Reply to comments by Edward Cook

Interactive comment on "The importance of input data quality and quantity in climate field reconstructions – results from a Kalman filter based paleodata assimilation method" by Jörg Franke et al.

Edward Cook (Referee)

drdendro@ldeo.columbia.edu
Received and published: 6 December 2019

To begin, let me declare that I am not an expert on the new data assimilation methods (DA) being used for climate field reconstruction (CFR) now. Therefore, I will not comment on the way in which the "state-of-the-art paleo data assimilation approach" has been applied in this paper. Rather, I will stick more so to what the title of the paper indicates, i.e. "the importance of input data quality and quantity in climate field reconstructions" as a generic problem that spans all methods of CFR. In so doing, I will point out what I regard as a problem with one of the main conclusions of this paper.

This study is based on three collections of tree-ring records: "(1) 54 of the best temperature sensitive tree-ring chronologies chosen by experts; (2) 415 temperature sensitive tree-ring records chosen less strictly by regional working groups and statistical screening; (3) 2287 tree-ring series that are not screened for climate sensitivity." These are the N-TREND, PAGES2K, and B14 data sets, respectively. I will not get into the issue of how the tree-ring series were processed (detrended and standardized) for temperature reconstruction other than to say that it is crucial to the recovery of multi-decadal to centennial timescale variability. This is possible from tree rings as numerous published studies have shown, but it is a difficult problem nonetheless. Regarding this study, the processing methods used are likely to vary considerably between the three data sets used, with the 54 best N-TREND tree-ring chronologies processed best by the experts, but the effects of these differences are not possible to determine in this paper. This is not a criticism. It is just the way it is given the data used.

Thank you for your feedback. We are aware of the spectral differences in proxies due to detrending, standardization, etc. and the first author of this paper even published on this topic (Franke et al., 2013). This inconsistency within the proxy data is actually the reason why we have another approach than most previous reconstructions, including the methodologically similar data assimilation approach used in the framework of the Last Millennium Reanalysis project (Hakim et al., 2016). A majority of reconstructions methods will probably be affected in their multi-decadal to centennial scale variability depending on the chosen input data set as shown by Tardif et al. (2019).

To avoid such issues and to be able to use the more reliable inter-annual to decadal variability that many tree-ring proxies contain even if they have not been specifically reconstructed to retain low frequency variability, we only assimilate anomalies around 71-year running means. As we explain in line 122, low frequency variability in our reconstruction is purely the model response to the external forcings and is consistent with model physics. As a consequence, the reviewer is right that we do not make use of the specific advantage of N-TREND to include probably the most realistic low frequency variability that can be obtained from tree-ring data.

The importance of input data quality and quantity in climate field reconstructions is at a basic level a given, so much of what this paper demonstrates is not terribly surprising. Thus, as a first-order conclusion, data quality and quantity do matter and more of both is better than less. However, as the authors show, quantity does not necessarily help if the quality of climate signal in the tree rings is not also considered given the target variable being reconstructed, in this case temperature. Thus, data screening for the signal of interest can have a big impact on the quality of the climate field reconstructions produced. The generic process of data screening in dendroclimatology goes back many years of course (e.g., Fritts, 1962), so again there is no surprise here. What is more controversial is the use of precipitation-sensitive tree-ring series to reconstruct past temperature through an inverse evapotranspiration demand mediated temperature signal rather than through a direct temperature effect on tree growth. I will not dwell on this here because it appears to work okay in certain cases, e.g. Trouet et al. (2013). However, there remains some concern about how the power spectrum of temperature reconstructions based on these quite different tree growth signals may differ. Let's just say that an inverse temperature signal is not as optimal as the direct one used in Wilson et al. (2016) and should be used with caution.

We completely agree that it is rather obvious and not new that more quantity and more quality is desirable. Nevertheless, there are constantly new collections of proxy data sets published and these are and will be used to

generate climate field reconstruction because the compilation a comprehensive proxy data set is a vast amount of work and requires experts from another field. All compilations have different strength and weaknesses and specific purposes, which they should be used for. Nevertheless, precipitation sensitive proxies remain in data sets that are specifically assembled for temperature reconstructions (Emile-Geay et al., 2017), which may in some instances be alright but not ideal as pointed out by the reviewer. Hence, proxy compilations are commonly used in other than just the best suited way, for instance because new methods allow to reconstruct multivariate 3-dimensional states of the atmosphere instead of surface temperature only. Hence, we find it useful to make users of these data sets aware of possible issues because not everyone involved in the development of reconstruction methods may be an expert in the proxy input data. Furthermore, it is not so obvious, how methods can deal with mixed temperature and precipitation signals in proxies and if the transfer of information into the multivariate atmospheric state works well. That is the reason why we include skill score maps for precipitation and sea level pressure, too.

What has not been adequately considered in the paper, however, are differences in the size and location of the proxy domains used in the CFR experiments relative to the size and location of the climate field being reconstructed. For example, there is a great difference in size and location between the domain occupied by the 54 N-TREND series and the temperature domain being reconstructed. This basic issue was investigated by Kutzbach and Guetter (1980) in their classic paper on paleoenvironmental network design. It is not often cited today, yet should be mandatory reading for anyone who wishes to engage in CFR. In it, Kutzbach and Guetter (1980) show that reconstructing a large climate field from a much smaller proxy field is likely to be far less effective compared to the case where the proxy field is large and extends beyond the limits of the climate field being reconstructed. Such is clearly not the case regarding the N-TREND data used in this paper's CFR experiments.

Thank you for suggesting this important study. Obviously, we also do not expect the N-TREND data set to produce a great reconstruction outside of the area covered although our method makes use of covariances in the model simulations between spatially distant locations. Our conclusions are not meant to criticize N-TREND for covering less space. Rather the opposite, we show that having this set of best reconstruction is greatly enhancing reconstruction skill in the covered areas and that these records get most weight in our assimilation procedure and therefore strongly influence the reconstruction. However, we wanted to highlight that a combination of data sets with less strictly selected records helps in regions where such high-quality information is not available. We find it noteworthy that at the same time data sets with less high-quality information do not blur the highest-quality information in regions covered by N-TREND.

The N-TREND data are exclusively from the 40°-75°N region rather than over the much larger domains of the other two tree-ring data sets. As such, those 54 tree-ring chronologies were never intended to be used in the way done in this paper because the temperature signals in many of the N-TREND series are comparatively local and therefore most reliable at that spatial scale of the overall N-TREND domain. See Anchukaitis et al. (2017) for Part 2 of the N-TREND study and the maps contained therein. Thus, the statement in the Abstract ''. . . nor the small expert selection [N-TREND] leads to the best possible climate field reconstruction" is really quite unfair because the experiments in this paper were set up in almost the worst possible way for N-TREND to succeed well. Thus, I find the results of this study difficult to interpret because of the vastly different spatial sampling that exists between N-TREND and the other two tree-ring datasets relative to the temperature field being reconstructed.

We will rephrase the statement in the abstract and also clearly this point in the discussion to avoid such a misunderstanding. However as explained above, we intend to show the consequences if proxy data compilations are used for climate field reconstruction in a way that was not really intended but occurs in practice. Nevertheless, even if not used in an ideal manner, N-TREND appears to be the most influential and important data set for all the region, which it covers. In this sense, our message is rather that such efforts as compiling the N-TREND are extremely important and helpful. However, if we aim at a global multivariate climate field reconstruction, we can add additional information if we combine data sets without blurring the information from high quality data sets.

The authors also talk about assessments of reconstruction skill or skill improvement, but this is not really true in the classical sense where estimates are compared to actual data not used in the model calibration exercise. So, there is no true out-of-sample skill assessment made in their analyses and estimates of true reconstruction skill remain unknown. This is basically acknowledged by the authors in lines 127-128: "it must be noted that the final reconstruction is consistent only in the model world." Yet, true model validation tests could have been made by reserving a traditional validation interval for testing as is typically done in classical statistical CFR. This can be done in the context of data assimilation for CFR too as discussed in Steiger et al. (2018). The authors could, for example, calibrate the proxies only back to 1920 and check performance of the reconstructions over the withheld interval for skill and clues of overfitting. However done, some form of out-of-sample model validation testing should be mandatory when applying and testing any CFR method.

We will add an additional figure, which shows the absolute skill of the reconstruction with respect instrumental data.

However, our methods is based on transient simulations as a prior in contrast to Steiger et al. (2018). Hence, our simulations already have skill and show for instance a greenhouse gas warming in the 20th century.

As we already explained to the first reviewer: "There is a lack of independence which comes from 1) the regression model and 2) the residuals. Concerning 1), regression coefficients are estimated from gridded instrumental data sets to translate grid cell temperature (and moisture) anomalies to local tree-ring measurements. The optimization is done on tree rings, not on the climate data, and it is done on many local scales and not the large scale. In that sense the effects of the dependence are rather indirect. In contrast the statistical reconstruction methods, which directly estimate a climate variable such as temperature through the regression parameter estimate, our assimilation method is less far affected by the calibration procedure. Nevertheless, we agree with the reviewer that using the same data for validation probably leads to a slight overestimation in reconstruction skill and this is the reason, why we made additional "leave-one-out"-experiments in the publication of the original reconstruction (Franke et al., 2017). Concerning 2), we use these regression residuals as an estimate of error covariance, i.e. the larger the residuals, the smaller the weight of the proxy observation in the assimilation process. Again, the error estimate concerns tree ring width, not climate parameters.

Note that in this study we just compare the relative skill of various inputs data sets, so the impact of dependencies will be the same for all. We do not see any reason how the relative skill should be influenced by not having a fully independent validation data.

In the revised manuscript, we will explain in the methods section, why this lack of independence cannot influence the finding of this study."

More specifically, a statistic called the root-mean-square-error skill score (RE) is used in this paper to compare the relative performances of the tree-ring data sets used in the DA experiments. But there is some unwanted and unnecessary confusion here. The 'true' RE (Reduction of Error) has a long history of use in both meteorology (Lorenz, 1956) and paleoclimatology (Fritts, 1976) as a measure of skill of 'out-of-sample' forecasts and hindcasts, respectively. To use the RE as classically defined requires an explicit calibration interval for model development and estimation of its mean state (climatology) and an explicit validation interval for testing the skill of the model estimates against withheld or 'out-of-sample' data. In this case, the minimum benchmark for model skill is RE > 0, i.e. skill > climatology. This does not appear to be the case here. Rather the authors seem to be using the model ensemble mean without proxy assimilation as the reference. As such, there are not any explicitly defined calibration and validation intervals, and the authors are just assessing whether the simulations that assimilate the proxies do better than simulations that are merely forced with SSTs. Thus, the RE in this paper is very different from the classical RE of Lorenz (1956) and Fritts (1976) and should be called something else to avoid confusion.

The idea of the "reduction of error" (RE) is to compare the error of a forecast to the error of a reference forecast. (Lorenz, 1956) used the root mean square error (RMSE) and climatology as a reference forecast. The same concept is also known as RMSE SS (Skill Score, Wilks, 2011) where the reference cannot only be climatology but also persistence or as in our case a transient model simulation forced not only by SSTs but solar variability, aerosols, land use changes, greenhouse gases, etc. and which already have RMSE SS > 0. Hence, achieving an RMSE SS > 0 in our case is much harder than just being better than climatology.

To avoid any confusion with the RE and CE definition used in the tree-ring community, we will call this skill score RMSE SS in the revised version.

References

Emile-Geay, J., McKay, N. P., Kaufman, D. S., Gunten, von, L., Wang, J., Anchukaitis, K. J., Abram, N. J., Addison, J. A., Curran, M. A. J., Evans, M. N., Henley, B. J., Hao, Z., Martrat, B., McGregor, H. V., Neukom, R., Pederson, G. T., Stenni, B., Thirumalai, K., Werner, J. P., Xu, C., Divine, D. V., Dixon, B. C., Gergis, J., Mundo, I. A., Nakatsuka, T., Phipps, S. J., Routson, C. C., Steig, E. J., Tierney, J. E., Tyler, J. J., Allen, K. J., Bertler, N. A. N., Björklund, J., Chase, B. M., Chen, M.-T., Cook, E., de Jong, R., DeLong, K. L., Dixon, D. A., Ekaykin, A. A., Ersek, V., Filipsson, H. L., Francus, P., Freund, M. B., Frezzotti, M., Gaire, N. P., Gajewski, K., Ge, Q., Goosse, H., Gornostaeva, A., Grosjean, M., Horiuchi, K., Hormes, A., Husum, K., Isaksson, E., Kandasamy, S., Kawamura, K., Kilbourne, K. H., Koç, N., Leduc, G., Linderholm, H. W., Lorrey, A. M., Mikhalenko, V., Mortyn, P. G., Motoyama, H., Moy, A. D., Mulvaney, R., Munz, P. M., Nash, D. J., Oerter, H., Opel, T., Orsi, A. J., Ovchinnikov, D. V., Porter, T. J., Roop, H. A., Saenger, C., Sano, M., Sauchyn, D., Saunders, K. M., Seidenkrantz, M.-S., Severi, M., Shao, X., Sicre, M.-A., Sigl, M., Sinclair, K., St George, S., St Jacques, J.-M., Thamban, M., Thapa, U. K., Thomas, E. R., Turney, C., Uemura, R., Viau, A. E., Vladimirova, D. O., Wahl, E. R., White, J. W. C., Yu, Z. and Zinke, J.: Data Descriptor: A global multiproxy database for temperature reconstructions of the Common Era, Sci. Data, 4, doi:10.1038/sdata.2017.88, 2017.

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, 170076, doi:10.1038/sdata.2017.76, 2017.

Franke, J., Frank, D., Raible, C. C., Esper, J. and Brönnimann, S.: Spectral biases in tree-ring climate proxies, Nature Climate change, 3(4), 360–364, doi:10.1038/nclimate1816, 2013.

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., Steiger, N. and Perkins, W. A.: The last millennium climate reanalysis project: Framework and first results, J. Geophys. Res. Atmos., 121(1), 6745–6764, doi:10.1002/2016JD024751, 2016.

Lorenz, E. N.: Empirical Orthogonal Functions and Statistical Weather Prediction. 1956.

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, 180086–15, doi:10.1038/sdata.2018.86, 2018.

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., Anderson, D. M., Steig, E. J. and Noone, D.: Last Millennium Reanalysis with an expanded proxy database and seasonal proxy modeling, Climate of the Past, 15(4), 1251–1273, doi:10.5194/cp-15-1251-2019, 2019.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, Academic Press. 2011.