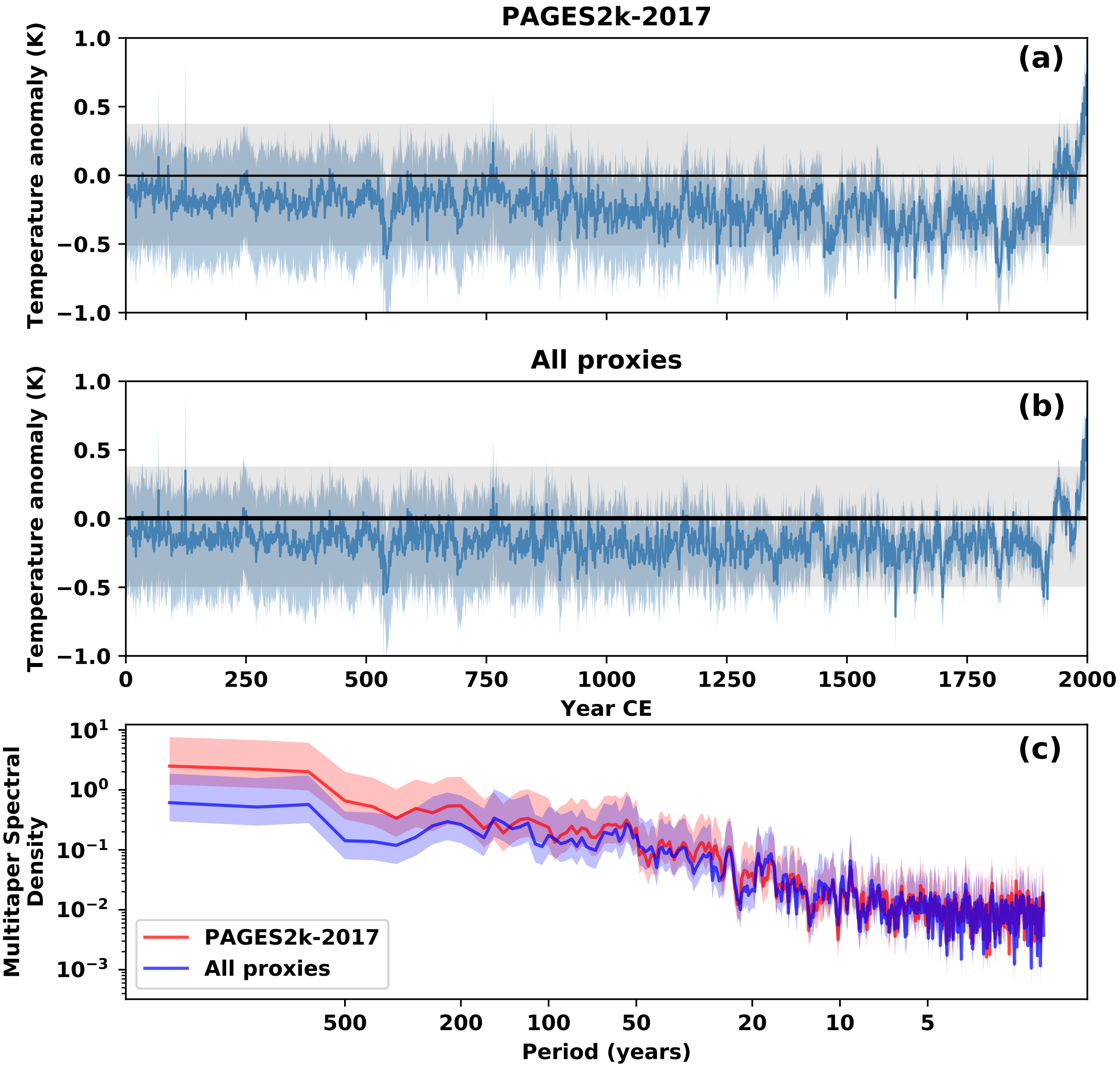


Figure 11. Northern Hemisphere temperature (NHMT) grand ensemble mean (blue solid lines) and 5–95% percentile range (blue shading) from (a) experiment performed with PAGES2k-2017 proxies and (b) experiment with the addition of proxies from Anderson et al. (2019) . The prior ensemble is shown by the solid black lines (ensemble mean) and 5–95% percentile range by the gray shading. (c) Spectra of NHMT from both experiments (solid lines), along with the χ^2 95% highest density regions (shading).

Table 1. Summary of instrumental–era verification results for the prototype and updated reanalyses. Verification scores shown are correlation (r) and coefficient of efficiency (CE), for the annual global mean temperature (GMT) and detrended GMT verified against the consensus of instrumental–era analyses, the global mean of gridpoint r and CE characterizing the spatially reconstructed temperature, 500 hPa geopotential height (Z500) and Palmer Drought Severity Index (PDSI). LMR spatial temperature is verified against the Berkeley Earth analysis (Rohde et al., 2013), Z500 is verified against the 20CR-V2 reanalysis (Compo et al., 2011) and PDSI is verified against the Dai (2011) analysis.

Reanalysis	Annual GMT		Detrended GMT		Spatial temperature		Spatial Z500		Spatial PDSI	
	r	CE	r	CE	r	CE	r	CE	r	CE
Prototype	0.91	0.79	0.71	0.32	~~0.49~~ 0.47	~~0.15~~ 0.10	0.41	0.07	0.05	-0.03
Updated	0.93	~~0.87~~ 0.86	0.77	0.59	0.52	0.22	0.45	0.18	0.09	0.00

Table 2. Verification of LMR prototype and updated reanalyses against independent (withheld from assimilation) proxies. Skill scores shown are the median of distributions for correlation (r), the fraction of proxy records characterized by a positive ΔCE (%+CE), and the median of the ΔCE distribution, where ΔCE is the difference in the coefficient of efficiency (CE) between the posterior (reanalysis) and the prior. The median of the ensemble calibration ratio (ECR) distribution is also shown. Statistics are compiled over 51 Monte-Carlo realizations, and cover different time periods, including the 1880–2000 PSM calibration period.

Verification period (years of Common Era)	Prototype				Updated reanalysis			
	r	%+CE	ΔCE	ECR	r	%+CE	ΔCE	ECR
1–499	0.00	56.0	0.00	0.78	0.03	55.9	0.00	0.96
500–999	0.08	62.1	0.01	1.00	0.13	65.3	0.02	1.00
1000–1499	0.11	63.0	0.01	1.10	0.16	67.3	0.05	1.06
1500–1879	0.14	64.1	0.02	1.06	0.28	72.7	0.10	1.02
1880–2000	0.23	72.6	0.03	0.97	0.40	82.7	0.13	0.89

Table 3. Summary of instrumental-era verification results for reconstruction experiments performed with various PSM configurations. Verification scores shown are the trend over the twentieth century (in K/100yrs), correlation (r) and coefficient of efficiency (CE), for the annual global mean temperature (GMT) and detrended GMT verified against the consensus of instrumental–era analyses. The GMT trend in the consensus of instrumental-era analyses is 0.56 K/100 yrs.

PSM configuration	GMT trend	Annual GMT		Detrended GMT	
		r	CE	r	CE
Prototype	0.61	0.91	0.79	0.71	0.32
Univariate - temperature (annual)	0.85	0.93	0.61	0.74	0.39
Univariate - temperature (seasonal meta.)	0.72	0.93	0.77	0.73	0.43
Univariate - temperature (seasonal obj.)	0.72	0.93	0.80	0.75	0.51
Univariate - temperature or moisture (TRW) (seasonal meta.)	0.71	0.92	0.78	0.72	0.44
Univariate - temperature or moisture (TRW) (seasonal obj.)	0.74	0.93	0.77	0.74	0.48
Bivariate - temperature and moisture (TRW) (seasonal meta.)	0.62	0.93	0.84	0.76	0.50
Bivariate - temperature and moisture (TRW) (seasonal obj.)	0.60	~~0.56~~ 0.93	~~0.26~~ 0.86	~~0.46~~ 0.77	0.54

Table 4. Verification of LMR reconstructions against independent (withheld from assimilation) proxies, for experiments using various PSM configurations. Skill scores shown are the median of distributions for correlation (r), the fraction of proxy records characterized by a positive ΔCE (%+CE), and the median of the ΔCE distribution. Statistics are compiled over 51 Monte-Carlo realizations, for two distinct periods: 1880–2000 (PSM calibration period) and 0–1879 (pre-calibration period).

PSM configuration	1880–2000			1–1879		
	r	%+CE	ΔCE	r	%+CE	ΔCE
Univariate - temperature (annual)	0.28	75.2	0.05	0.17	~~0.12~~ 66.0	~~0.01~~ 0.03
Univariate - temperature (seasonal meta.)	0.32	78.7	0.06	0.21	69.6	0.04
Univariate - temperature (seasonal obj.)	0.34	80.6	0.09	0.21	69.4	0.06
Univariate - temperature or moisture (TRW) (seasonal meta.)	0.30	76.1	0.06	0.19	67.7	0.04
Univariate - temperature or moisture (TRW) (seasonal obj.)	0.33	77.6	0.08	0.19	66.3	0.04
Bivariate - temperature and moisture (TRW) (seasonal meta.)	0.32	77.9	0.07	0.20	68.1	0.04
Bivariate - temperature and moisture (TRW) (seasonal obj.)	0.36	78.9	0.11	0.22	66.0	0.06

Table 5. ~~Ensemble calibration ratios characterizing ensembles~~ Twentieth century trend of global-mean temperature ~~from~~ (GMT), correlation (r) and coefficient of efficiency (CE), for the annual and detrended GMT, as well as the global mean of the spatial (i.e. gridpoint) r and CE of reconstructed temperature verified against the consensus of instrumental-era analyses, for reconstruction experiments performed with covariance localization ~~applied with~~ using various ~~values of the~~ localization ~~radius R_L~~ cut-off radii L_R. ~~The ratio from~~ Verification statistics for an experiment without covariance localization ~~is~~ are also shown for comparison. ~~The result~~ Results from the prototype ~~reanalysis is~~ are shown for reference. The GMT trend in the consensus of instrumental-era analyses is 0.56 K/100 yrs.

~~Localization~~	~~Localization~~ L_R	~~Localization~~ L_R	~~Localization~~ L_R	~~Localization~~ L_R	~~No localization~~ L_R	L_R
~~R_L =~~	5000 km	~~R_L =~~ 10000 km	~~R_L =~~ 15000 km	~~R_L = 20000~~ 25000 km	~~R_L= 25000~~ 35000 km	45000 km
~~0.66~~ Trend (K/100 yrs)	0.17	0.31	0.40	0.49	0.51	0.56
Annual GMT r	0.92	0.93	0.93	0.93	0.93	0.93
Annual GMT CE	0.46	0.71	0.82	0.86	0.87	0.87
Detrended GMT r	0.74	0.77	0.77	0.77	0.77	0.76
Detrended GMT CE	0.35	0.53	0.59	0.59	0.58	0.56
Mean spatial r	0.36	0.46	0.50	~~0.68~~ 0.52	~~0.81~~ 0.52	~~1.06~~ 0.53
Mean spatial CE	0.11	0.17	0.19	0.22	0.21	0.21

Table 6. As in Table 4, but statistics compiled for tree-ring wood-density (MXD) proxies only, and for experiments using the PAGES2k-2017 proxies only, PAGES2k-2017 with the addition of all proxies from A19 (All proxies), and PAGES2k-2017 plus only a subset of A19 records obtained after removing all but 188 TRW records from B14 (B14 subset). See text for selection details. Skill scores are the median of the correlation (r) distributions, the fraction of proxy records characterized by a positive ΔCE (%+CE), and the median of ΔCE distributions. Statistics are compiled over the 51 Monte-Carlo realizations, for the following periods: 1880–2000 (PSM calibration period) and 1600–1879 (pre-calibration period with a significant number of MXD records and covering a significant portion of the Little Ice Age).

PSM configuration	1880–2000			1600–1879		
	r	%+CE	ΔCE	r	%+CE	ΔCE
PAGES2k-2017	0.62	93.2	0.37	0.58	91.4	0.39
All proxies	0.43	88.0	0.18	0.46	95.4	0.26
B14 subset	0.56	92.5	0.30	0.53	93.8	0.34

Table A1. Mean differences in Bayesian Information Criterion (ΔBIC) corresponding to PSMs for records within the proxy categories considered in LMR, between models calibrated using proxy seasonal responses from the metadata or derived objectively during calibration, with respect to the reference of annual seasonality. Calibration dataset: GISTEMP v4.

Proxy types	Number of records	Seasonal (metadata)	Seasonal (objective)
Tree ring width (PAGES2k-2017)	347	-1.34	-4.84
Tree ring width (Breitenmoser et al.)	2156	-1.72	-5.24
Tree ring wood density	59	-23.28	n/a
Coral $\delta^{18}O$	75	+0.02	n/a
Coral Sr/Ca	30	-0.01	n/a
Coral Rates	11	+0.03	n/a
Ice core $\delta^{18}O$	89	+0.02	n/a
Ice core δD	12	0.00	n/a
Ice core accumulation	3	0.00	n/a
Ice core melt	1	0.00	n/a
Lake core varve	7	-0.52	n/a
Lake core misc.	2	-2.32	n/a
Bivalve $\delta^{18}O$	1	0.00	n/a
Tree ring $\delta^{18}O$	1	+11.81	n/a

Table A2. Mean differences in Bayesian Information Criterion (ΔBIC) for tree ring width univariate "temperature or moisture" and bivariate PSMs, calibrated using metadata seasonality or derived objectively during calibration, against their respective univariate temperature-only PSMs as reference. Calibration datasets: GISTEMP v4 and GPCC v6.

PSM formulation	Seasonal (metadata)		Seasonal (objective)	
	PAGES 2k trees	Breitenmoser trees	PAGES 2k trees	Breitenmoser trees
Univariate - temperature or moisture	-0.86	-1.41	-2.59	-6.65
Bivariate	+2.63	+1.73	-2.35	-6.88