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

General comments:

The manuscript describes results from a series of data assimilation experiments aimed at identifying the "best" tree-ring input data set, i.e. the one leading to the largest improvements in paleoclimate reconstructions of temperature, sea-level pressure and precipitation. Three data sets are primarily tested, differing in the level of screening applied to tree-ring proxy records with respect to their climate sensitivity. The topic addressed in the manuscript is an important one and should be part of the published literature on climate field reconstructions. Several issues are found however, which should be addressed before the manuscript is published.

We appreciate that the reviewer considers this topic to be of general importance.

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 believe that these results are relevant for various archive type. However, the reviewer is right that we only tested tree-ring proxies. Hence, we will choose a more precise title: "The importance of input data quality and quantity in climate field reconstructions – results from the assimilation of various tree-ring data 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?

The reviewer is right that there is a lack of independence which comes from 1) the regression model and 2) the residuals. Concerning 1), regression coefficients are estimated from gridded instrumental data sets to translate grid cell temperature (and moisture) anomalies to local tree-ring measurements. The optimization is done on tree rings, not on the climate data, and it is done on many local scales and not the large scale. In that sense the effects of the dependence are rather indirect. In contrast the

statistical reconstruction methods, which directly estimate a climate variable such as temperature through the regression parameter estimate, our assimilation method is less far affected by the calibration procedure. Nevertheless, we agree with the reviewer that using the same data for validation probably leads to a slight overestimation in reconstruction skill and this is the reason, why we made additional "leave-one-out"-experiments in the publication of the original reconstruction (Franke et al. 2017). Concerning 2), we use these regression residuals as an estimate of error covariance, i.e. the larger the residuals, the smaller the weight of the proxy observation in the assimilation process. Again, the error estimate concerns tree ring width, not climate parameters.

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

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

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 and just assimilate 71-year running anomalies based on the fact that many to tree-ring chronologies do not contain the correct centennial variability (Franke et al 2011). Hence, we cannot draw any conclusions on the impact of the input data set choice in low-frequency variability. We touch upon this point in the second and third paragraph of the discussion but will make it clearer in the revised version. Already in the introduction, we will better explain the differences between the methods used by Tardif et al. and us.

Based on the comments of both reviewers, we will extend the entire methods section. This will allow the reader to better understand this study without reading the previous publications, which explained more details about the used method.

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 will limit 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.

Our prior is a forced model simulation, which has by definition already some skill, e.g. positive correlation due to warming trend in the 20th century. Therefore, we present the improvements with regard to forced model simulations. This makes more sense as well with regard to comment 2.1.a. Nevertheless, we understand the point that it is of interest to the reader how well your prior already agrees with instrumental observations and how well the reconstruction in performs in comparison to instrumental data. We will add a figure with absolute skill of the prior and the reconstruction, highlighting that the method is able to generate a reasonable reconstruction.

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 described in section 2, line 90. The annual tree-ring data are translated into temperature (and moisture) based on a multiple regression model using monthly means of a six months growing season (April to September in the northern hemisphere). Thus, there are 6 regression coefficients in the experiments with temperature only and 12 regression coefficients in the experiments where growth can be limited by temperature and precipitation. Because of a lack of more sophisticated PSMs for tree-ring density and possible model biases, which would affect non-linear growth functions, we decided to use a regression model instead of a more sophisticated tree-growth model such as VSL (Tolwinski-Ward et al. 2013).

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?

Will be stated more carefully.

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

We will clarify 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. We will add appropriate references to make this literature more known and remove the word "probably".

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

We will explain that tree-ring proxies are the most widely used proxy types to reconstruct climate of the past centuries because they are the most abundant. Additionally, they are best suited for data-assimilation based reconstruction because they have hardly any dating uncertainties.

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

We will add 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 (Jones et al., 2012, Fig. 1). Hence, many locations of temperature sensitive tree-ring proxies, which are located in remote mountainous regions or high latitudes neither have station measurements nearby nor long overlapping periods.

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

We will change 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 will clarify 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, will be 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 will clarify this in the revised version.

- 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.

We will rewrite this paragraph to clarify the PSMs (see last point).

- 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.

The same tree-ring site may have been included in multiple collections. To prevent the same observations from being assimilated multiple times, we assume that different study sites should be more than 0.1° apart from each other. Hence, this procedure is only a removal of duplicate records in the experiments, where data sets are combined. These rare cases hardly affect the reconstruction skill and hence there is no need for further discussion. We will clarify this in the revised text.

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

We add the references to the ECHAM simulation ensemble (CCC400) published in Bhend et al. (2012) and Franke et al. (2017).

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

We will add 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 will describe this procedure in more detail in the revised version of the manuscript.

- 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).

- 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.

- 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 will check the sign of the regression coefficients of the assimilated observations in this region to evaluate how many temperature-sensitive observations are assimilated there. These should show wide rings in warm and dry years.

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

Will be corrected.

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

Exactly

- 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.

- 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 will be clear, when we have already explained our method in more detail in the 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)

Will be added.

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

Will be corrected.

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

Will be added.

References, which had not been included in the discussion paper:

P. D. Jones, D. H. Lister, T. J. Osborn, C. Harpham, M. Salmon, and C. P. Morice, "Hemispheric and large-scale land-surface air temperature variations: An extensive revision and an update to 2010," J. Geophys. Res., vol. 117, no. 5, p. D05127, Mar. 2012.

E. R. Cook, R. Seager, M. A. Cane, and D. W. Stahle, "North American drought: Reconstructions, causes, and consequences," Earth Science Reviews, vol. 81, no. 1, pp. 93–134, Mar. 2007.