The importance of input data quality and quantity in climate field reconstructions – results from the assimilation of various tree-ring collections

Franke, Jörg[1,2], Valler, Veronika[1,2], Brönnimann, Stefan[1,2], Neukom, Raphael[1,2,3,4], Jaume Santero, Fernando[5]

[1] Institute of Geography, University of Bern, Switzerland.

[2] Oeschger Centre for Climate Change Research, University of Bern, Switzerland.

[3] Department of Geography, University of Zurich, Switzerland.

[4] Department of Geosciences, University of Fribourg, Switzerland

[5] Universidad Complutense de Madrid.

Correspondence to: Jörg Franke (franke@giub.unibe.ch)

Abstract. Differences between paleoclimatic reconstructions are mainly caused by two factors, the method and the input data. While many studies compare methods, we will focus in this study on the consequences of the input data choice in a state-of-the-art Kalman-filter paleo data assimilation approach. We evaluate reconstruction quality in the 20th century 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. The three data sets cover the range from small sample size, small spatial coverage and strict screening for temperature sensitivity to large sample size and spatial coverage but no screening. Additionally, we explore a combination of these data sets plus screening methods to improve the reconstruction quality.

A large, unscreened collection leads generally to a poor reconstruction skill. A small expert selection of extratropical northern hemisphere records allows for a skillful high latitude temperature reconstruction but cannot be expected to provide information for other regions and other variables. We achieve the best reconstruction skill across all variables and regions by combing all available input data but rejecting records with insignificant climatic information (p-value of regression model > 0.05) and removing duplicate records. It is important to use a tree-ring proxy system model that includes both major growth limitations, temperature and moisture.

1 Introduction

In the past 20 years, a lot of effort has been invested in improving climate reconstructions for the last centuries to millennia based on indirect climate information – so-called "proxies". Focus has been on both, large-scale averages as well as the reconstructions of regional to global fields (Masson-Delmotte et al., 2013; Smerdon and Pollack, 2016). Temporal and spatial resolution varied with the included paleoclimatic archives. However, most reconstructions for the past centuries rely heavily on the most abundant indirect climate archive, tree rings, and specifically on tree-ring width (TRW) and late-wood density (MXD). Differences between reconstructions have

mostly been discussed with differences in reconstruction methodology in mind (Christiansen and Ljungqvist, 2017). However, a new study shows good agreement between a wide range of methods, if reconstructions are based on the same input data set (Neukom et al., 2019a; 2019b). Another recent study found that temperature sensitive tree-ring proxies from the PAGES2k database (Emile-Geay et al., 2017) lack multi-centennial trends, which are found in other proxy archives (Klippel et al., 2019). This suggests that the input data play a crucial role for differences between reconstructions. This fact is also seen in data assimilation for weather prediction, e.g. at the switch from radiosonde to satellite observations (Swinbank et al., 2012, p.365). Today, many proxy data archives are available, hence compiling input data for reconstruction is not only a matter of the amount of proxy data, but also of their selection, i.e., screening.

In this study, we therefore aim at evaluating the effect of various tree-ring data collections and their screening on the final reconstructions. Tree-ring proxies are by far the most numerous climate information source for the past centuries and additionally chosen because our methodology relies on annual data without dating uncertainties. Due to the relevance of temperature in the climate change discussion and the fact that many biological proxies react to temperature stress, temperature has so far been the variable of most interest. However, to study the underlying processes a multi-variable perspective is required. Therefore, we evaluate the effects of the input data choice, using a state-of-the-art data assimilation technique, which allows for multi-variable climate reconstructions in form of model simulations that are in optimal agreement with proxy information (Bhend et al., 2012; Franke et al., 2017).

Previous studies based on data assimilation techniques proposed that a higher quantity of input data would always be beneficial (Matsikaris et al., 2016; Steiger et al., 2018; Tardif et al., 2019). The idea is that regression-based proxy system models weight each proxy series by their regression residuals. Hence, proxies that do not contribute information will be downweighted automatically. However, this weighting may not work perfectly because of two factors: (1) the regression depends on overlapping paleodata and instrumental measurements, which often results in a small sample (Fig. 1 in Jones et al., 2012), uncertain residuals and possible model overfitting; (2) moisture and temperature sensitive proxies may correlate and hence moisture sensitive paleodata will be used to correct temperature and vice versa. However, these two variables probably have very different multi-decadal to centennial variability (Franke et al., 2013). The growth limiting factor may even change over time (Babst et al., 2019).

In this study, we use the Kalman filter based state-of-the-art data assimilation technique introduced in Bhend et al. (2012), which is very similar to the methodology used in the last millennium reanalysis (LMR) project (Hakim et al., 2016; Tardif et al., 2019). In contrast to LMR, our method is a transient-offline method, in which the background state is time-dependent due to the external forcing prescribed to the climate model simulations. In our experiments, we focus on the effect of the input data choice on the final reconstruction. We compare three published collections of tree-ring records (focusing on TRW and MXD), of which at least two are commonly used for climate reconstructions. These have very different characteristics: (1) The B14 collection of 2287 consistently detrended TRW chronologies from the International Tree Ring Data Base (ITRDB), not screened for climate sensitivity (Breitenmoser et al., 2014); (2) TRW and MXD series from the PAGES2K database version 2 (Emile-Geay et al., 2017), with a selection of 415 temperature sensitive records, most selected by a statistical screening for positive correlation with instrumental temperature; and (3) the N-TREND tree-ring collection of 54 TRW,

MXD or blended TRW-MXD time series (Wilson et al., 2016), selected by experts to be the best temperature recorders. Thus, the three data sets cover the range from large sample size and spatial coverage but no screening for temperature sensitivity to small sample size and small spatial coverage but strict screening. Note, that these collections were generated with slightly different aims, which affects their use in reconstructions. Thus, for instance, we cannot expect to achieve the best global scale field reconstruction already from a proxy collection covering a much smaller area (Kutzbach and Guetter, 1980). However, all data sets are used for climate reconstruction.

In the next section the method and data sets are introduced in greater detail before we show our results. Then we discuss the possible reasons for our results and the differences compared to previous studies. Finally, we draw our conclusion how an optimal proxy selection process should look like.

2 Data and Methods

We use three input data sets for comparison, all consist of annually resolved tree-ring measurements, which have hardly any dating uncertainties:

1. B14 is a collection by Breitenmoser et al. (2014) of 2287 uniformly detrended and standardized TRW measurements from the ITRDB (Zhao et al., 2018). We use the full collection without any further screening for climate/temperature sensitivity. Hence, this represents the data set with the highest quantity of records. However, the weighting of temperature information in the paleodata is completely up to the reconstruction method.
2. PAGES2k is a collection of 415 TRW and MXD series from PAGES2k data base version 2 (Emile-Geay et al., 2017). These are all records that correlate significantly ($p<0.05$) with nearby instrumental temperature measurements and/or have been described by experts to represent temperature variability. This compilation represents a compromise of good quantity, large spatial coverage and good quality paleodata, based on global selection criteria. However, experts from various regional groups were differently strict in their screening procedure, which led to varying data density in the different regions.
3. N-TREND is a collection of 54 tree-ring chronologies based on TRW, MXD or a combination of both. They were chosen by experts to be just the best tree-ring paleodata for temperature reconstructions (Wilson et al., 2016). Hence, they are our low quantity, highest quality input data set with least spatial coverage.

Climate fields are reconstructed by assimilating these tree-ring observations into an ensemble of climate model simulations using a Kalman filter technique, Ensemble Kalman Fitting (Bhend et al., 2012; Franke et al., 2017). The simulations, which serve as a background (sometimes called first guess or prior) of the atmospheric state at each point in time, are given by a 30-member initial condition ensemble of atmospheric model simulations (ECHAM5.4, (Roeckner, 2003)). All simulations follow the same external forcings (volcanic (Crowley et al., 2008), solar (Lean, 2000), greenhouse gases (Yoshimori et al., 2010), land use (Pongratz et al., 2008), tropospheric aerosols (Koch et al., 1999)) and sea surface temperatures boundary conditions based on a reconstruction by(Mann et al., 2009) plus additional El Niño Southern Oscillation variability (Franke et al., 2017). The data assimilation method is “transient offline”. “Transient” refers to the fact the our prior at each point in time consists of 30 ensemble members that are in agreement with forcings and boundary conditions. “Offline” assimilation means

that the simulations are calculated for the full period before the assimilation is conducted. This is possible in the paleo-climatological setup because we have only one observation per year. Predictability on these time scales only comes from the boundary conditions and not from the atmospheric model.

EKF is the offline variant of the ensemble square root filter (Whitaker and Hamill, 2002) in which the observations (y) are assimilated serially. The assimilation procedure is divided into an update of the ensemble mean ($\bar{x}$) and an update of the anomalies with respect to the ensemble mean (x'):

(1) $$\bar{x}^a = \bar{x}^b + K(\bar{y} - H\bar{x}^b)$$

(2) $$x'^a = x'^b + \widetilde{K}(y' - Hx'^b) = \left(I - \widetilde{K}H\right)x'^b \; with: y' = 0$$

where the superscript [a] refers to the analysis and [b] to the background of the atmospheric state x, which is a vector with values of multiple variables at all grid boxes. H denotes an operator which maps x^b to the observation space (see proxy system model PSM below). K and $\widetilde{K}$ are the Kalman gain matrices (Whitaker and Hamill, 2002):

(3) $$K = P^b H^T (HP^b H^T + R)^{-1}$$

(4) $$\widetilde{K} = P^b H^T \left[\left(\sqrt{HP^b H^T + R}\right)^{-1}\right]^T \times \left(HP^b H^T + R + \sqrt{R}\right)^{-1}$$

The K matrices control how the information from the observations updates the background. It depends on the observation error covariance matrix R and the background error covariance matrix P^b. R is estimated from the regression residuals of the PSM and errors are assumed to be uncorrelated. Background error covariances P^b are calculated from the 30-member ensemble n_{ens} of ECHAM5.4 simulations (CCC400) at each time step with row number i, column number j and ensemble member k (Bhend et al. 2012, Franke et al. 2017):

(5) $$P^b_{i,j} = \frac{1}{n_{ens}-1} \sum_{k=1}^{n_{ens}} x'^b_{i,k}\, x'^b_{j,k}$$

This has the advantage of taking time-dependent covariance structures into account, for instance during El Niño vs La Nina years. The disadvantage is the small sample of 30 ensemble member for covariance estimation. To deal with spurious correlations caused by the relatively small ensemble, we apply a distance-dependent localization, i.e. updates are only possible with a certain radius around the observations (Valler et al., 2019):

(6) $$C = \exp\left(-\frac{|d_i - d_j|^2}{2L^2}\right)$$

d_i and d_j describe the zonal and meridional distances from the selected grid box. L is the length scale parameter used for localization. It has been estimated based on the spatial correlation in the simulations and is variable dependent, e.g. 1500/450 km in case of temperature/precipitation (Franke et al., 2017).

A recent comparison of Valler et al. (2018) has shown superior performance when using an improved covariance estimation, which blends 50% of the 30-member time-dependent covariance with 50% of a 250-member "climatological" time-independent covariance (Experiment: 50c_PbL_Pc2L in Valler et al. (2018)). In this paper we use both the original setting as in Franke et al. (2017) as well as the improved setting proposed by Valler et al. (2018).

Our paleo-reanalysis is based on anomalies from a 71-year period around the current year. Low frequency

variability is a function of the models' response to the prescribed external forcings and background conditions, which include sea-surface temperatures. Because low frequency variability is not consistently preserved in paleodata (Franke et al., 2013; Klippel et al., 2019), but reasonably well represented in the model simulations of the last millennium (Franke et al., 2017), this approach is expected to provide consistent skill at all time scales. Note that the assimilation of anomalies retains possible model biases. This circumvents a big problem in data assimilation approaches with temporally varying input data networks. Observations that gradually pull the model away from its biased state, can lead to artificial trends or step functions in time-series.

We use a linear multiple regression PSM to simulate tree-ring observations using modelled temperature and/or precipitation. The regression model is calibrated with gridded instrumental data (CRU TS 3.1, Harris et al., 2014) in the period 1901-1970. It includes monthly temperature (and precipitation) during the growing season April to September. In this study, we limit the analysis to the northern hemisphere because the majority of the tree-ring observations can be found there. In the first four experiments (Table 1), which only use temperature (T) in the PSM, we have 6 independent variables (i.e., local, monthly mean temperature of April to September). If we assume that tree growth was limited by temperature and moisture (TR) variability (experiments 5 and 6 in Table 1), we have 12 independent variables (i.e., local, monthly mean temperature of April to September, and monthly precipitation sums of April to September). Note that regression coefficients can be zero and thus growth can still be limited to just temperature or just precipitation and to less than 6 months. In experiments 7 and 8 (Table 1) we additionally consider only regression models, in which the growth occurs in consecutive months. Therefore, we fit all possible combinations of consecutive months and choose the PSM with the lowest Akaike information criterion (AIC). Temperature and precipitation limitations can occur in a different sequence of months for each variable (e.g. precipitation limits growth from April to June and temperature limits growth from June to September). The variance of the regression residuals is used to specify the observation error covariance matrix (assumed diagonal) in the assimilation, i.e. the larger the residuals, the less weight an observation gets and the less the model simulations get corrected.

In this study, the period in which the regression coefficients of the PSM are estimated as well as the calculations of the regression residuals overlaps with the period when the reconstruction skill is estimated. This apparent lack of independence is negligible in this case because: 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. In that sense the effects of the dependence are rather indirect. In contrast to statistical reconstruction methods, which directly estimate a climate variable such as temperature through the regression parameter estimate, our assimilation method is far less affected by the calibration procedure. Nevertheless, using the same data for validation probably lead to a slight overestimation in reconstruction skill. However, 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. Concerning the regression residuals, again the error estimate concerns tree ring width, not climate parameters. We use the residuals as an estimate of error covariance. In case we slightly underestimate the residuals, proxy observation would have a little too much weight in the assimilation process compared to the simulations.

If multiple data collections are combined, there may be duplicates of the same proxy, possibly in differently treated/detrended versions. We conduct experiments where we prevent single sites from being assimilated twice

by only choosing the best proxy (smallest regression residuals irrespective of series length) in a 0.1ºx0.1º (ca. 10km) grid. This is a rare case, however, and hardly effecting the results.

We evaluate the quality of the reconstruction based on correlation with gridded instrumental observations of temperature, precipitation (Harris et al., 2014) and sea level pressure (Allan and Ansell, 2006) in the period 1901-1990 as a reference (x^{ref}, where x is the state vector). After showing absolute correlation coefficients of the analysis, we focus on correlation improvements over the original model simulations, because these forced simulations already correlate positively with the gridded observations in many locations. Correlation focuses on the co-variability, i.e. the correct sign of the anomaly. Additionally, we use a root-mean-square-error skill score *RMSESS* that describes the improvement of the analysis x^a over the original model simulations (background) x^b over all time steps i:

$$RMSESS = 1 - \frac{\Sigma\left(x_i^a - x_i^{ref}\right)^2}{\Sigma\left(x_i^b - x_i^{ref}\right)^2} \quad (7)$$

It is more difficult to reach positive *RMSESS* values than correlation improvements, because this score penalizes a wrong amplitude of variability (e.g., an uncorrelated reconstruction with correct variance would yield *RMSESS* = -1). Because it is based on squared errors, too high variability is penalized more than little variability, which the ensemble mean of the original model simulations has. We only present correlation improvements and *RMSESS* of the ensemble mean. In contrast to correlation coefficients, which tend to be higher for the ensemble mean than for the ensemble members, *RMSESS* of single ensemble members tends to be slightly higher than *RMSESS* of the ensemble mean (Fig. 6 in (Bhend et al., 2012)).

To evaluate the influence of the input data on the final reconstruction, we conducted the following set of experiments:

Table 1: Experiments

	Name	Proxy system model	Description
1.	NTREND_T	6 regression coeff. for Apr. to Sep monthly temperature (T)	Using the best tree-ring chronologies for temperature reconstruction, which have been chosen by experts, i.e. very strict selection of few, best records
2.	PAGES_T	Same as above	Using a selection of temperature sensitive proxies selected by the regional PAGES working groups. The mostly statistical screening for a temperature signal involved a sign correction, i.e. if temperature and moisture are negatively correlated, tree-ring chronologies can remain as temperature sensitive in the data set. Therefore, more records but less strictly screened than NTREND.
3.	B14_T	Same as above	Consistently detrended tree-ring data from the ITRDB by B14. This proxy set includes the largest amount of proxy series. However, many of them do not include any climate signal.
4.	ALL_T	Same as above	All three data sets together, largest data set with greatest spatial coverage. However, duplicate proxies cannot be excluded
5.	ALL_TR	12 regression coeff. For Apr. to Sep. monthly	Same as above

		temperature and precipitation (R)	
6.	ALL_TR_scr0.05	Same as above	Same as above but with additional screening, i.e. only records with a climate signal (p-value < 0.05) will be assimilated. In the experiments above these series got little weight due to large errors (regression residuals) but were still assimilated.
7.	ALL_TR_scr0.05_ AIC_NOdup	Maximum of 12 regression coefficients but only consecutive months are allowed, still mixed temperature and precipitation signals are possible	Same as above but we chose with the AIC the regression model under the precondition that only climate from consecutive months can influence tree growth, which is more realistic due to local growing season length. Additionally, we remove duplicate proxies by only considering the best proxy (lowest regression residuals) within a 0.1°x0.1° (ca. 10 km) grid. In each grid box we keep both, the best mainly temperature limited and the best mainly moisture sensitive proxy if both exist.
8.	ALL_TR_scr0.05_ AIC_NOdup_ClimCovar	Same as above	Same as above but with background error-covariance estimate not only from the 30 ensemble members of the current year. Instead we use a mix of 50% error covariance coming from 250 random ensemble members and years.

3 Results

Temperature correlation coefficients between the analyses and gridded instrumental data are positive nearly all around the globe and for all three proxy collections (Fig. 2a,b,c) because the transient simulations follow forcings and boundary conditions and hence show proper multidecadal variability and a 20th century warming trend. However, this is not the case for precipitation, which does not show a warming trend (Fig. 2d,e,f). In contrast to the assimilation of PAGES and NTREND (Fig. 2d,e), we can observe clearly higher correlations in the United States if the B14 proxies are assimilated (Fig. 2f). Although these first three experiments only use a temperature PSM, information can spread to other variables through the covariance matrix.

To evaluate the differences between the experiments due to the data assimilation we focus on correlation improvement over the background (i.e. the model simulations, which already correlate with the reference data set mainly due to the specified SSTs and external forcing). First, we compare the role of the choice of the three input data sets assuming only temperature dependence and no constraint on the regression model structure (Fig. 3a,b,c; experiments NTREND_T, PAGES_T and B14_T). The highest local improvements are reached with the NTREND data set, however the largest spatial coverage of improvement is found with the B14 data set. Note that temperature correlation improves with all data sets and decreases nowhere, although some proxy records in the B14 data set do not contain any temperature signal. This has been identified with negative regression coefficients for the majority of B14 tree-ring series in the United States of America. In terms of correlation the data assimilation scheme appears to weight the input data appropriately. Looking at precipitation correlation improvements (Fig. 4a,b,c), we find hardly any improvements with the NTREND collection. In contrast, the B14 data set leads to some precipitation correlation improvements over North America, where no NTREND series are located.

The correct sign of the anomaly, measured by correlation, is only telling us one aspect of the reconstruction quality. To see if the amplitude of the anomaly is also reconstructed correctly, we look at the *RMSESS* skill score (see methods). Here, we find large differences between the proxy collections. With NTREND_T we find improvements everywhere, whereas B14_T shows more regions with negative than positive skill (note that we use PSM with only temperature). The PAGES data set is again in the middle. This suggest that using moisture sensitive proxies to reconstruct temperature as in B14_T, which works just because temperature and precipitation are correlated at a given location, is not ideal. Hence, further experiments with an improved PSM and upgraded screening procedure were conducted to take the proxies' temperature or moisture sensitivity better into account and to find an option to use the PAGES and B14 collection at locations, where no expert selected proxies are available but rather keep the quality of the expert selected data, where it is available.

Before we come to a more sophisticated PSM and more sophisticated input data screening, we simply combine all three data sets using still a model with only temperature (ALL_T). This experiment performs well. Temperature correlation now reaches levels of the NTREND_T experiment, where NTREND data is available and additionally correlation improvements cover the regions, where only PAGES or B14 have data. *RMSESS* values are positive in most regions, too. However, around India and the Himalaya as well as in the US southwest, skill is negative. Precipitation correlations improved only marginally (Fig. 6d) and precipitation *RMSESS* (Fig. 7d) is mostly negative.

The obvious change to improve precipitation reconstruction skill is to use a PSM that includes precipitation, i.e. a multiple regression model with 12 coefficients for temperature and precipitation influence during the 6-months growing season (experiment ALL_TR). Temperature skill remains at the same high level, but precipitation skill clearly improves (Fig. 6e and 7e). Correlations improve everywhere and *RMSESS* values become positive in most regions with the exception of the Himalaya region and most northeast of Russia.

So far, we have not excluded any proxies from the data assimilation. We trust that proxies with no or a weak climate signal simply have regression coefficient close to zero and large residuals. This way they hardly affect the analysis. However, in a regression model with 12 independent variables and only 70 years of overlapping data, some records may just by chance get more weight than they deserve. Therefore, our next step is the introduction of a weak screening. In a first step, we only assimilate proxies with p-values < 0.05 for the full regression model (ALL_TR_scr0.05). This removes ca. 16% of the proxies and hardly affects correlations (Fig. 3f, 4f, 5f) but removes most of the negative *RMSESS* values in both, temperature and precipitation (Fig. 6f and 7f).

This result appears good, but this could also be a result of overfitting the regression model because any combination of growing season months was allowed to affect tree growth. It would not make physiological sense if a tree would be limited for instance by May, July and September temperatures but not by June and August temperatures. Hence, the next step is to further constrain the model. The tree growth should be affected by climate conditions in a locally varying growing season of consecutive months. We fit all possible combinations of temperature and precipitation influences in consecutive months and chose the model with the lowest AIC (note that additionally duplicates are removed; experiment ALL_TR_scr0.05_AIC_NOdup). As a result of this more physically based growth model, reconstruction skill decreases slightly (Fig. 6g and 7g). This suggests that the previously noted improvement in *RMSESS* was indeed partly due to overfitting. Nevertheless, correlations remain

on the same high level everywhere (Fig. 3g, 4g, 5g). Only *RMSESS* decreases in some regions with a high number of paleodata such as parts of China and parts of North America.

Recently, Valler et al. (2018) could show that major improvements of the method used in this study can be achieved by using a background error covariance matrix, which is not only calculated from the 30 ensemble members for the current year (Franke et al. 2017) but blended with a climatological error covariance matrix based on random years and ensemble members from the original model simulations (see methods, experiment ALL_TR_scr0.05_AIC_NOdup_ClimCovar). Using improved covariance information increases *RMSESS* values again and only very few grid boxes with negative skill remain. Moreover, the largest effects of the better error covariance estimation appear in variables that have not been assimilated such as sea level pressure (Fig. 5h). This is very important because one of the reasons for using data assimilation instead to traditional statistical reconstruction techniques is the possibility to gain knowledge about further variables in a physically consistent way, which allows for a better dynamic interpretation of the identified climatic variations.

4 Discussion

Correlations of the reconstructions with temperature improved as it would be expected after the assimilation of the three data sets and using a temperature PSM. We calculate the regression coefficients based on instrumental temperature. Hence, all proxies that correlate in some way with instrumental temperature will be used to update the analysis temperature. The analysis has highest correlations improvements with instrumental temperature if the proxies themselves had highest correlations, which is the case for the NTREND data set with the best temperature proxies only. Correlations improvements are lower but cover a larger area with the B14 collection.

Note that correlation improvements can be a result of a negative relationship between tree-ring width and instrumental temperature if local growth is moisture limited and growing season temperature and precipitation are negatively correlated. This can be a benefit because through the covariance we use the extra information that dry summers are also warm and vice versa. Hence, we find much better precipitation correlation with the B14 collection than with the NTREND data set. However, using moisture sensitive trees to update temperature fields may cause problems. Precipitation variability shows high inter-annual variability in many locations but neither the same inter- to multi-decadal variability as temperature nor its centennial trend (Hartmann et al., 2013; Landrum et al., 2013). Although we cannot study this with our methods, another study shows that updating other than the assimilated variables through the covariance matrix can cause problems on longer than inter-annual scale (Tardif et al., 2019).

The regression model is calibrated on the interannual time scale assuming that TRW limitations are time-independent. However, this may not be the case (Babst et al. 2019), and therefore decadal-to-multidecadal variability may be less well represented. A similar argument holds for the update introduced by the model covariance matrix, which, although state dependent, may yield optimal estimates only for seasonal and not decadal time scales. However, our approach avoids these pitfalls in two ways. First, at multidecadal and longer time-scales, the model takes over, and therefore relations in our reconstructions are not constrained to be stationary across time scales. Furthermore, with our approach, the stationarity assumption is restricted to the regression model, thus it is a local stationarity - no further stationarity assumption concerning spatial variability is introduced except for experiment ALL_TP_scr0.05_AIC_NOdup_ClimCovar, where 50% of the background error

covariance matrix is climatological and thus stationary. Most other approaches assume stationary spatial covariances.

Theoretically, it would be optimal to assimilate all available data and let each record be weighted based by its error. However, the true observation error is unknown and its estimation is uncertain. In our case, we use a multiple regression proxy-system model with 6/12 variables (six months of temperature and optionally six months of precipitation) in a 70-year period of overlapping instrumental data and proxy measurements to estimate regression coefficients. This rather short period and large number of independent variables can lead to overfitting the model and thus underestimating the observation error, which is defined by the regression residuals. Together with the low signal-to-noise ratio of many tree-ring chronologies, this can lead to an over- or under-correction of the model field in the assimilation step. An additional experiment with doubled observation error (not shown) increases *RMSESS* values clearly. This suggests that PSM overfitting is part of the reason for the negative *RMSESS* skill scores in the B14_T experiment in contrast to the NTREND_T experiment (Fig. 4a and c).

In the following experiments (ALL_TR_scr0.05, ALL_TR_scr0.05_AIC_NOdup, ALL_TR_scr0.05_AIC_NOdup_ClimCovar) we tried to reduce the consequences of uncertain error estimates step by step. Excluding proxies without a significant climate signal ($p<0.05$) for the full regression model, clearly improves the *RMSESS* skill score for temperature and precipitation in large parts of Asia (Fig. 6f and 7f). This highlights the negative effects of spurious correlation – even if it is very weak – on the analysis. Hence, screening the data appears to be important, especially in data sparse regions, where there is no chance for better records with smaller errors to correct errors introduced due to spurious covariances. In other reconstruction methods, for instance principal component regression or the search for the best analogs, screening of records will additionally be necessary to avoid spatial biases due to non-homogeneous proxy distributions (Bradley, 1996; Rutherford et al., 2005). However, this is negligible in the data assimilation framework because the number of assimilated records has a regional instead of global impact and because the method provides a measure of uncertainty in form of ensemble spread at each grid cell.

In the experiment, in which we only allow for a single growing season (ALL_TR_scr0.05_AIC_NOdup) per year instead of a statistically optimal selection of months and by removing duplicate records that are in more than one of the data collections, correlations improve slightly but *RMSESS* decreases slightly. Obviously, we continue with this more realistic setup, but note that the choice what is "best" depends on the chosen statistic or the reconstruction characteristics that are wished by the user. For instance, correlation just measures covariance whereas *RMSESS* is based on squared errors and hence penalizes especially large biases, i.e. it favors an underestimation variability over an overestimation.

Finally, we introduce an improved background error covariance estimation scheme (ALL_TR_scr0.05_AIC_NOdup_ClimCovar, Valler et al. 2019). Because assimilated information is spread in space and in between variables through the covariance matrix, it is important to estimate covariances well. Estimating covariance from both, the 30 members at the current time step and from climatology and then blending both information, especially improves our results for variables, which have not been assimilated such as sea level pressure (Fig. 5h).

In reality, climate signals in tree-ring proxies may be even more complicated than a function of moisture availability and growing season temperature. Limiting factors may change over time (Babst et al., 2019) or light

availability may be important and not always be highly correlated with temperature, i.e. more diffuse light after volcanic eruptions may stimulate growth (Stine and Huybers, 2014). More sophisticated proxy system forward models such as VS-Lite (Tolwinski-Ward et al., 2011) could be used in data assimilation (Acevedo et al., 2016; Dee et al., 2016). In fact, we have applied VS-lite to all TRW records in B14 (Breitenmoser et al., 2014). Although these model are more realistic and represent for instance non-linear responses, they introduce new problems mainly related to model biases. This currently prevents them for being used more broadly (Dee et al., 2016).

Finally, we tested the order of assimilated data, because we assimilate data serially. In combination with using covariance localization, the order could influence the final reconstruction (Greybush et al., 2011). Assimilating the data from the best to worst record in terms of regression residuals and in opposite order from worst to best, hardly influenced correlation and *RMSESS* skill scores (not shown). Hence, we continue to assimilate records starting with the best ones, similar to traditional reanalysis, which sort observations from the largest to smallest expected variance reduction in the reanalysis (Slivinski et al., 2019; Whitaker et al., 2008).

Although our results are specifically valid only for the data assimilation method used herein, it is likely that methods with a similar structure, i.e. using PSMs and variations of Kalman filters, will have similar sensitivities to the selection of input data. We expect that they are even valid for most climate field reconstruction techniques, because the basic principles of transferring proxy information to climatic variables and dealing with errors share common concepts across these methods. Even though all such methods include some routines to separate climatic information from non-climatic noise, in practice results can almost always be improved by pre-selecting the records with the highest information content, independent from the reconstruction technique applied (Neukom et al., 2019a; 2019b; Smerdon and Pollack, 2016 and references therein). This suggests that our results are qualitatively transferrable to climate field reconstruction methods in general.

Conclusion

In this study, we use existing proxy data collections to generate climate field reconstructions, as it is common practice. We are aware that this is not in all cases the main aim for which these data collections were compiled. Hence, we want to highlight the consequences of using the data set for field reconstructions. These results are not meant the rank any data set above another. Disadvantages of one data set in our setup are most probable a result of unintended usage.

How to choose input data for paleo data assimilation? We address this question by comparing three paleodata compilations of different sizes as well as using all data sets together in combination with various screening approaches.

Just using a large collection of proxy data (B14) does not lead to a skillful reconstruction. In contrast, just using a small expert selection of the best temperature proxies (NTREND) leads to a good high latitude temperature reconstruction but wastes the potential of modern data assimilation technique to reconstruct the 4-dimensional multi-variate state of the atmosphere. However, simply combing all available input data and leaving the weighting completely to a statistical model does not lead to optimal results, either. Rejecting records without a clear climatic signal, removing duplicates and using a physically plausible PSM altogether lead to a better reconstruction.

Hence the answer to our research question if it is better to assimilate all available proxy data or just the best expert selection has to be answered with: neither of the two is optimal. We achieve the best results in terms of correlation and *RMSESS*, if we use a large collection of proxy records. However, to make proper use of input data, which was not screened by experts, it is crucial to:

1. use proxy system models that properly represent the paleodata, here taking possible temperature and moisture limitations of tree growth into account.
2. use correct physical assumptions, in our case about tree growth, to avoid statistical overfitting.
3. remove input data with random, not significant climate signals.
4. care about overfitting (underestimation of errors)

For a future project, it would be very interesting to study how different reconstruction methods handle these three differently screened data sets to see, if these results are valid for other reconstructions methods, too?

Author contribution

JF had the initial idea for this paper and performed most of the analysis and drafted the manuscript. VV contributed to the code development. SB helped to shape the manuscript and experimental design. RN contributed additional analysis and all authors provided critical feedback and contributed to the writing of the manuscript.

Competing interests

There are no competing interests present.

Acknowledgements

This project was supported by the Swiss National Science Foundation project 162668 (RE-USE) and EU ERC project 787574 (PALAEO-RA). We like to thank CSCS for their support in conducting the ECHAM simulations.

References

Acevedo, W., Reich, S. and Cubasch, U.: Towards the assimilation of tree-ring-width records using ensemble Kalman filtering techniques, Climate Dynamics, 46(5), 1909–1920, doi:10.1007/s00382-015-2683-1, 2016.

Allan, R. and Ansell, T.: A New Globally Complete Monthly Historical Gridded Mean Sea Level Pressure Dataset (HadSLP2): 1850–2004, JCLI, 19, 5816–5842, doi:10.1175/JCLI3937.1, 2006.

Babst, F., Bouriaud, O., Poulter, B., Trouet, V., Girardin, M. P. and Frank, D. C.: Twentieth century redistribution in climatic drivers of global tree growth, Science Advances, 5(1), doi:10.1126/sciadv.aat4313, 2019.

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

Bradley, R. S.: Are there optimum sites for global paleotemperature reconstruction?, in Climatic Variations and Forcing Mechanisms of the Last 2000 Years, vol. 3, pp. 603–624, Springer, Berlin, Heidelberg, Berlin, Heidelberg. 1996.

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(2), 437–449, doi:10.5194/cp-10-437-2014, 2014.

Christiansen, B. and Ljungqvist, F. C.: Challenges and perspectives for large-scale temperature reconstructions of the past two millennia, RG, 55(1), 40–96, doi:10.1002/2016RG000521, 2017.

Crowley, T., Zielinski, G., Vinther, B., Udisti, R., Kreutz, K., Cole-Dai, J. and Castellano, E.: Volcanism and the little ice age, PAGES news, 16, 22–23, 2008.

Dee, S. G., Steiger, N. J., Emile-Geay, J. and Hakim, G. J.: On the utility of proxy system models for estimating climate states over the common era, Journal of Advances in Modeling Earth Systems, 8(3), 1164–1179, doi:10.1002/2016MS000677, 2016.

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.

Greybush, S. J., Kalnay, E., Miyoshi, T., Ide, K. and Hunt, B. R.: Balance and Ensemble Kalman Filter Localization Techniques, Monthly Weather Review, 139(2), 511–522, doi:10.1175/2010MWR3328.1, 2011.

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.

Harris, I., Jones, P. D., Osborn, T. J. and Lister, D. H.: Updated high-resolution grids of monthly climatic observations - the CRU TS3.10 Dataset, Int. J. Climatol., 34(3), 623–642, doi:10.1002/joc.3711, 2014.

Hartmann, D. L., Tank, A. K., Rusticucci, M., Alexander, L. V., Brönnimann, S., Charabi, Y., Dentener, F. J., Dlugokencky, E. J., Easterling, D. R., Kaplan, A., Soden, B. J., Thorne, P. W. and Wild, M.: Observations: atmosphere and surface. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, in Climate Change 2013 - The Physical Science Basis, edited by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, pp. 159–254, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 2013.

Jones, P. D., Lister, D. H., Osborn, T. J., Harpham, C., Salmon, M. and Morice, C. P.: Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010, J. Geophys. Res., 117(D5), D05127, doi:10.1029/2011JD017139, 2012.

Klippel, L., St George, S., Büntgen, U., Krusic, P. J. and Esper, J.: Differing pre-industrial cooling trends between tree-rings and lower-resolution temperature proxies, Clim. Past Discuss, 1–21, doi:10.5194/cp-2019-41, 2019.

Koch, D., Jacob, D., Tegen, I., Rind, D. and Chin, M.: Tropospheric sulfur simulation and sulfate direct radiative forcing in the Goddard Institute for Space Studies general circulation model, J. Geophys. Res. Atmos., 104(D), 23799–23822, doi:10.1029/1999JD900248, 1999.

Kutzbach, J. E. and Guetter, P. J.: On the Design of Paleoenvironmental Data Networks for Estimating Large-Scale Patterns of Climate, Quaternary Research, 14(2), 169–187, doi:10.1016/0033-5894(80)90046-0, 1980.

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, J Climate, 26(4), 1085–1111, doi:10.1175/JCLI-D-11-00326.1, 2013.

Lean, J.: Evolution of the Sun's Spectral Irradiance Since the Maunder Minimum, Geophys. Res. Lett., 27(1), 2425–2428, doi:10.1029/2000GL000043, 2000.

Mann, M. E., Woodruff, J. D., Donnelly, J. P. and Zhang, Z.: Atlantic hurricanes and climate over the past 1,500 years, Nature, 460(7257), 880–885, doi:10.1038/nature08219, 2009.

Masson-Delmotte, V., Schulz, M., Abe-Ouchi, A., Beer, J., Ganopolski, A., Gonzalez-Rouco, F. J., Jansen, E., Lambeck, K., Luterbacher, J., Naish, T., Osborn, T., Otto-Bliesner, B., Quinn, T., Ramesh, R., Rojas, M., Shao, X. and Timmermann, A.: Information from paleoclimate archives, in Working Group I Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change, edited by T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex, and P. M. Midgley, pp. 383–464, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA. 2013.

Matsikaris, A., Widmann, M. and Jungclaus, J.: Influence of proxy data uncertainty on data assimilation for the past climate, Climate of the Past, 12(7), 1555–1563, doi:10.5194/cp-12-1555-2016, 2016.

Neukom, R., Barboza, L. A., Erb, M. P., Shi, F., Emile-Geay, J., Evans, M. N., Franke, J., Kaufman, D. S., Lücke, L., Rehfeld, K., Schurer, A., Zhu, F., Brönnimann, S., Hakim, G. J., Henley, B. J., Ljungqvist, F. C., McKay, N., Valler, V. and Gunten, von, L.: Consistent multidecadal variability in global temperature reconstructions and simulations over the Common Era, Nature Geosci, 536(8), 411, doi:10.1038/s41561-019-0400-0, 2019a.

Neukom, R., Steiger, N., Gómez-Navarro, J. J., Wang, J. and Werner, J. P.: No evidence for globally coherent warm and cold periods over the preindustrial Common Era, Nature, 571(7766), 550–554, doi:10.1038/s41586-019-1401-2, 2019b.

Pongratz, J., Reick, C., Raddatz, T. and Claussen, M.: A reconstruction of global agricultural areas and land cover for the last millennium, Global Biogeochem. Cycles, 22(3), GB3018, doi:10.1029/2007GB003153, 2008.

Roeckner, E.: The Atmospheric General Circulation Model ECHAM5, Hamburg. 2003.

Rutherford, S., 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, J Climate, 18(13), 2308–2329, 2005.

Slivinski, L. C., Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Giese, B. S., McColl, C., Allan, R., Yin, X., Vose, R., Titchner, H., Kennedy, J., Spencer, L. J., Ashcroft, L., Brönnimann, S., Brunet, M., Camuffo, D., Cornes, R., Cram, T. A., Crouthamel, R., Castro, F. D., Freeman, J. E., Gergis, J., Hawkins, E., Jones, P. D., Jourdain, S., Kaplan, A., Kubota, H., Le Blancq, F., Lee, T. C., Lorrey, A., Luterbacher, J., Maugeri, M., Mock, C. J., Moore, G. W. K., Przybylak, R., Pudmenzky, C., Reason, C., Slonosky, V. C., Smith, C., Tinz, B., Trewin, B., Valente, M. A., Wang, X. L., Wilkinson, C., Wood, K. and Wyszyn'ski, P.: Towards a more reliable historical reanalysis: Improvements for version 3 of the Twentieth Century Reanalysis system, Quarterly Journal of the Royal Meteorological Society, qj.3598, doi:10.1002/qj.3598, 2019.

Smerdon, J. E. and Pollack, H. N.: Reconstructing Earth's surface temperature over the past 2000 years: the science behind the headlines, Wiley Interdisciplinary Reviews: Climate Change, 7(5), 746–771, doi:10.1002/wcc.418, 2016.

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.

Stine, A. R. and Huybers, P.: Arctic tree rings as recorders of variations in light availability, Nature Communications, 5(1), 3836, doi:10.1038/ncomms4836, 2014.

Swinbank, R., Shutyaev, V. and Lahoz, W. A.: Data Assimilation for the Earth System, edited by R. Swinbank, V. Shutyaev, and W. A. Lahoz, Springer Science & Business Media, Dordrecht. 2012.

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.

Tolwinski-Ward, S. E., Evans, M. N., Hughes, M. K. and Anchukaitis, K. J.: An efficient forward model of the climate controls on interannual variation in tree-ring width, Climate Dynamics, 36(1), 2419–2439, doi:10.1007/s00382-010-0945-5, 2011.

Valler, V., Franke, J. and Brönnimann, S.: Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation, Clim. Past Discuss, 1–27, doi:10.5194/cp-2018-168, 2018.

Valler, V., Franke, J. and Brönnimann, S.: Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation, Climate of the Past, 15(4), 1427–1441, doi:10.5194/cp-15-1427-2019, 2019.

Whitaker, J. S. and Hamill, T. M.: Ensemble Data Assimilation without Perturbed Observations, Monthly Weather Review, 130, 1913–1924, doi:10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2, 2002.

Whitaker, J. S., Hamill, T. M., Wei, X., Song, Y. and Toth, Z.: Ensemble data assimilation with the NCEP Global Forecast System, Monthly Weather Review, 136(2), 463–482, doi:10.1175/2007MWR2018.1, 2008.

Wilson, R., Anchukaitis, K., Briffa, K. R., Büntgen, U., Cook, E., D'arrigo, R., Davi, N., Esper, J., Frank, D., Gunnarson, B., Hegerl, G., Helama, S., Klesse, S., Krusic, P. J., Linderholm, H. W., Myglan, V., Osborn, T. J., Rydval, M., Schneider, L., Schurer, A., Wiles, G., Zhang, P. and Zorita, E.: Last millennium northern hemisphere summer temperatures from tree rings: Part I: The long term context, Quaternary Science Reviews, 134, 1–18, doi:10.1016/j.quascirev.2015.12.005, 2016.

Yoshimori, M., Raible, C. C., Stocker, T. F. and Renold, M.: Simulated decadal oscillations of the Atlantic meridional overturning circulation in a cold climate state, Climate Dynamics, 34(1), 101–121, doi:10.1007/s00382-009-0540-9, 2010.

Zhao, S., Pederson, N., D'Orangeville, L., HilleRisLambers, J., Boose, E., Penone, C., Bauer, B., Jiang, Y. and Manzanedo, R. D.: The International Tree-Ring Data Bank (ITRDB) revisited: Data availability and global ecological representativity, J Biogeogr, 46(2), 355–368, doi:10.1111/jbi.13488, 2018.

Figure 1: Proxy locations of the three collections.

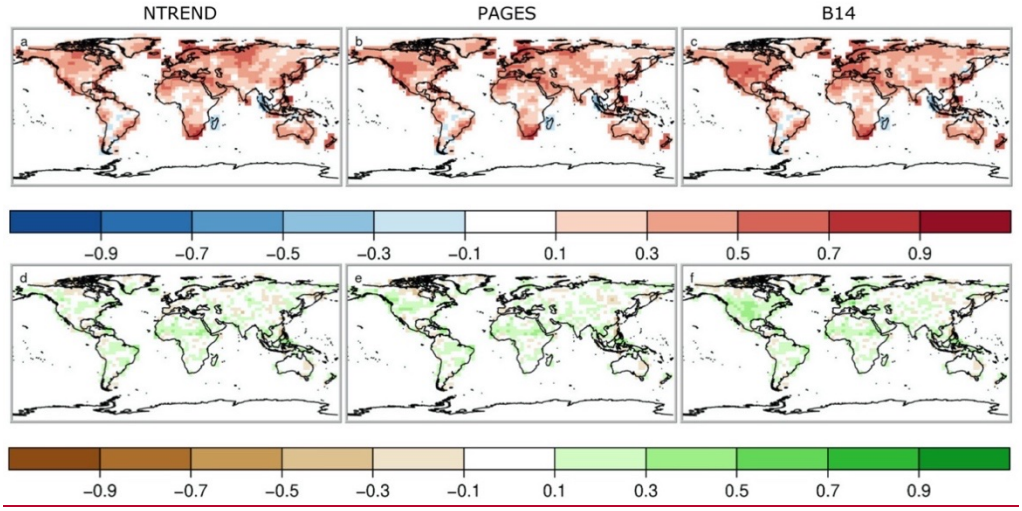

Figure 2: Pearson correlations coefficients between the analysis and gridded instrumental data in the 20th century. The top panels show temperature and the bottom panels precipitation correltation. This figure shows results from experiments 1 to 3 (Table 1), i.e. after assimilation of the three proxy data collections using the proxy system model that assumes only growth limitation by temperature.

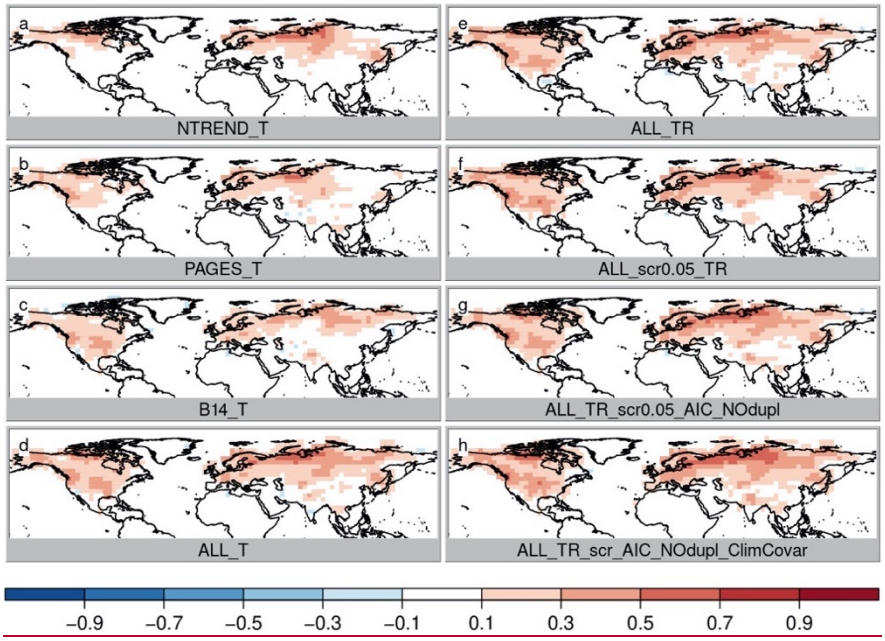

Figure 3: Temperature correlation improvement of the analysis over the original model simulations, i.e. correlation between analysis and CRU TS minus correlation between simulations and CRU TS, where red colors indicate an improvement of the analysis. All maps show the Apr. to Sep. growing season of the northern hemisphere.

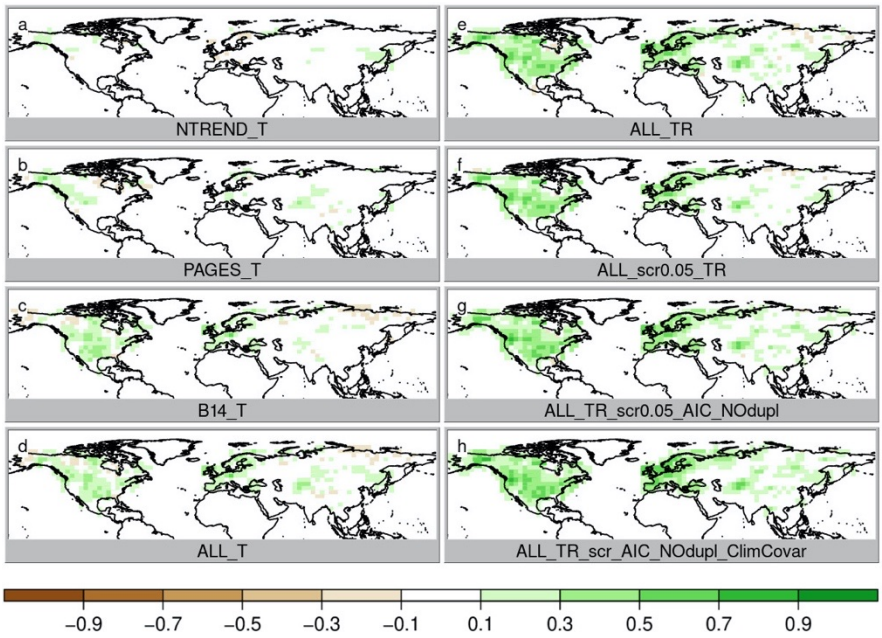

Figure 4: Same as Fig. 3 for precipitation correlation, where green colors indicate an improvement of the analysis.

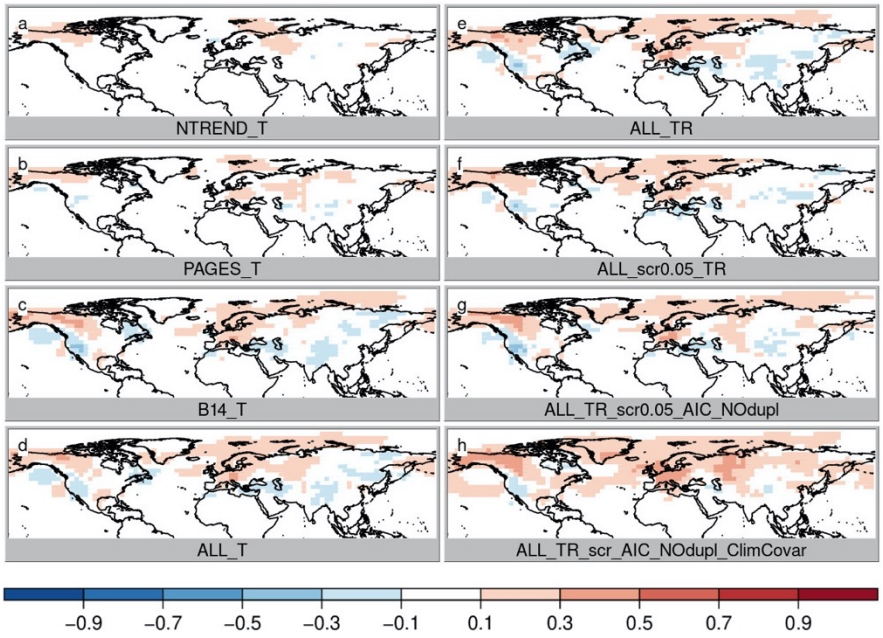

Figure 5: Same as Fig. 3 for SLP correlation, where red colors indicate an improvement of the analysis.

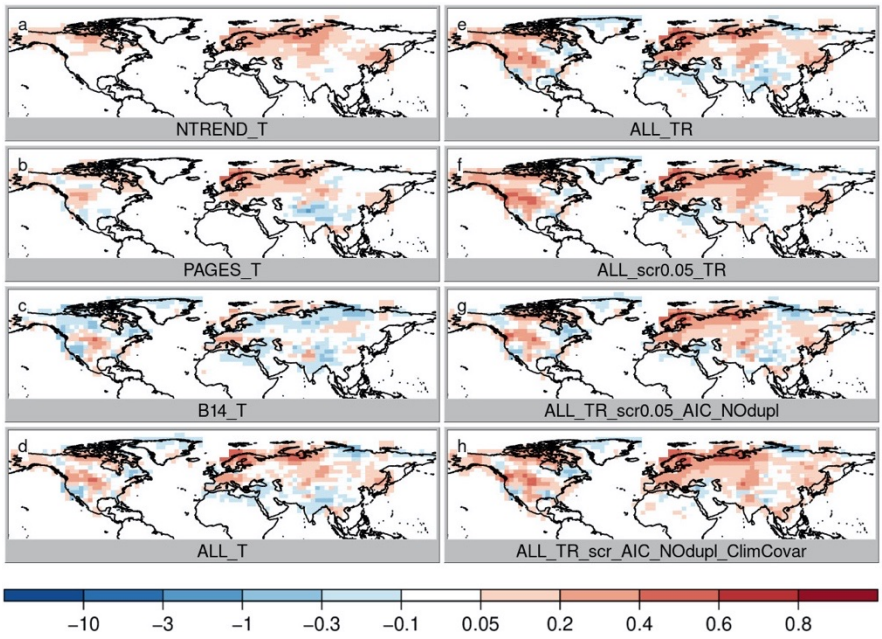

Figure 6: Temperature RMSESS skill score, where red colors indicate an improvement of the analysis.

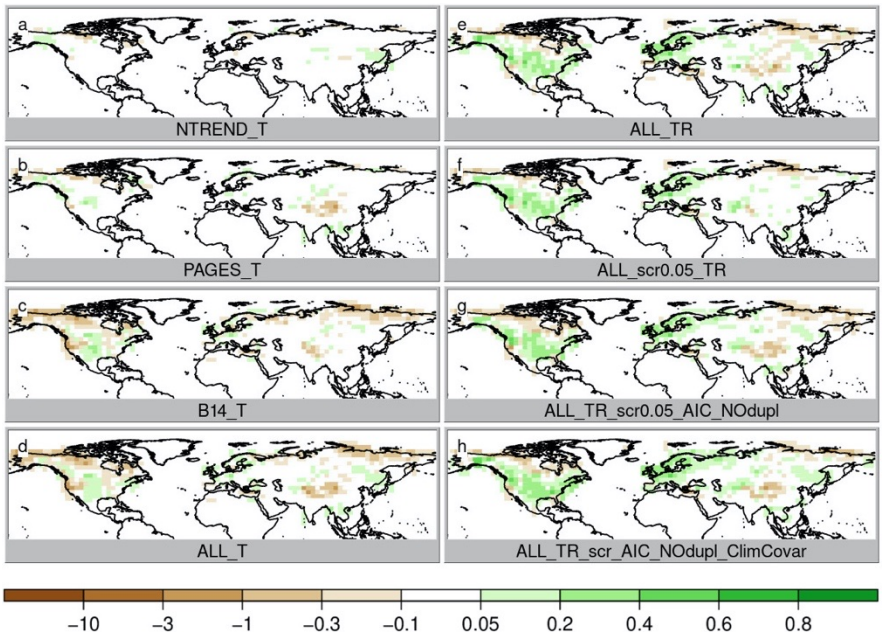

Figure 7: Precipitation RMSESS skill score, where greens colors indicate an improvement of the analysis.

Reviewer 1

***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. Anonymous Referee #1**

Received and published: 11 September 2019

Issues believed to be most important are:

1- The title of the manuscript is somewhat misleading. It implies that a more comprehensive evaluation is presented, covering a wider range of proxy archives, while the study is restricted to tree-ring data.

We changed the title to: "The importance of input data quality and quantity in climate field reconstructions – results from the assimilation of various tree-ring collections"

2- The presentation of results is problematic in several aspects:

2.1- The evaluation of the reconstructions is restricted to the instrumental-era, based on comparisons with the CRU data set for temperature and precipitation. This leads to several questions:

a. The validation is performed with the same data set used for the calibration of the forward models. What is the impact of this lack of independence on the overall conclusions of the study?

We added a discussion how this apparent lack of independence could affect our reconstruction and why it cannot have any influence on the conclusions drawn in the study (line 168ff).

b. An evaluation of reconstructions limited to the instrumental era does not provide a solid perspective of variability over longer time scales. For instance, Tardif et al. (2019) have recently shown that the selection of assimilated tree-ring width data sets leads to noticeable differences in reconstructed temperature variability at multi-decadal to centennial time scales, including the representation of notable epochs such as the LIA. Is the long-term variability in your reconstructions affected by the use of various tree- ring data sets and how? Are your results consistent with dependencies to assimilated data shown by Tardif et al.?

Our results are complementary to the research of Tardif et al. (2019). They use a method in which the prior consists of an ensemble combined from random model years of the past millennium, i.e. the prior has no multi-decadal or centennial variability. In contrast, our method uses a transient ensemble simulation as a prior, i.e. the prior in each year is in agreement with the model forcings. We keep the multi-decadal to centennial variability of the model response to the forcings in our reconstruction.

We explain this in detail in our new methods section and refer to this point in the discussion section, too.

2.2- The use of global maps in the presentation of the results is not optimal. The proxy data sets and related impact are confined to northern Hemisphere as is noted in the paper, with no signal elsewhere. Please use maps of NH only, which would show the results more clearly.

Thanks for the suggestion, we limited the maps to the northern hemisphere in the revised version.

2.3- The results are shown from the perspective of changes in verification scores in the reconstructions over corresponding values from the prior. You should show and discuss prior verification scores to cast your results in their proper context.

The new Fig. 2 shows correlation between our reconstructions and gridded instrumental data.

3- The impact of assimilated data is usually tied to the particular forward models (here proxy system models or PSMs) used. Yet, there is lack of information about the performance of the various proposed PSMs, nor a reference to prior work is given which would provide the necessary information. A characterization of the PSMs themselves would help the reader gain a more complete perspective on the results.

The PSM is now described in more detail in line 151ff.

More specific comments/questions are:

- Page 1, line 19: the use of "best possible" seems an overstatement. Perhaps you mean the best reconstruction given the parameters tested?

We rephrased this paragraph.

- Page 1, line 23: how is "insignificant" defined in the present context? Please clarify.

We clarified that the p-value of 0.05 of the regression-based proxy system model was used as a threshold.

- Page 2, line 37: The use of "probably" is not appropriate. There is a large body of literature on the impact of input data on data assimilation results (mostly focused on weather applications however). I believe a more unequivocal statement would better convey what is already known about the importance of the quantity and quality of input data to data assimilation systems.

We refer to the effect of input data quality in mostly statistical climate reconstructions, which is discussed to a much smaller degree in literature than in the case of data assimilation for weather forecasting. However, the reviewer is right that paleo-climatologists could learn from research done in meteorology. Therefore we added new references and removed the word "probably".

- Page 2, line 41: Could you better explain/justify why the study is restricted to tree-ring data?

We explain this now in lines 44f.

- Page 2, lines 48-49: Reference to specific studies which support your "would always be beneficial" statement would help improve the manuscript.

We added a reference.

- Page 2, lines 52-53: The statement including "which often results in a small sample, uncertain residuals and possible model overfitting" lacks support. Can you include references or show results that highlights these problems?

Many tree-ring measurements were already done in the 1970th to 1990th. Hence, the number of proxy data drops rapidly from the 1970th to the present (e.g. J. Emile-Geay et al (2017). Instrumental station networks outside Europe and the United States of America, however, were very sparse before 1900 and only reach roughly present-day spatial coverage around 1950. We added a reference for clarification.

- Page 2, line 75: Is your statement about "no dating uncertainties" accurate? Perhaps "small dating uncertainties" would be more appropriate?

We changed it to "hardly any dating uncertainties".

- Page 3, line 85: About the statement "experts from various regional groups were differently strict in their screening procedure", has this been characterized in a more formal way? Please provide support for this statement.

The PAGES2k data set is a global proxy data collection gathered by multiple regional groups. While we use the entire collection, Emile-Geay et al. (2017) provide additionally multiple screening levels of the data (their Tab. 2). The amount of screened records varies by region if these stricter rules are applied (see supplementary Fig. S2, S3 and S4 in Emile-Geay et al., 2017).

The difference in the amount of data in the various regions is also caused by different priorities. The European group, for example, only included the longest and highest quality records from the wealth of datasets that exist in this region. In contrast, most other regional groups included all available records that fulfill the global minimum selection criteria. Therefore, the number of records in the PAGES2k database from Asia and North America is much higher than from Europe.

We have amended the text to read: "This compilation represents a compromise of good quantity, large spatial coverage and good quality paleodata, based on global selection criteria. However, experts from various regional groups were differently strict in their screening procedure, which lead to varying data density in the different regions".

- Page 3, line 86: You mention that "N-TREND is a collection of 54 tree-ring reconstructions". Do you assimilate the reconstruction data or the tree-ring data underlying the reconstructions? Please clarify. If you use the reconstructions, please justify.

We clarified in the text that we used the N-TREND tree-ring chronologies. In neither case we work with raw tree-ring measurement but use processed chronologies, in which

multiple samples from one site have been combined, growth trends have been removed etc.

- Page 3, line 93: Statement with "...simulate tree-ring observations using modeled temperature or precipitation": I believe you also use PSMs that include both temperature and precipitation as input. A more accurate statement would therefore include "temperature and/or precipitation".

Yes, this has been corrected accordingly.

- Page 3, line 95: You use a single seasonal response for all records, and for temperature and precipitation. Please justify.

No, we allow for all six months of a hemispheric growing season to potentially contain regression coefficients to be different from one. However, this also allows for having growths influenced by a few or a single month only. We explain this now in detail in the new methods section.

- Page 3, lines 97-103: I do not easily understand the information provided in this paragraph. I would suggest revising the description of the PSMs, perhaps using equations or illustrations, to provide a description the reader will more easily understand.

This has been clarified in the new methods section.

- Page 3, line 109: The procedure described here amounts to some screening of the data that is not evaluated nor discussed further here. Perhaps it should be.

This in now explained in the methods section.

- Page 3, line 112: Please specify what is the source of the 30 ensemble members. This is not clearly identified here.

We added a short description of the simulations and references to the more detailed description of the simulations.

- Page 3, line 115: What is the localization applied when precipitation is involved? Please specify.

We added the equation used for localization including parameters for temperature and precipitation.

- Page 4, line 120: I am failing to understand the justification for using anomalies about 71-yr mean values, or the prior model states? proxies? Please describe and justify in more detail so the reader can understand.

The general problem in many tree-ring chronologies is the fact that they were not specifically created with the aim to retain realistic variability at all time scales. For instance, if a study aimed at interannual variations, multidecadal to centennial variability may have been filtered out. Or already the sampling strategy may not have been appropriate to retain such low frequency variability. Therefore, we only assume

that tree-ring chronologies contain a reliable interannual to decadal signal. Accordingly, we assimilate anomalies around a 71-year mean. We described this procedure in more detail in the methods sections of the new manuscript version.

- Page 4, line 124: Can you support the statement that the method is "expected to provide consistent skill at all time scales"?

Most tree-ring chronologies can be expected to represent interannual variability similarly well. However, centennial scale variability is not similarly well retained in all records, (see last comment or Franke et al. 2011). Therefore, we only assimilate anomalies around the 71-year mean and update the same 71-year anomaly field in the model. The model climatology is added again after the assimilation is finished. This way, the centennial-scale variability in our paleo-reanalysis is just a function of the model response to the external forcings and the model remains physically consistent but biased. We prefer this procedure not only because of proxy data characteristics but also because it does not introduce artificial biases when new observations become available (see Franke et al. 2017). This is now explained in the methods section, too.

- Page 5, top row of table, rightmost frame: Can you provide some evidence to support your statement that records "Probably included some moisture or partly moisture sensitive" ones?

Many tree-ring records have a mixed climate signal, i.e. are not pure temperature or precipitation recorders. Many of the records in the PAGES2k database may have a significant precipitation signal, which may be even stronger that the temperature signal. Inclusion criteria were mainly that a record needs to be temperature sensitive, independent from potential relations to other variables, even if those are stronger. If you look at the proxy distribution maps in Emile-Geay et al. (2017), you can find many sites in warm and dry locations such as the south-western United States. These sites have been used in hydroclimatic reconstructions (Steiger et al. 2018, Cook et al, 2007). In Emile-Geay et al. (2017) a sign correction is done, i.e. if temperature and precipitation are negatively correlated, tree-ring chronologies can remain as temperature sensitive in the data set, no matter if they show an anomalously wide or a narrow ring in an anomalously warm growing season.

This information has been included in Table 1.

- Page 6, lines 162-163, statement that "B14 provides temperature information in places where temperature is correlated with precipitation": while most likely true, this statement seems incomplete. B14 also contains temperature sensitive records. One could argue that temperature improvements are mostly related to the assimilation of such records, more so than through the process you describe here. Can you support and quantify your statement?

We checked the sign of the regression coefficients and found mostly negative relationships in the United States of America and the Mediterranean. Here, narrow rings indicate dry and warm growing seasons. This information has been included in the results.

- Page 6, line 183: regions (plural).

This has been corrected.

- Page 7, line 195, “lost”: I believe you mean “lowest”.

This has been corrected.

- Page 7, line 198: Can you provide a more complete reasoning as to why you believe overfitting is the (main?) reason for the behavior described in this paragraph?

As mentioned in one of the earlier comments, our regression model always contains coefficients for the six months of the growing season although growth may in many cases be limited to the shorter period. In such cases, regression coefficients will be close to zero but will not be exactly zero because we work with a relatively small sample size. Hence, in many cases we use a regression model with too many independent variables, which do not contain information. In multiple regression, this is known to cause model overfitting.

This is now explained in more detail in lines 256ff.

- Page 7, lines 225-226, statement about problems on longer time scales: The experiments discussed in the manuscript are not evaluated on that particularly sensitive issue, an important shortcoming of the study in my opinion. The fact that results from your experiments can not provide a clear contribution toward characterizing or resolving this issue should be acknowledged.

As explained above, this study is complementary to Tardif et al. 2019 in this respect and we cannot draw any conclusion in this regard with our method because our low-frequency variability is the response of the model to the forcings and not proxy dependent.

This should be clear after reading the new methods section.

- Page 9, lines 278-279: A more complete statement should include a reference to the work of Dee et al. (2016) to indicate that application of VS-lite has additional limitations related to model biases. (Dee, S., Steiger, N. J., Emile-Geay, J., and Hakim, Gregory J., 2016: The utility of proxy system modeling in estimating climate states over the Common Era, J. Adv. Model. Earth Sys., 8, 1164–1179, doi: 10.1002/2016MS000677)

This has been added in line 345.

- Page 9, line 288: data sets (plural)

This has been corrected.

- Including a figure showing the location of the proxy records from the various data sets would strengthen the presentation.

We added a new Fig. 1 with the proxy locations.

Reviewer 2

Review of '**The importance of input data quality and quantity in climate field reconstructions – results from a Kalman filter based paleodata assimilation method.'** by J. Franke V. Valler, S. Brönnimann, R. Neukom, and F. Jaume Santero

Specific comments

1) The period which is analysed should be stated in the abstract. At the moment the first time this information is given is in line 130. The abstract should also briefly mention the type of assimilation method.

The method and analyzed time period are now mentioned in the abstract.

2) Line 21, 'improved' relative to what? This does become clear later in the text, but the abstract needs to make sense on its own.

Has been changed to "…but fail to provide information for other regions and other variables".

3) The paper links the results in several places to terms in the Kalman filter, namely to the proxy system model, the observation error covariance matrix and the background error covariance matrix. These are important comments, but readers who are not experts in data assimilation will probably not understand them, because the Kalman filter equation that is used is not given in the paper. I do appreciate that these details are given in previous publications, but with respect to its core elements a paper should be self-contained. Please add the equation to the method section and discuss there how the terms are calculated and how information is spread by the various terms from the proxy data to the different reconstructed meteorological variables. When presenting the results please refer back to this discussion where appropriate.

We added a detailed methods section, which allows reader to understand this paper without reading previous publications about the method.

4) Line 92, 'we need a forward model that simulates them in the model state vector'. This is not well phrased.

This has been rephrased.

5) There should be clear comments to what extent the findings can be transferred to other data assimilation methods used in paleoclimatology. It is likely that methods with a similar structure, i.e. using PSMs and variations of Kalman filters, will have similar sensitivities to the selection of input data, while others, for instance particle filters, may not.

We think that the quality/selection of input data has similar consequences not only in other data assimilation methods but in statistical reconstruction, too. However, there will be method dependent differences.

We added a paragraph at the end of the discussion (line 354ff).

6) Line 57-59. The statement on the similarity between the method used in this paper and the method used in the last millennium project is misleading. The method used in the paper uses a 'transient offline method', in which the background state is time-dependent due to the signal of the external forcing. This aspect is actually highlighted by the authors in lines 120-124 and lines 231-232. In contrast, the last millennium project uses a 'stationary offline method' in which the background state does not depend on time. This crucial difference should be mentioned.

We already added this information to the introduction to avoid confusion (line 63ff).

7) Lines 121-122. The comments on low-frequency variability should include a discussion of the setup for the simulations that provide the background state. Why is sea surface temperature mentioned as a forcing? Are the simulations done with atmosphere-only GCMs? If so, which sea surface temperatures are used? These comments should also discuss the role of random, internal, low-frequency variability.

As mentioned above, we added a much more detailed methods section to the revised manuscript.

It should also be clarified that the validation measures are calculated from annual values and are thus dominated by inter-annual variability. This fact and the short evaluation period imply that an evaluation of low-frequency variability is not possible in this study.

Yes, as explained in the answers to the first reviewer's comments, we explain in the revised version that our methodology is not suitable to draw conclusions on the proxy data sets' influence on centennial-scale climate variability. This study is complementary to the results found by Tardif et al. (2019) on low frequency variability effects caused by the input data selection.

8) Line 125, it is not clear from which data the running mean is calculated and what 'model' refers to.

As mentioned before, everything is now explained in detail in the new method section.

9) Line 131, 'just at correlation itself' sounds strange.

This has been corrected to "Instead of analyzing absolute correlation coefficients, we analyze correlation improvements …"

10) Lines 137/139/264, 'punishes' should not be used in scientific writing.

This has been replaced by "penalize".

11) Line 138-140, please include a more detailed justification of why the evaluation is based on ensemble means rather than on individual ensemble members followed by averaging of the skill scores. This should include explicit statements on the effect of the reduced variability in ensemble means on the RE; the current statement is unclear.

Thanks for this suggestion. You are right that we can expect reduced variability in the ensemble mean compared to the ensemble members and compared to the validation

data, too. Nevertheless, most user will be interested in the ensemble mean and its skill. One would expect that the ensemble mean of the transient simulations before assimilation has little variability and that the assimilation would lead to an increase in variability bringing it closer to the observations. This is also what we observe with the generally positive RE skill scores. This way, a perfect skill score of one could only be reached, if the ensemble would have no variability left and perfectly matches the observations. We do not expect that the pattern or sign of the skill score would change in case of averaging the skill of the single members. However, the skill scores of the single members tend to be larger than the skill score of the ensemble mean because variance is not systematically underestimated.

We discuss this now in lines 199ff.

12) Line 227, 'TRW limitations remain the same' is not clear.

This has been changed to: "TRW limitations are time-independent".

13) Line 254, it should be 'principal' not 'principle'

This has been corrected.

14) The use of hyphens is inconsistent and often wrong. Adjectives that are constructed from two words should usually be hyphenated. Examples are 'temperature-sensitive', 'regression-based', 'time-dependent' (which is better than 'time-variant' used in line 113), 'low-frequency' (if used as an adjective), 'inter-annual', 'multi-decadal, 'multi-variate' etc. In some case it is also correct to combine the two words, e.g. 'multivariate'

We checked the document and hopefully corrected it everywhere.

15) Line 195, replace 'lost' with 'lowest'

This has been corrected.

16) Lines 212-213, This is not a proper sentence.

This sentence has been rephrased.

17) The paper Matsikaris, A., Widmann, M. and Jungclaus, J., 2016. Influence of proxy data uncertainty on data assimilation for the past climate. *Climate of the Past*, *12*(7), pp.1555-1563. addresses similar questions and should be included in the introduction and/or the discussion.

We now refer to this paper in line 53.

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. 2011). This inconsistency within the proxy data is actually the reason why we have another approach than most previous reconstructions. 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 146 of the revised manuscript, 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 such as the PAGES data base. This 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 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.

We cite the study of Kutzbach and Guetter (1980) now and also explain that we cannot expect skill from the NTREND data set far outside the region that is covers.

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 rephrased the abstract 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. This message should become clearer in the revised text.

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 added a new Fig. 2, which shows the absolute skill of the reconstruction with respect to gridded instrumental data.

As we already explained to the first reviewer: "We added a discussion how this apparent lack of independence could affect our reconstruction and why it cannot have any influence on the conclusions drawn in the study (line 168ff).

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 an error of a forecast for instance the root mean square error (RMSE) to the error of a reference forecast. Lorenz 1956 used 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. 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 called this skill score RMSE SS in the revised version.