Reviewer #1

This paper tests several different methodological choices that are typically made (or could be made) in paleoclimate reconstructions using DA. I think this presents a good and valuable presentation and discussion of these choices. The findings and suggestions for future reconstructions are very helpful to the community performing these types of reconstructions.

We thank the Reviewer for the positive comments on our work and for the suggestion how to improve our figures.

Section 3.1: Is there any specific justification for the choice of the ratio of L_z and L_m being 2:1? Could this, or some other ratio, be justified by looking at the correlation length scale in observational data?

> The idea behind a longer length scale in zonal direction than in meridional direction is based on the zonal flow in the atmosphere. On multi-annual to multi-decadal time scales multiple processes act in meridional direction, e.g. a widening/shrinking of the Hadley cell, shifts of the ITCZ or changes in atmospheric modes like the AMO or the NAO. These can shift the zonal circulation northward or southward but the zonal coherence will be less affected.
>
> In principle the correlation length scale found in observational data or reanalyses could be used, but the anisotropy depends strongly on season, variable and location (e.g., it is much larger in the tropics than in the extratropics). Between experiments “no localization” (of Pclim) and localization of the standard setup, a 2:1 experiment seems like a good intermediate.
>
> We explained the hypothesis in more details in the revised manuscript.

Section 3.1: Is there any justification for the specific localization values that you chose for each variable that was reconstructed? Are these values data-driven or just educated guesses? Were any experiments done to test on optimal localization value? I would assume that if these values were used based on weather DA experiments, they might not apply on the longer paleo time scales where one would generally expect the correlation length scales to be larger.

> The localization length scale parameters were defined based on the spatial correlation of the variables in the monthly ECHAM model simulation fields. In Section 2.4 we refer to the paper by Franke et al. 2017 how the localization was done in the original setup. We used the same localization length scale parameters for localizing the sample covariance in most of our experiments to evaluate improvements in comparison with this initial setup. For this study, we calculated the latitudinal dependency of correlation of the state variables from a bigger ensemble of the model than in Franke et al (2017). The result suggested that the longer length scale parameters can be applied in the tropics and the predefined length scale parameter of precipitation is probably too strict. Based on the rather strict correlation length scale parameters in the previous study and the assumption that the covariances can be better estimated from a bigger ensemble, we used doubled length scale parameters in some of the experiments for localizing the climatological covariances. In this case, the L for temperature is 3000 km, which means that the correlation is decreased close to zero approximately 6000 km away from the observation. We did not carry out further experiment to find the optimal localization value because even double the localization distance hardly changed the reconstruction skill, not even at locations that newly fall inside the radius of influence of at least a single observation. Hence, our setup does not appear to be very sensitive on the localization distance as long as it remains in a reasonable range. On the one hand, we do not

further restrict the localization because that would limit updates to a small regions around observations. On the other hand, our experiments without localization showed negative reconstruction skill in locations far away from observations, even with the error covariance matrix is calculated from climatology. We provide further explanation about localization in Section 2.4 and Section 3.2 in the revised manuscript.

Section 4.2.1: When you are comparing the distributions, you say that for example,the most skillful reconstruction is obtained from the 100c_PcL experiment. What is the basis for saying it's the best? What aspect of the distribution are you comparing? The median or some other specific value(s)?

Yes, for comparison we used only the median.

Many of the distributions shown in the figures look very similar so it was hard for me to feel confident about the statement that one particular set of reconstruction choices was better than another. Are the distributions statistically distinct?

We agree with the Reviewer that the distributions of the skill of the experiments over the extratropical Northern hemisphere look similar. We have not checked whether the distributions are statistically distinct. In the revised paper we provide an evaluation of the skill scores of the experiments compared to the original analysis skills, using a permutation test.

Instead of comparing the distributions, would it be possible to show the differences compared to the "original"reconstruction (i.e., you'd compute the difference in the skill score for each location and then summarize this distribution of differences in the plots)? I'm wondering if this,or something similar, might make the differences more clear. Because currently when I look at the distributions, many of them look very similar and perhaps even statistically indistinguishable.

Thank you for your suggestion. We made the plots as it was suggested and replaced Fig. 4-6 and Fig. 8 showing the differences between the experiments and the original analysis.

Fig 8 & 10: It would be very helpful to give a little more explanatory information/labeling on each panel, such as was done in Fig 3.

We added more labels to the figures (in the revised manuscript Fig. 7 and Fig. 9).

Reviewer #2

This manuscript identifies the best choices from a number of different spatial and temporal localization approaches and from different inflation techniques for the background error covariance matrix in Ensemble Kalman Filters used in paleoclimatic applications. The optimization of these technical details in data assimilation is important for the growing paleoclimate data assimilation community. The results are systematically derived and the manuscript is in general well written. I support publication after the points listed below have been clarified or corrected.

We thank the Reviewer for the careful revision of the manuscript. We will follow his/her recommendations and add further details at the indicated points to provide better understanding how ensemble Kalman filter methods work, how we set up the blending experiment and how covariance inflation techniques can help to better estimate the background-error covariance matrix. We are also thankful for improving the English of the manuscript.

Specific comments

Page 1, Line 2, replace ‘of the assimilation system’ with ‘of some assimilation systems’

> We replaced it.

Page 1, Lines 17/18, 24/25, ‘boundary conditions’ are specifications of state variables at the boundaries of a model domain, and thus not the same as ‘forcings’, which are external influences on the system. The two should be distinguished throughout the text. It seems that here the statement are about forcings. If so, reformulate avoiding the use of ‘boundary condition’.

> Was replaced as:
>
> Climate models constrained with realistic, time-dependent external forcings provide fields that are consistent with these forcings and the model physics.
>
> If information from observations is optimally blended with climate model simulations, the result is the best estimate of the climatic state, given the observations, given the external forcings, and given the known climate physics.

Page 2, line 6, ‘linear models’ of what? I think it should by ‘linear dynamical systems’. A short comment on why KFs are used with non-linear systems, including GCMs, would be good. ‘Gaussian distributions’ of what? The state variables?

> Yes, the KF is optimal for linear systems (linear dynamical model and linear observation operator) with Gaussian error distributions (model, background, and observation). In atmospheric data assimilation the most commonly used KFs are the ensemble-based Kalman filters that can handle some non-linearity in the dynamics and in the observation operator. Moreover, in our approach we use an offline implementation of the ensemble square root filter, i.e. we only adjust the precomputed model simulations with the observations and never deal with the question of model dynamics. Nevertheless, we removed this sentence because it was not necessary here and just caused confusion.

Page 2, line 13-16, I suggest using ‘stationary offline’ and ‘transient offline’ for the two approaches.

Thank you for the suggestion. We used the terms as stationary (forcing-independent) offline and transient (forcing-dependent) offline to better distinguish between the two methods.

Page 2, lines 17/18, 'The true climate state is not known, therefore it has to be estimated'. Does 'it' refer to 'the true climate state', as the sentence suggests or to 'the uncertainty of the background state', which would link better to the first sentence in this paragraph? This sentence should be clarified, or it could simply be deleted (which I think is the better option).

We deleted the sentence.

Page 2, line 20, What is a non-simplified KF? A KF with a 'true' background error covariance matrix? If so, how can this exist? The background error covariance has always to be estimated somehow. Please clarify the statement.

We repalced **'**Ensemble-based KFs are simplification of the KF' to 'Ensemble-based KFs are approximations of the KF'.

In the Kalman filter, the background-error covariance matrix is given as

$$P^b = \overline{(x^b - x^t)(x^b - x^t)^T}$$

the superscripts b and t indicate the background state and the true state, while the overline marks the expectation value.

In practice, as was said, the background-error covariance matrix has to be estimated. In ensemble-based Kalman filter techniques, the ensemble covariance will provide the background-error covariance matrix and the true state is estimated by the ensemble mean.

$$P_e^b = \overline{(x^b - \overline{x^b})(x^b - \overline{x^b})^T}$$

Page 2, line 23, It seems that the sampling error for the background error is not only a random error, but leads to a systematic underestimation of the background error, otherwise inflation would not be a suitable approach. Please explain better.

Ensemble-based filters with flow-dependent covariances can diverge due to sampling error. Our small 30 member ensemble may result in an error underestimation, because we do not assess the uncertainties in SST boundary conditions and external forcings. In an online ensemble-based KF approach, after the update step all ensemble members are propagated forward according to the model dynamics. However, if the uncertainty of the analysis is underestimated in the update step, the background error may be underestimated in the next time step and the method will trust the model more and more, while giving less and less weight to the observations in the following time steps. This should not be a problem in offline approaches. However, we conduct this experiment for completeness because as we say in the manuscript, there are two commonly used methods to reduce the negative effect of sampling error: inflation (e.g., Anderson and Anderson, 1999), and localization (e.g., Hamill et al., 2001) of the background-error covariance matrix."

The multiplicative inflation part was rewritten as:

A simple inflation technique is the multiplicative inflation (Anderson and Anderson, 1999), which compensates for potential underestimation of the analysis error. Multiplicative

inflation helps to maintain a more realistic distribution of the ensemble members by increasing the deviation of the members from the ensemble mean at each DA cycle (Anderson and Anderson, 1999). However, the underestimation of the analysis error is of minor importance in offline approaches, because the ensemble members are not propagated forward in time.

Page 2, lines 25/26. The statement on distribution of ensemble members refers to online approaches, but the approach used by the authors is an offline approach. This is confusing. Please briefly explain how the ensembles are generated in an online KF, and that in offline approaches the ensemble is given, but that the background error covariance still needs to be inflated.

To avoid any confusion, we changed the sentence to: "Two methods are commonly used in online ensemble-based KF approaches to reduce the negative effect of sampling error:" On page 2, line 8-11 the main steps of an online KF is already described. These steps remain the same in online ensemble-based Kfs. In the update step the ensemble members are updated with the observation when they become available. In the forecast step these updated ensemble members are propagated forward by the model to the next time step, then the mean and the covariance are estimated again. Finally, with regard to the need to inflate the background error covariance, please see our answer to the last comment on page 2, line 23.

Page 3, line 14, replace 'other method' with 'additive method'

We replaced it.

Page 3, line 25, replace ''form' with 'from'

We replaced it.

Page 3, line 28, Don't use 'forced by boundary conditions', as forcings and boundary conditions are different (see comment above). If I understand correctly for all ensemble members the same greenhouse gas, solar and volcanic forcings have been used, as well as the same SST boundary conditions. Please clarify.

Yes, the same forcings and boundary conditions were used in the 30 members.

Was replaced as:

The 30 ensemble members were forced and driven with the same external forcings and with the same boundary conditions.

Page 3, Lines 29-31, The SST reconstruction can be expected to strongly influence the results of this data assimilation approach with an atmosphere-only GCM. There should be some comments on how the SST reconstructions have been made, what is known about their uncertainties, and why this approach is taken rather than data assimilation with a coupled atmosphere-ocean GCM.

We use the SSTs from Mann et al. (2009), augmented as described in Bhend et al. (2012). The main reason why we use the SST reconstruction by Mann et al. 2009 is that it was the only available reconstruction at the time when we ran the simulation ensemble, like we already explained in the paper: "For sea-surface temperatures (SSTs), which have a

particularly large effect on the simulations, the reconstruction by Mann et al. (2009) was used. This is the only global gridded SST reconstruction that dates back till 1600."

We already describe limitations of the reconstruction and how we tried to improve them in the discussion paper, too: "The SST reconstruction by design captures interdecadal variations (Mann et al., 2009), hence intra-annual variability dependent on a El Niño/Southern Oscillation reconstructions (Cook et al.,2008) was added to the SST fields." We will add a reference to the paper of Bhend et al. 2012, which explains the simulations in a little more detail.

There are multiple reasons why we do not use a coupled atmosphere-ocean model:

1. We understand a paleo-reanalysis to be a product that best describes the atmospheric states of the past globally. 400 years of just externally forced coupled atmosphere-ocean simulations would not be in close agreement with the true state of the ocean in the past and e.g. not produce El Niño events in the years when they actually occurred. Therefore such simulations would not capture related teleconnections patterns in the corresponding years, either. We can not gain the same information that currently comes from the SST boundary conditions with our data assimilation setup because we only assimilate absolutely dated tree-ring proxies on land and mostly at mid to high latitude. It would be an option to assimilate coral data in the future but there are few records that back to 1600 and their assimilation has not been tested yet. That is why coral data assimilation in not an option for the presented sensitivity experiments. The focus of this study is to show, how the paleo data assimilation setup published by Franke et al. (2017) can be further improved without introducing further changes that would blur the effect of the improved covariance estimation.

2. If we would like the ocean to be in closer agreement with the observations in a coupled model, we would need to do an online assimilation scheme that could take the longer memory of the ocean into account whereas our atmospheric states have no memory that extends beyond a month or year when the next proxy observation becomes available. Hence, running a coupled ocean-atmosphere model would not allow us to work with our offline assimilation system and do all the sensitivity experiments, which we present in this study.

3. Finally, an ensemble of coupled simulations plus the additional spin-up time for the ocean to reach equilibrium would have simply not been computationally affordable.

Page 3, line 30, I think 'till' should not be used in formal writing and should be replaced with 'until' (also later in the text).

We replaced it, also in page 4, line 10-12

Page 4, line 1, replace 'boundary conditions' with 'forcing', and make a separate statement on land-surface boundary conditions, including which variables are prescribed.

Was replaced as:

Further forcings include solar irradiance, volcanic activity, and greenhouse gas concentrations (for more details see Bhend et al., 2012; Franke et al., 2017). The land-use reconstruction by Pongratz et al. (2008) was used to derive the land-surface parameters.

We would suggest only mentioning that the land-use reconstruction by Pongratz et al. (2008) was used in the paper because we would not like to give very different levels of detail for each forcing or boundary condition. The reconstruction by Pongratz is the standard land-use

reconstruction used in basically all ECHAM model simulations since the reconstruction was published. For the Reviewer interest, the 14 land-cover classes defined by Pongratz et al (2008) were mapped to land surface parameters closely following the procedure of Hagemann et al. (1999).

Page 4, line 16, 'CCC400' has not been introduced

In the title of section 2.1 we try to indicate that the model simulation is called CCC400. CCC400 stands for "Chemical climate change over the past 400 years". The name is misleading since the model was run without chemistry due to a lack of computing resources.

To make it more clear we will replace page 3, line 25-26 as:
The model simulation, termed as Chemical Climate Change over the Past 400 years (CCC400), has 30 ensemble members, that are used as background to reconstruct monthly climate states between 1600 and 2005.

Page 4, line 23, If I understand correctly the deviations of the ensemble members from the ensemble mean are updated in the online EnSRF according to equation 2, but not in the offline EKF, which uses an existing ensemble. Please clarify.

In both online and offline EnSRF methods, the mean and the deviation from the mean are updated, according to Eq. (1) and Eq. (2). However, in an online application each ensemble member is propagated forward to the next time steps. In case of an offline approach the process stops after all available observations at the current time steps have been assimilated.

Page 4, lines 28-30, the statements on the 6 month periods are partly redundant.

Was replaced as:

In the EKF, the length of the assimilation window is 6 month (October-March and April-September), which were adapted to the southern and northern hemispheric growing seasons to effectively incorporate the proxy records stored in trees. Due to the 6-monthly assimilation window, x^b contains the variables of 6 months.

Page 5, lines 5-7, How have the error variances been chosen? Should sigma^2 be K^2? If so, why is the error for documentary data smaller than for instrumental data? Which multiple regression?

The error variance of documentary indices without units is set to 0.25 standard deviations^2. Because in the experiments described in this paper, we did not assimilate documentary data, we will delete the information regarding to this source in the revised manuscript.

Was replaced as:

We set the error variances of instrumental temperature observations to 0.9 K^2, and of instrumental pressure data to 10 hPa^2. The error variances are rough estimates that include for instance measurement uncertainties, temporal inhomogeneities, and the fact that a station is not representative for a grid cell (see Frei, 2014; Franke et al., 2017). The errors of tree-ring proxy data are calculated as the variance of the multiple regression residuals of the H operator.

We added further explanation about the H operator:

H is the forward operator that maps the model state to the observation space (here, it is linear). H differs depending on the type of observation being assimilated. In case of tree-ring width data H extracts temperature between May and August and precipitation between April and June from the model, then these fields are transformed to observational space by using a multiple regression approach (for more details see Franke et al., 2017).

Page 5, lines 16 – 19. The notation is not clean. In line 16 it is said that R is a diagonal matrix, in line 19 that R is a scalar. The problem is that the same notation is used for an equation using the full set of observations (where R is a diagonal matrix) and for the equation when the individual observations are assimilated sequentially. Please reformulate.

It is true that the same notation is used for describing the whole system or the serial approach, but in other papers (including the original paper by Whitaker and Hamill, 2002) the notation remains unchanged. Therefore, we would like to follow the common usage of the notation and keep it as it was in the submitted paper.

Page 6, line 1, replace 'localization function' with 'the localization function'.

We replaced it.

Page 6, line 19/20, replace 'additive inflation' with 'the additive inflation', and 'hybrid' with 'the hybrid'.

We replaced it.

Page 6, line 23-25, The explanation is confusing. One can select ensemble members for the whole period or some or all ensemble members for some time steps; how exactly are the climatological state vector and the associated error covariance matrix calculated? The simulations have already been performed; why are there substantial computational costs for using a large number of ensemble members?

We added the following explanation:

The climatological state vector is created as: 1. Define the ensemble size (n) of x^{clim} ; 2. Select n random years between 1601 and 2005; 3. Every year has 30 members from which one member is randomly selected and kept; 4. The chosen members are combined in x^{clim} . x^{clim} is randomly resampled after every second assimilation cycle. Using x^{clim} in the assimilation leads to increased computational cost, which partly comes from the creation of x^{clim}. The other time consuming part comes from the updating of the climatological part after each observation is assimilated. (The standard way when observations are assimilated serially). We tested numbers between 100 and 500. From x^{clim} a climatological background-error covariance matrix (P^{clim}) can be obtained by using the ensemble perturbations.

Page 6, line 29, H^T is at the end of all terms in the equation. Can it not simply be deleted?

We deleted them.

Page 7, line 4-5. It is not clear how x^{clim} is calculated and updated, what n is, and what 'propagated' means in an offline assimilation scheme.

Please see our reply to comment Page 6, line 23-25.

We assimilate observations serially, that is observations are processed one at a time. After the first observation is assimilated and the analysis is obtained, this analysis field will become the background state for the next observation. On Figure 2 the arrows pointed from the analyses to the background state vectors are meant to indicate this process. Here, by propagating we refer to this process, that the information from the observation is also incorporated into x^{clim} and not only in x^b in most of our experiments (see Table 2).

We replaced propagate with update to avoid confusion.

Page 8, line 7, replace 'isotropic localization' with 'an isotropic localization'.

We replaced it.

Page 8, line 15, ENH is not defined.

In page 8 line 2, we introduced ENH, which stands for extratropical Northern hemisphere.

Page 9, line 3, Explain why P^b does not have full rank, why this is a problem, and what this has to do with inflation.

In the Introduction (page 2, line 20-29), some applications of covariance inflation were mentioned and how they help to avoid the negative effect of sampling error. Inflation techniques were discussed later again in section 3.2.

We rewrote the sentence as:

The main problem of ensemble-based DA techniques is the computationally affordable limited ensemble size. Due to the finite ensemble size the estimation of P^b suffers from sampling error. Applying inflation techniques is one method to mitigate its effect (see Sec. 3.2).

Page 9, lines 5-6. This is not well phrased; it is not the covered model space that is multiplied with the inflation factor.

Thank you for pointing out the mistake.

We rewrote the sentence as:

Using the multiplicative inflation method, the deviations from the ensemble mean are multiplied with a small factor (y).

Page 9, line 16, It has not yet been mentioned in the main text in section 3 that 250 members have been chosen and how they have been selected (see also earlier comment on this).

In most of our experiment, we indeed used 250 members to create the climatological state vector. We wrote that the ensemble size of the climatological part ranged between 100 and 500 members in the experiments.

Here, we added a sentence:

In most of our experiment n is 250.

Page 12, line 2, replace ‘verification’ with ‘validation’ (this is used elsewhere in the text) or ‘performance’

We replaced it.

Page 12, line 6, missing full stop after ‘2’.

We corrected it.

Page 12 and also in main text, Add a comment on whether the skill differences found are substantial and practically relevant.

We agree that this statistic was missing. In the revised paper we provide an evaluation of the skill scores of the experiments compared to the original analysis skills, using a permutation test.

Page 14, line 11, ‘kalman’ should be upper case.

We corrected it.

Page 14, line 16/17, typos and missing spaces

We corrected it.

Figs. 1, 3, 8, 10 should be bigger

We replaced the figures in the revised version.

References

Frei, C.: Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances, International Journal of Climatology, 34, 1585–1605, 2014.

Relevant changes

- All figures were modified and Figure 4 and Figure 5 were combined into one Figure.

- Eq. 1, Eq. 2 and Eq. 6 were rewritten.

- A permutation test was conducted to test significant difference between the experiments and the original version.

Impact of different estimations of the background-error covariance matrix on climate reconstructions based on data assimilation

Veronika Valler[1, 2], Jörg Franke[1, 2], and Stefan Brönnimann[1, 2]

[1]Institute of Geography, University of Bern, Bern, Switzerland
[2]Oeschger Centre for Climate Change Research, University of Bern, Bern, Switzerland

Correspondence: Veronika Valler (veronika.valler@giub.unibe.ch)

Abstract. Data assimilation has been adapted in paleoclimatology to reconstruct past climate states. A key component of ~~the assimilation system~~ some assimilation systems is the background-error covariance matrix, which controls how the information from observations spreads into the model space. In ensemble-based approaches, the background-error covariance matrix can be estimated from the ensemble. Due to the usually limited ensemble size, the background-error covariance matrix is subject to the so-called sampling error. We test different methods to reduce the effect of sampling error in a published paleo data assimilation setup. For this purpose, we conduct a set of experiments, where we assimilate early instrumental data and proxy records stored in trees, to investigate the effect of 1) the applied localization function and localization length scale; 2) multiplicative and additive inflation techniques; 3) temporal localization of monthly data, which applies if several time steps are estimated together in the same assimilation window. We find that the estimation of the background-error covariance matrix can be improved by additive inflation where the background-error covariance matrix is not only calculated from the sample covariance, but blended with a climatological covariance matrix. Implementing a temporal localization for monthly resolved data also led to a better reconstruction.

1 Introduction

Estimating the state of the atmosphere in the past is important to enhance our understanding of the natural climate variability, the underlying mechanisms of past climate changes and their impacts. To infer past climate states, two basic sources of information are available: observations, and numerical models. Climate models constrained with realistic, time-dependent ~~boundary conditions~~ external forcings provide fields that are consistent with ~~the external~~ these forcings and the model physics. Observations can be instrumental meteorological measurements, which are mainly available from the mid 19th century. Prior to this time, information from proxies stored in natural archives (like trees, speleothems, marine sediments, ice cores) or documentary data can be exploited. Observations provide important local information, however their spatial and temporal coverage is sparse.

In recent years, a novel technique, the data assimilation (DA) approach, has been adapted for paleoclimatological research. DA creates a framework to combine information from different sources. If information from observations is optimally blended with climate model simulations, the result is the best estimate of the climatic state, given the observations, given the ~~boundary conditions~~external forcings, and given the known climate physics. The field of paleo data assimilation (PDA) has undergone

profound developments, and many DA techniques have been implemented to reconstruct past climate states, such as forcing singular vectors and pattern nudging (Widmann et al., 2010), selection of ensemble members (Goosse et al., 2006; Matsikaris et al., 2015), particle filters (e.g., Goosse et al., 2010), the variation approach (Gebhardt et al., 2008), the Kalman filter and its modifications (e.g., Bhend et al., 2012; Hakim et al., 2016; Franke et al., 2017; Steiger et al., 2018). However, there are still unresolved problems, and thus, a need for improvements how to best combine observations with climate model simulations.

One popular DA method is the Kalman filter (KF; Kalman, 1960). ~~The KF provides an estimate of the state that can be shown to be optimal with linear models and Gaussian distributions (Ghil and Malanotte-Rizzoli, 1991).~~ In standard applications, the processes of the KF can be summarized in two main steps (Ide et al., 1997). In the update step, the background state and the uncertainty of the background state provided by the model simulation are adjusted by assimilating new observations. In the forecast step, the updated state, called the analysis, and the uncertainty of the analysis are propagated forward in time. These processes are repeated when new observations become available. However, in PDA, the forecast step is usually neglected, that is the filter is used offline (e.g., Franke et al., 2017). Because the process is not cycled, the background state is obtained from a pre-computed model simulation. In some previous PDA studies, the background state is constructed once from the model simulation, and later, the same state is used in every assimilation window (Steiger et al., 2018, and references therein); we refer to them as stationary (forcing-independent) offline DA methods. In other PDA studies, the background state is specific for the current assimilation window, that is, the state changes in each assimilation window according to the forcings (Bhend et al., 2012; Franke et al., 2017); we call them transient (forcing-dependent) offline DA methods.

An essential component of the KF is the uncertainty of the background state. ~~The true climate state is not known, therefore it has to be estimated.~~ In ensemble-based approaches, an ensemble of the background state provides estimation of the truth, represented by the ensemble mean, and the perturbations from the mean are used to estimate the uncertainty, represented by the background-error covariance matrix. Ensemble-based KFs are ~~simplification~~ approximations of the KF, because the true state is usually sampled with a few tens to a few hundreds of ensemble members. The limited ensemble size leads to errors in the estimation of the background-error covariance matrix. This effect is known as the sampling error.

Two methods are commonly used in online ensemble-based KF approaches to reduce the negative effect of sampling error: inflation (e.g., Anderson and Anderson, 1999), and localization (e.g., Hamill et al., 2001) of the background-error covariance matrix. A simple inflation technique is the multiplicative inflation (Anderson and Anderson, 1999), which compensates for potential underestimation of the analysis error. Multiplicative inflation helps to maintain a more realistic distribution of the ensemble members by increasing the deviation of the members from the ensemble mean at each DA cycle (Anderson and Anderson, 1999)~~, which~~ . However, the underestimation of the analysis error is of minor importance in offline approaches, because the ensemble members are not propagated forward in time. Covariance inflation, besides reducing the sampling error, ~~also accounts~~ can also account for underestimated model error. In the additive inflation technique, the covariances are inflated by e.g., adding an additional error term to the background-error covariances (Houtekamer et al., 2005). Covariance localization removes long-range spurious covariances in the background-error covariance matrix that occur by chance due to a limited sample size. Several localization techniques have been proposed, from a simple cut-off radius approach (Houtekamer and Mitchell, 1998) to more sophisticated ones (Houtekamer and Mitchell, 2001; Hamill et al., 2001). By applying covariance

localization methods, the elements of the background-error covariance matrix are modified, and in the standard approach the covariances are forced to approach zero at a certain separation length from the location of the observation. This is achieved by multiplying the background-error covariance matrix element-wise with a distance-dependent function. In practice, this function is often estimated by a Gaussian localization function, recommended by Gaspari and Cohn (1999).

In stationary offline PDA studies, the time-dependent background-error covariance matrix is ~~often~~ replaced by a constant covariance matrix (e.g., Steiger et al., 2014). By using a constant background-error covariance matrix in the update step, the dependence on the climate state is lost. However, it is possible to estimate the covariance matrix from a much larger ensemble size, which reduces the sampling error. If the constant covariance matrix is built from a large enough sample size, representing different climate states, it can be successfully used in the assimilation process (Steiger et al., 2014).

Covariance inflation and localization techniques are used and under improvement in weather forecasting (e.g., Bowler et al., 2017), but have not been yet sufficiently explored for PDA. In this paper, we discuss three possibilities to improve the estimates of background error, relevant to our PDA method:

- using a two-dimensional multivariate Gaussian function as a horizontal localization function to test the hypothesis of longer correlation length scales in zonal than meridional direction.
- applying covariance inflation techniques. In the multiplicative inflation technique, a constant factor is used to inflate the deviations from the ensemble mean. In the ~~other~~ additive method, the background-error covariance matrix is calculated as the sum of the sample covariance matrix plus a climatological background matrix, where the climatological background is based on all ensemble members of multiple years. This larger sample size decreases the chances of spurious correlations.
- adding temporal localization to the background-error covariance matrix. Multiple time steps are combined in one assimilation window to efficiently assimilate seasonal paleodata. In case of monthly observations, covariances between the months have been used to update all six months (Franke et al., 2017).

This paper is structured as follows: An overview of our PDA approach, introducing the model, the observational network and the offline DA technique is given in Sect. 2. Section 3 describes the experimental framework. In Sect. 4 the results are presented and each experiment followed directly by a discussion. We summarize our experiments in Sect. 5.

2 Ensemble Kalman Fitting Framework

2.1 Model Simulation: CCC400

We start ~~form~~ from an existing DA system, which is described in Bhend et al. (2012) and Franke et al. (2017). ~~It uses a~~ We use the same atmospheric model simulation as in the previous studies. The model simulation, termed as Chemical Climate Change over the Past 400 years (CCC400), has 30 ~~member ensemble of atmospheric model simulations as~~ ensemble members, that are used as background to reconstruct monthly climate states between 1600 and 2005. Simulations were performed with the

ECHAM5.4 climate model (Roeckner et al., 2003) at a resolution of T63 with 31 levels in the vertical. The 30 ensemble members were forced and driven with the same external forcings and with the same boundary conditions. For sea-surface temperatures (SSTs), which have a particularly large effect on the simulations, the reconstruction by Mann et al. (2009) was used. ~~This is the only~~ At the time when the model simulation was run, this was the only available global gridded SST reconstruction that ~~dates back till~~ dated back until 1600. The surface temperature reconstruction by Mann et al. (2009) is based on a multiproxy network and was produced by a climate field reconstruction method. The SST reconstruction by design captures interdecadal variations (Mann et al., 2009), hence intra-annual variability dependent on a El Niño/Southern Oscillation reconstructions (Cook et al., 2008) was added to the SST fields. Further ~~boundary conditions~~ forcings include solar irradiance, ~~land-surface parameters,~~ volcanic activity, and greenhouse gas concentrations ~~(for more details see Bhend et al., 2012; Franke et al., 2017) . The~~ (for more details see Bhend et al., 2012; Franke et al., 2017). The land-use reconstruction by Pongratz et al. (2008) was used to derive the land-surface parameters. The 6-hourly output fields provided by the model were transformed to monthly means. To reduce the computational burden only every second grid points in the latitude and longitude were selected. We limit the analysis in this study to 2m-temperature, precipitation and sea-level pressure.

2.2 Observational network

In this study, we use the same observational network of tree-ring proxies, documentary data and early instrumental measurements as described in Franke et al. (2017) (Fig. 1); but we only assimilate tree-ring proxies and instrumental data. The temporal resolution of the instrumental air temperature and sea-level pressure measurements ~~, as well as the documentary temperature data,~~ is monthly. The tree-ring proxy records have annual resolution. Trees respond to a locally varying growing seasons. We consider temperature from May ~~till~~ until August and precipitation from April ~~till~~ until June to possibly affect tree-ring width data. The maximum latewood density proxies were considered to be affected by temperature over May ~~till~~ until August. The observations were quality checked before the assimilation, and outliers which were more than 5 standard deviation away from the calculated 71-year running mean were discarded, both for instrumental and proxy data. ~~The documentary data were manually screened.~~

2.3 Assimilation method

In our paleoclimate reconstruction, we combine the CCC400 model simulation with the observations as described above by implementing a modified version of the ensemble square root filter (EnSRF; Whitaker and Hamill, 2002). This ensemble-based DA method is called ensemble Kalman fitting (EKF; Franke et al., 2017). In fact, the EKF is an offline version of the EnSRF; and the update step of the EKF remains the same as of the EnSRF. EKF is described in more detail in Bhend et al. (2012) and Franke et al. (2017). Here we shortly highlight the most important aspects of the EKF. The update ~~equation~~ step in the EnSRF

scheme has two parts: updating the mean ($\overline{\boldsymbol{x}}$), and for each member, the deviation from the mean ($\boldsymbol{x'}$). They are calculated as

$$\overline{\boldsymbol{x}}^a = \overline{\boldsymbol{x}}^b + \mathbf{K}\left(\boldsymbol{y} - \mathbf{H}\overline{\boldsymbol{x}}^b\right) \tag{1}$$

$$\boldsymbol{x'}^a = \boldsymbol{x'}^b + \tilde{\mathbf{K}}\left(\boldsymbol{'} - \mathbf{H}\boldsymbol{x'}^b\right), \text{with } \boldsymbol{'} = 0 \tag{2}$$

where $\mathbf{K}$ and $\tilde{\mathbf{K}}$ are

$$\mathbf{K} = \mathbf{P^b H}^{\mathrm{T}}\left(\mathbf{H P^b H}^{\mathrm{T}} + \mathbf{R}\right)^{-1} \tag{3}$$

$$\tilde{\mathbf{K}} = \mathbf{P^b H}^{\mathrm{T}}\left(\left(\sqrt{\mathbf{H P^b H}^{\mathrm{T}} + \mathbf{R}}\right)^{-1}\right)^{T} \times \left(\sqrt{\mathbf{H P^b H}^{\mathrm{T}} + \mathbf{R}} + \sqrt{\mathbf{R}}\right)^{-1} \tag{4}$$

The background state vector ($\boldsymbol{x^b}$) contains the variables of interest from CCC400 (Table 1). In the EKF, the length of the assimilation window is 6 month (October-March and April-September), which ~~were~~ was adapted to the southern and northern hemispheric growing seasons to effectively incorporate the proxy records stored in trees. ~~Hence~~ Due to the 6-monthly assimilation window, $\boldsymbol{x^b}$ contains the variables of 6 months~~(October–March, April–September)~~. $\boldsymbol{x^a}$ stands for the analysis state vector. $\mathbf{H}$ is the forward operator that maps the model state to the observation space (here, it is linear). $\mathbf{H}$ differs depending on the type of observation being assimilated~~(see Franke et al., 2017).~~. In case of tree-ring width data $\mathbf{H}$ extracts temperature between May and August and precipitation between April and June from the model, then these fields are transformed to observational space by using a multiple regression approach (for more details see Franke et al., 2017). $\boldsymbol{y}$ represents the observations. $\mathbf{K}$ is the Kalman gain matrix, and $\tilde{\mathbf{K}}$ is the reduced Kalman gain matrix. $\mathbf{P^b}$ is the background-error covariance matrix, estimated from the 30 ensemble members. A common assumption is to treat the observation-error covariance matrix ($\mathbf{R}$) as a diagonal matrix: it is presumed that the elements of $\mathbf{R}$ are uncorrelated. Therefore, the observations can be processed serially. We set the error variances of instrumental temperature observations to 0.9 K^2, and of instrumental pressure data to 10 ~~. The defined error variance of documentary temperature data is 0.25 σ^2, while the~~ hPa^2. The error variances are rough estimates that include for instance measurement uncertainties, temporal inhomogeneities, and the fact that a station is not representative for a grid cell (see Frei, 2014; Franke et al., 2017). The errors of tree-ring proxy data are calculated as the variance of the multiple regression residuals of the $\mathbf{H}$ operator. The assimilation is conducted on the anomaly level: we subtract both from model and from observational data their 71-yr running mean in order to deal with the biases related to systematic model errors and inconsistent low-frequency variability in the paleodata.

The use of DA in an offline manner is typical in paleoclimate reconstructions (e.g., Dee et al., 2016). Bhend et al. (2012) argue that the assimilation step is too long for initial conditions to matter, whereas there is some predictability from the boundary conditions. In addition, Matsikaris et al. (2015) found that both online and offline DA methods perform similarly in their paleoclimate reconstruction setup. Furthermore, the offline DA is advantageous as it allows using the pre-computed simulations. In our case, we can use CCC400 (Bhend et al., 2012) and test the method without having to repeat the simulations.

2.4 Spatial localization

As $\mathbf{R}$ is a diagonal matrix the EKF can be used to assimilate the observations one by one~~.~~, that is after the first observation is assimilated and the analysis is obtained, this analysis field becomes the background state for the next observation (see the arrow pointing from $\boldsymbol{x}^{\boldsymbol{a}}$ to $\boldsymbol{x}^{\boldsymbol{b}}$ on Fig. 2). This serial implementation makes the calculation of $\mathbf{P^b}$ simpler. $\mathbf{H}$ ~~is then~~ becomes a vector (not a matrix) of the same length as $\boldsymbol{x}^{\boldsymbol{b}}$. It is zero everywhere except for few elements (those required to model the observation). This translates to only a few columns of $\mathbf{P^b}$ that are actually required. $\mathbf{HP^bH^T}$ and $\mathbf{R}$ are then scalars (Whitaker and Hamill, 2002). This procedure also makes the localization simpler, as it needs to be applied only to those columns. In the original setup the elements of $\mathbf{P^b}$ were Schur-product with a distance-dependent function (see Eq. (7) in Franke et al., 2017). For all the variables in the state vector, the same Gaussian function was used but with different localization length scale parameters (Table 1). The localization length scale parameters are defined based on the spatial correlation of the variables in the monthly CCC400 model simulation fields. For the cross-covariances between two variables, the smaller localization length scale of the two variables is applied. With the serial implementation, the calculation and localization of $\mathbf{P^b}$ is significantly simplified.

3 Experiment design

Franke et al. (2017) produced a monthly global paleoclimatological data set by using the EKF method. We leave most of the original setup unchanged and mainly focus on the estimation of $\mathbf{P^b}$. To investigate the performance of the EKF some aspects involving localization and estimation of the $\mathbf{P^b}$ ~~matrix~~ were tested. An overview of all experiments conducted in this study is given in Table 2. The results of the various experiments are evaluated in terms of performance measures, which then compared to those obtained with the original setup.

3.1 Spatial localization

In most of the studies, the localization function is implemented in an isotropic manner. In the original setup, the same horizontally isotropic localization function was used with different localization parameters. However, such spatial symmetries may not be realistic. In the real atmosphere, correlation lengths might be longer in the zonal than in the meridional direction, due to the prevailing winds and the weaker large-scale temperature gradients in this direction. On multi-annual to multi-decadal time scales multiple processes act in meridional direction, e.g. a widening/shrinking of the Hadley cell, shifts of the Inter Tropical Convergence Zone or changes in atmospheric modes like the Atlantic Multi-Decadal Oscillation or the North Atlantic Oscillation. These can shift the zonal circulation northward or southward but the zonal coherence will be less effected. Hence, instead of using a circular Gaussian function, we conducted an experiment with a spatially anisotropic localization function

$$C = \exp\left(-\frac{1}{2}\left(\frac{d_z^2}{L_z^2} + \frac{d_m^2}{L_m^2}\right)\right), \tag{5}$$

where d_z and d_m are the distances from the selected grid box in the zonal and meridional directions, respectively. L_z and L_m are the length scale parameters used for localization in the zonal and meridional directions, respectively. As a first experiment

we tested a 2:1 ratio for L_z:L_m. We used the values from Table 1 in the meridional direction and doubled them in the zonal direction. Thus, the resulting localization function has an elliptical shape.

3.2 Inflation techniques

Covariance inflation techniques are another possible method to compensate for errors in the DA system (Whitaker et al., 2008). The multiplicative inflation technique uses a small factor γ ($\gamma > 1$) with which the $\boldsymbol{x'^b}$ is multiplied (Anderson and Anderson, 1999). This type of covariance inflation accounts for filter divergence due to sampling error (Whitaker and Hamill, 2002), but can be also applied to take into account system errors (Whitaker et al., 2008). We conducted some experiments using multiplicative inflation, although in our offline approach, filter divergence is not the main concern as $\mathbf{P^b}$ ~~is~~ ensemble members are not propagated in time.

The other methodology that we adapt, shows similarities with the additive inflation technique (e.g., Houtekamer and Mitchell, 2005) and with the hybrid DA scheme (e.g., Clayton et al., 2013). In both methods $\mathbf{P^b}$ is modified by either adding model error (Houtekamer and Mitchell, 2005) or a so-called climatological covariance matrix (Clayton et al., 2013) to $\mathbf{P^b}$. This has given rise to the idea of generating a climatological ensemble in order to alleviate the effect of the small ensemble size. In the original setup $\mathbf{P^b}$ is approximated from only 30 members. Here, we additionally build a climatological state vector ($\boldsymbol{x^{clim}}$) from randomly selected ensemble members from our 400-year long model simulation. The number of ensemble members should be higher than the original ensemble size, but still computationally affordable. The climatological state vector is created as: 1. Define the ensemble size (n) of $\boldsymbol{x^{clim}}$; 2. Select n random years between 1601 and 2005; 3. Every year has 30 members from which one member is randomly selected and kept; 4. The chosen members are combined in $\boldsymbol{x^{clim}}$. $\boldsymbol{x^{clim}}$ is randomly resampled after every second assimilation cycle. Using $\boldsymbol{x^{clim}}$ in the assimilation leads to increased computational cost, which partly comes from the creation of $\boldsymbol{x^{clim}}$. The other time consuming part comes from the updating of the climatological part after each observation is assimilated. (The standard way when observations are assimilated serially). We tested numbers between 100 and 500. From $\boldsymbol{x^{clim}}$ a climatological background-error covariance matrix ($\mathbf{P^{clim}}$) can be obtained by using the ensemble perturbations. The background-error covariance matrix used in the blending experiments ($\mathbf{P^{blend}}$) is built as a linear combination of the sample covariance matrix ($\mathbf{P^b}$) and the climatological covariance matrix ($\mathbf{P^{clim}}$):

$$\mathbf{P^{blend}}\mathbf{H}^{\mathrm{T}} = \beta_1 \mathbf{P^b}\mathbf{H}^{\mathrm{T}} + \beta_2 \mathbf{P^{clim}}\mathbf{H}^{\mathrm{T}}, \tag{6}$$

where β_1, β_2 mean the weights given to the covariance matrices. The sum of the weights is unity.

Figure 2 shows the main steps of the blending assimilation process. First, the covariance matrices were localized separately, then we blended them according to the given weights. We conducted several experiments to tune the ratio between the two covariance matrices while using different localization length scale parameters (L) (Table 2). We used the same Ls for localizing $\mathbf{P^b}$ in most of our experiments to evaluate improvements in comparison with the original setup. For this study, we calculated the latitudinal dependency of correlation of the state variables from a bigger ensemble of the model than in Franke et al. (2017) . The result suggested that longer Ls can be applied in the tropics and the L of precipitation is probably too strict. Based on the rather strict Ls in the previous study and the assumption that the covariances can be better estimated from a bigger ensemble,

we used doubled length scale parameters (2L) in some of the experiments for localizing the climatological covariances. In this case, the L for temperature is 3000 km, which means that the correlation is decreased close to zero approximately 6000 km away from the observation.

Since observations are assimilated serially, we also update $\boldsymbol{x}^{clim}$ after an observation is assimilated with the same Kalman gain matrices as $\boldsymbol{x}^{b}$. Thus, in the assimilation process we ~~propagate~~ update 30 + n ensemble members~~, which leads to an increased computational time.~~.

3.3 Temporal localization

Localizing observations in time is a special feature of the EKF due to its 6-month assimilation window. Having the state vector in half-year format, every month within the October–March or April–September time window is updated by each single observation. To test whether the covariances between a single observation and the multivariate climate fields are correctly captured, we ran an instrumental-only experiment with temporal localization. We set covariances between different months to zero.

3.4 Skill scores

The EKF method is tested with different localization functions and with a set of mixed background-error covariance matrices as described above. We have performed the experiments by assimilating either only proxy records (proxy-only) or only instrumental data (instrumental-only). The proxy-only experiments were carried out between 1902 and 1959, because many proxy records already end in the 1960s, while the instrumental-only experiments were tested over the 1902–2002 period. We separated the different observation types to see whether different settings perform better depending on the type of data being assimilated. We do not compare proxy-only results with instrumental-only results, hence the difference in time periods used does not matter; we simply use the longest possible time period. To evaluate the reconstructions we examined two verification measures: correlation coefficient, and reduction of error (RE) skill score (Cook et al., 1994). We use the CRU TS 3.10 dataset (Harris et al., 2014) for reference in the validation process. The presented verification measures are functions of time. Correlation is calculated between the absolute values of the ensemble mean of the analysis and the reference series at each grid point. The RE compares, in our case, the reconstruction with ~~a no knowledge prediction (such as a climatology)~~the model simulation, both expressed as deviations from a reference.

$$RE = 1 - \frac{\sum(x_i^u - x_i^{ref})^2}{\sum(x_i^f - x_i^{ref})^2} \tag{7}$$

where x^u is the ensemble mean of the analysis, x^f is the ensemble mean of the model background state, x^{ref} is the reference dataset and *i* refers to the time step. The RE skill scores are computed based on anomalies with respect to the 71-year running climatologies. Note that x^f comes from a forced model simulation, therefore it already has skill compared with a climatological state vector. The RE is 1 if the x^u is equal to x^{ref}. Negative RE values indicates that the background state is closer to the reference series than the analysis.

To test which experiments have significantly different skill compared with the original skill, we carried out a permutation test following the method described in DelSole and Tippett (2014). Permutation was performed 10000 times. If the difference between the median of the experiment and the median of the original data falls outside of the 95% confidence level of the interval calculated from permuted data, then the experiment is significantly different from the original data.

In the next section, we will focus on analysing the result of the experiments mainly over the extratropical Northern hemisphere (ENH), because most of the data are located in this region. The skill scores refer to seasonal averages of the ensemble mean.

4 Results and Discussions

4.1 Localization function

4.1.1 Results

We compared the original setup applying an isotropic localization function and the experiment in which an anisotropic localization function was used, to test whether we can obtain a more skilful reconstruction by implementing anisotropic localization method. As an example of the spatial reconstruction skill, we show the RE values of temperature (Fig. 3). The figures reveal that the type of localization function only resulted in small differences in both experiments. Nonetheless, there are larger areas of negative RE values (Greenland, Siberia) with the anisotropic localization function in the proxy-only experiment (Fig. 3~~f~~). In the instrumental-only experiment the decrease of RE values occur in the northern high latitudes and in the Tibetan plateau in both seasons (Fig. 3~~d, Fig. 3e~~). To have a better overview how the skill scores changed we summarize their distributions with the help of box plots. Figure 4 shows ~~how the correlation coefficients of~~ the differences of skill scores between the aniso experiment and the original skill for the three variables (temperature, precipitation and sea-level pressure) ~~were affected~~ in the ENH region~~by using the anisotropic localization function~~. In the instrumental-only experiment correlation values of temperature and sea-level pressure decreased in both season while for precipitation it remained mostly unchanged. The RE values show that the experiments with anisotropic localization function reduced the skill of the reconstructions, but the extent of the reduction varies with the variables and with the seasons (Fig. ~~??~~4). In general, the same holds for the proxy-only experiment (Fig. 4~~, Fig. ??~~).

4.1.2 Discussion

In a previous ozone reconstruction study, a seasonally and latitudinally varying localization method was tested which mostly positively affected the analysis (Brönnimann et al., 2013). Here, we increased the zonal distances to see if we can use the information of the observations for a larger region. However, the verification measures are shifted more to the negative direction. We assume that the degraded skill of the reconstruction is due to the choice of too long L_z, hence spurious correlations were not removed. Using anisotropic localization (doubling the Ls only in the zonal direction) consistently makes the reconstruction worse.

4.2 Inflation experiments

4.2.1 Results

~~The rank-deficiency of $\mathbf{P^b}$ is the~~ The main problem of ensemble-based DA techniques ~~. To improve this issue we have tested different inflation methods.~~ is the computationally affordable limited ensemble size. Due to the finite ensemble size the estimation of $\mathbf{P^b}$ suffers from sampling error. Applying inflation techniques is one method to mitigate its effect (see Sec. 3.2).

Using the multiplicative inflation method, the ~~model space covered by the ensemble is extended by being~~ deviations from the ensemble mean are multiplied with a small factor (γ). To find the optimal γ a set of experiment runs is required. We used γ = 1.02 and γ = 1.12 in our experiments, where only instrumental data were assimilated. We chose γ from a range that was previously tested by Whitaker and Hamill (2002). Multiplying the deviations from the ensemble mean with γ = 1.02 in the assimilation process hardly affected the skill of the reconstruction over the ENH region (not shown). When we increased the value of γ to 1.12, the RE values slightly decreased (not shown). We did not carry out further experiments since based on the results randomly increasing the error in background field did not lead to improvement.

In the other set of experiments, we used $\mathbf{P^{blend}}$ in the update equation (Eq. 6). The experiments were run with using β_2 equal to 0.25, 0.50, 0.75, and 1 to estimate the $\mathbf{P^{blend}}$ (denoted 25c, 50c, etc.). Besides the varying weight given to $\mathbf{P^{clim}}$, the applied Ls on $\mathbf{P^b}$ and $\mathbf{P^{clim}}$ differed as well. Three Ls were used: No localization (termed no), applying Ls as in Table 1 (L) and doubling these numbers (2L). Different combinations of the fraction of $\mathbf{P^{clim}}$ and Ls were termed accordingly (e.g., 50c_PbL_Pc2L).

We expect that estimating the covariances from ~~250 members~~ a bigger ensemble size (n=100-500) instead of 30 members leads to a more accurate background matrix. In most of our experiment n is 250. Hence, $\mathbf{P^{clim}}$ is likely less affected by the sampling error implying that long-range spurious correlations are less prominent, which makes localization less needed. We presume that using $\mathbf{P^{blend}}$ helps to better reconstruct areas which were characterized with lower skill score values in the original setup and to improve the estimation of unobserved climate variables. The reconstruction skill of the blending experiments is always calculated from $\boldsymbol{x}^{\boldsymbol{a}}$ (Fig. 2).

For the ENH region we present how the verification measures changed by replacing $\mathbf{P^b}$ with $\mathbf{P^{blend}}$ in the assimilation process. We conducted an experiment without localizing $\mathbf{P^{clim}}$ and using Ls from Table 1 on $\mathbf{P^b}$ in the ~~constructionn~~ construction of $\mathbf{P^{blend}}$. However, the skill of the reconstruction was largely reduced, implying that 250 members are not enough to avoid localization altogether (not shown).

Figure 5 and Figure 6 show the distribution of the differences of the skill scores between the experiments and the original analysis, for correlation coefficients and RE values ~~,~~respectively. Depending on the variables and the data type being assimilated, different setups perform ~~best~~better. In case of assimilating only instrumental data, ~~the most skilful~~ one of the largest increase of median for temperature reconstruction was obtained from the 100c_PcL experiment in both seasons (Fig. 5~~a and b~~, Fig. 6~~a and b~~). Precipitation records were not assimilated, thus a reasonable estimation of the cross-variable covariances is essential. The skill of the precipitation reconstruction with the original setup, in terms of correlation, is better than the forced simulation (~~Fig. 5d). However~~not shown); however, the RE ~~skill score are rather decreased with the original setup over the ENH region~~ values are negative over large regions in the ENH (Fig. 7~~a and b). The settings of 75c~~). Using $\mathbf{P^{blend}}$ in the assimilation,

with e.g. the settings of 50c_PbL_Pc2L experiment lead to ~~improved analysis~~ more positive RE skill (Fig. 7~~c and d~~). The biggest improvement, in terms of RE skill score, was found in Europe (Fig. 7~~c and d~~). The ~~75c~~50c_PbL_Pc2L analysis also has higher skill in North-America, especially in the summer season (Fig. 7~~d~~). The ~~largest improvement in~~ skill of the sea-level pressure reconstruction ~~was achieved~~ also improved in the 50c_PbL_Pc2L experiment (Fig. 5~~g and h~~, Fig. 6~~g and h~~). In the proxy-only experiments, 75c_PbL_Pc2L is among the best performing experiments for all the variables (Fig. 5~~c, f, i;~~ , Fig. 6~~c, f, i~~).

We also investigated the effect of the ensemble size in the estimation of $\mathbf{P}^{\mathbf{clim}}$. To test whether further improvements can be achieved by doubling the ensemble size of $\boldsymbol{x}^{clim}$, we ran an experiment with the following setup: β_1 and β_2 are equally weighted, and L and 2L is applied on $\mathbf{P}^{\mathbf{b}}$, $\mathbf{P}^{\mathbf{clim}}$, respectively (Table 2). In the experiment we assimilated only instrumental data. The skill scores of $\boldsymbol{x}^{a}$ (corr, RE) from the 500 ensemble members experiment showed no marked improvement compared with the same experiment with 250 ensemble members. An additional experiment was carried out with the same setup but using only 100 ensemble members in the construction of $\boldsymbol{x}^{clim}$. The verification measures of the 50c_PbL_Pc2L_100m experiment are higher than the original one, and the distribution of the skill scores over the ENH region is very similar to what we obtain by using 250 members in $\mathbf{P}^{\mathbf{clim}}$ for temperature and precipitation. However, the sea-level pressure fields from the 50c_PbL_Pc2L have higher skill than in the 50c_PbL_Pc2L_100m experiment (not shown).

Furthermore, we conducted two experiments in which only $\boldsymbol{x}^{b}$ was updated after an observation was assimilated, and $\boldsymbol{x}^{clim}$ was kept constant in the assimilation window. However, the ensemble members of $\boldsymbol{x}^{clim}$ were randomly reselected for each year (October–September). The advantage of this setup compared to the setup described in Sect. 3.2 is that, it is computationally less demanding since only the original 30 members keep being updated with the observations. In the first test, we give β_2=0.75 weight to $\mathbf{P}^{\mathbf{clim}}$ with 2Ls. In the second test β_2=1, that is only $\mathbf{P}^{\mathbf{clim}}$ used for updating $\boldsymbol{x}^{b}$ and for localization the Ls in Table 1 were applied. By comparing the skill of the reconstructions without and with updating the climatological part, we see that the skill scores are higher when the climatological part is also updated with the information from the observations (Fig. 8). The only exception is the correlation values of sea-level pressure: when keeping the climatological part constant, they are slightly higher in both seasons (Fig. 8~~e and f~~). Nonetheless, by keeping the climatological part static in one assimilation window, the experiments still outperform the original reconstruction (Fig. 8).

4.3 ~~Discussion~~

4.2.1 Discussion

We have tested a number of configurations of the mixed covariance matrix $\mathbf{P}^{\mathbf{blend}}$ to evaluate the effect of the sampling error. In numerical weather predication (NWP) applications, various methods have been designed to better estimate the errors of the background state. In hybrid DA systems, the advantages of variational and ensemble Kalman filter techniques are combined (Hamill and Snyder, 2000; Lorenc, 2003). In another method, the background-error covariances are obtained from an ensemble of assimilation experiments performed by a variational assimilation system (Pereira and Berre, 2006). In an additive inflation experiment, a term is added to the $\boldsymbol{x}^{a}$ to account for the errors of the DA system (Whitaker et al., 2008).

In our implementation, $\mathbf{P}^{\mathbf{blend}}$ is calculated from $\boldsymbol{x}^{b}$ and $\boldsymbol{x}^{clim}$. Using $\mathbf{P}^{\mathbf{blend}}$ in the assimilation process improved on the reconstruction performed with the original setup. The skill scores show the largest improvement in the sea-level pressure reconstruction. Moreover, the skill of the precipitation reconstruction also improved, indicating that $\mathbf{P}^{\mathbf{clim}}$ helps to better estimate the cross-covariances of the background errors between the variables. In general, increasing the weight of $\mathbf{P}^{\mathbf{clim}}$ in forming $\mathbf{P}^{\mathbf{blend}}$, positively affected the skill of the analysis. The 100c_PcL experiment, in which $\mathbf{P}^{\mathbf{blend}}$ is equal to $\mathbf{P}^{\mathbf{clim}}$, is similar to the DA technique used in the last millennium climate reanalysis (LMR) project (Hakim et al., 2016). In the LMR, 100 randomly chosen ensemble members form a climatological state vector, which is used in each assimilation window and is updated with the observations. In this study, $\boldsymbol{x}^{clim}$ is randomly resampled every year and primarily used in the estimation of $\mathbf{P}^{\mathbf{blend}}$. The ~~analysis of the 100c_PcL experiment is more skilful, than the original reconstruction. The~~ settings used in the 100c_PcL experiment lead to ~~the best~~ one of the largest increase in the median for temperature reconstruction when only instrumental measurements are assimilated. However, other settings ~~performed better~~ resulted in larger increase of median for different variables and observation types. By applying no localization on $\mathbf{P}^{\mathbf{clim}}$ in the 50c_PbL_PcnoL experiment we obtained a less skilful reconstruction than by using the other two localization schemes. The skills reduced especially over the areas where no local observations were assimilated. Using 2Ls for localizing the covariances of $\mathbf{P}^{\mathbf{clim}}$ in the instrumental-only experiments resulted in ~~better analysis~~ higher correlation values of sea-level pressure (50c_PbL_Pc2L) and helped to ~~better reconstruct summer precipitation~~ obtain higher correlation scores of precipitation in summer. Among the proxy-only experiments, 75c_PbL_Pc2L shows the ~~best skill~~ largest increase of median for pressure reconstruction. Here, pressure data are not assimilated, and the result suggests that by applying longer Ls, the cross-variable covariances are better treated. We tested whether the skill of the experiments performed with various settings is significantly different from the skill of the original analysis. We compared the median value of the skill scores from the experiments and the original data, and with most of the settings a significant difference was obtained for all the variables. The results of the experiments show that with a mixed covariance matrix implementation a major drawback of the ensemble-based DA system, due to the limited ensemble size, can be improved.

4.3 Localization in time

4.3.1 Results

Since six monthly time steps were combined in one state vector (one assimilation window), covariances between different months also need to be considered. An additional experiment was conducted in which the (localized) $\mathbf{P}^{\mathbf{b}}$ was multiplied with a temporal localization function when instrumental data were assimilated. This is a specific experiment due to the structure of EKF. The assimilation window in the EKF is 6-month, hence a single observation is enabled to adjust all the meteorological variables in $\boldsymbol{x}^{b}$ in a half-year time window. In the temporal localization experiment, the information from a given observation can only modify the different climate fields in its current month, while leaving all other fields of the 5 months unchanged (Table 2). In general, the skill scores indicate an improvement. The difference of RE values between the temp_loc and original experiments are mostly positive over the northern high latitude areas (Fig. 9).

4.3.2 Discussion

The higher skill scores with temporal localization (Fig. 9) indicate that the cross-covariances in time were not correctly represented by $\mathbf{P}^{\mathbf{b}}$. Hence, it is likely that in the original setup some non-physical covariances were taken into account. Applying the same assimilation scheme to another problem (estimating the two-dimensional ozone distribution from an ensemble of chemistry-climate models and historical observations), Brönnimann et al. (2013) used a localization time scale of 3 months based on empirical studies. It may be worth considering or allowing for temporal covariance in specific cases (e.g. in the case of ozone concentrations) which vary on longer time scale.

5 Conclusions

In this study, ~~an~~ a transient offline data assimilation approach was used to test the effect of the estimation of the background-error covariance matrix in a climate reconstruction. Several experiments were evaluated with different ~~verification~~ validation measures to see which background-error covariance matrix estimation techniques improve the skill of the reconstruction. The ~~validation~~ evaluation of the presented techniques suggests the following: 1. Applying an anisotropic localization function on the sample covariance matrix did not improve the reconstruction; ~~2 Constructing the background-error covariance matrix from the sample and climatological covariance matrices, allows using longer localization length scales, and it leads to higher skill scores~~2. Most of the settings, which make use of covariance estimates from a larger climatological sample, result in significantly improved skills compared to an estimation from the 30 member ensemble; 3. Assimilating early instrumental data with temporal localization leads to a better analysis. To which extent the different techniques helped in the estimation of the background-error covariance matrix varies geographically and also depends on the climate variable being reconstructed. The cross-variable covariances of the background-error covariance matrix can provide information from unobserved climate variables. Including climatological information in the estimation of precipitation has lead to a better reconstruction, especially in Europe. Estimating sea-level pressure with the blended $\mathbf{P}^{\mathbf{blend}}$ matrix also improved the skill of the reconstruction. For instance, the 50c_PbL_Pc2L experiment performs constantly better than the original setup. This study shows that results can be improved by better specifying the background-error covariance matrix. In the future we will combine all the techniques that lead to more skilful analyses to produce a climate reconstruction over the last 400 years.

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

Acknowledgements. The study was funded by the Swiss National Science Foundation (project 162668) and by the European Union (H2020/ERC grant number 787574 PALAEO-RA). The CCC400 simulation was performed at the Swiss National Supercomputing Centre CSCS. The comments of the two anonymous reviewers are gratefully acknowledged.

References

Anderson, J. L. and Anderson, S. L.: A Monte Carlo implementation of the nonlinear filtering problem to produce ensemble assimilations and forecasts, Monthly Weather Review, 127, 2741–2758, https://doi.org/10.1175/1520-0493(1999)127<2741:AMCIOT>2.0.CO;2, 1999.

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, https://doi.org/10.5194/cp-8-963-2012, 2012.

Bowler, N., Clayton, A., Jardak, M., Jermey, P., Lorenc, A., Wlasak, M., Barker, D., Inverarity, G., and Swinbank, R.: The effect of improved ensemble covariances on hybrid variational data assimilation, Quarterly Journal of the Royal Meteorological Society, 143, 785–797, https://doi.org/10.1002/qj.2964, 2017.

Brönnimann, S., Bhend, J., Franke, J., Flückiger, S., Fischer, A. M., Bleisch, R., Bodeker, G., Hassler, B., Rozanov, E., and Schraner, M.: A global historical ozone data set and prominent features of stratospheric variability prior to 1979., https://doi.org/10.5194/acp-13-9623-2013, 2013.

Clayton, A. M., Lorenc, A. C., and Barker, D. M.: Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office, Quarterly Journal of the Royal Meteorological Society, 139, 1445–1461, https://doi.org/10.1002/qj.2054, 2013.

Cook, E., D'Arrigo, R., and Anchukaitis, K.: ENSO reconstructions from long tree-ring chronologies: Unifying the differences, in: Talk presented at a special workshop on Reconciling ENSO Chronologies for the Past, vol. 500, 2008.

Cook, E. R., Briffa, K. R., and Jones, P. D.: Spatial regression methods in dendroclimatology: a review and comparison of two techniques, International Journal of Climatology, 14, 379–402, https://doi.org/10.1002/joc.3370140404, 1994.

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, 1164–1179, https://doi.org/10.1002/2016MS000677, 2016.

DelSole, T. and Tippett, M. K.: Comparing forecast skill, Monthly Weather Review, 142, 4658–4678, https://doi.org/10.1175/MWR-D-14-00045.1, 2014.

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, Scientific data, 4, 170 076, https://doi.org/10.1038/sdata.2017.76, 2017.

Frei, C.: Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances, International Journal of Climatology, 34, 1585–1605, 2014.

Gaspari, G. and Cohn, S. E.: Construction of correlation functions in two and three dimensions, Quarterly Journal of the Royal Meteorological Society, 125, 723–757, https://doi.org/10.1002/qj.49712555417, 1999.

Gebhardt, C., Kühl, N., Hense, A., and Litt, T.: Reconstruction of Quaternary temperature fields by dynamically consistent smoothing, Climate dynamics, 30, 421–437, https://doi.org/10.1007/s00382-007-0299-9, 2008.

Ghil, M. and Malanotte-Rizzoli, P.: Data assimilation in meteorology and oceanography, in: Advances in geophysics, vol. 33, pp. 141–266, https://doi.org/10.1016/S0065-2687(08)60442-2, 1991.

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

Goosse, H., Crespin, E., de Montety, A., Mann, M., Renssen, H., and Timmermann, A.: Reconstructing surface temperature changes over the past 600 years using climate model simulations with data assimilation, Journal of Geophysical Research: Atmospheres, 115, https://doi.org/10.1029/2009JD012737, 2010.

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., Steiger, N., and Perkins, W. A.: The last millennium climate reanalysis project: Framework and first results, Journal of Geophysical Research: Atmospheres, 121, 6745–6764, https://doi.org/10.1002/2016JD024751, 2016.

Hamill, T. M. and Snyder, C.: A hybrid ensemble Kalman filter–3D variational analysis scheme, Monthly Weather Review, 128, 2905–2919, https://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2, 2000.

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

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, International journal of climatology, 34, 623–642, https://doi.org/10.1002/joc.3711, 2014.

Houtekamer, P. L. and Mitchell, H. L.: Data assimilation using an ensemble Kalman filter technique, Monthly Weather Review, 126, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2, 1998.

Houtekamer, P. L. and Mitchell, H. L.: A sequential ensemble Kalman filter for atmospheric data assimilation, Monthly Weather Review, 129, 123–137, https://doi.org/10.1175/1520-0493(2001)129_0123_ASEKFF_2.0.CO;2, 2001.

Houtekamer, P. L. and Mitchell, H. L.: Ensemble Kalman filtering, Quarterly Journal of the Royal Meteorological Society, 131, 3269–3289, https://doi.org/10.1256/qj.05.135, 2005.

Houtekamer, P. L., Mitchell, H. L., Pellerin, G., Buehner, M., Charron, M., Spacek, L., and Hansen, B.: Atmospheric data assimilation with an ensemble Kalman filter: Results with real observations, Monthly weather review, 133, 604–620, https://doi.org/10.1175/MWR-2864.1, 2005.

Ide, K., Courtier, P., Ghil, M., and Lorenc, A. C.: Unified Notation for Data Assimilation: Operational, Sequential and Variational, Journal of the Meteorological Society of Japan. Ser. II, 75, 181–189, https://doi.org/10.2151/jmsj1965.75.1B_181, 1997.

Kalman, R. E.: A new approach to linear filtering and prediction problems, Journal of basic Engineering, 82, 35–45, 1960.

Lorenc, A. C.: The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var, Quarterly Journal of the Royal Meteorological Society, 129, 3183–3203, https://doi.org/10.1256/qj.02.132, 2003.

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

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

Pereira, M. B. and Berre, L.: The use of an ensemble approach to study the background error covariances in a global NWP model, Monthly weather review, 134, 2466–2489, https://doi.org/10.1175/MWR3189.1, 2006.

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

Roeckner, E., Bäuml, G., Bonaventura, L., Brokopf, R., Esch, M., Giorgetta, M., Hagemann, S., Kirchner, I., Kornblueh, L., Manzini, E., et al.: The atmospheric general circulation model ECHAM 5. PART I: Model description, 2003.

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

Steiger, N. J., Smerdon, J. E., Cook, E. R., and Cook, B. I.: A reconstruction of global hydroclimate and dynamical variables over the Common Era, Scientific data, 5, 180 086, https://doi.org/10.1038/sdata.2018.86, 2018.

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

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, 463–482, https://doi.org/10.1175/2007MWR2018.1, 2008.

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

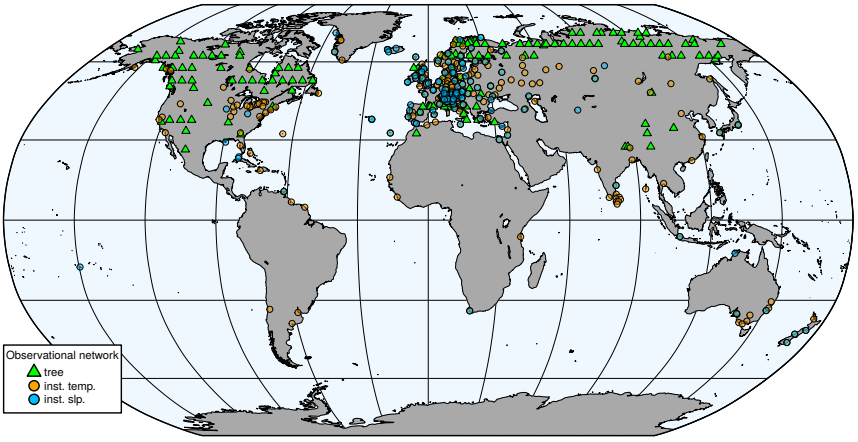

Figure 1. The observational network in 1904, before quality check.

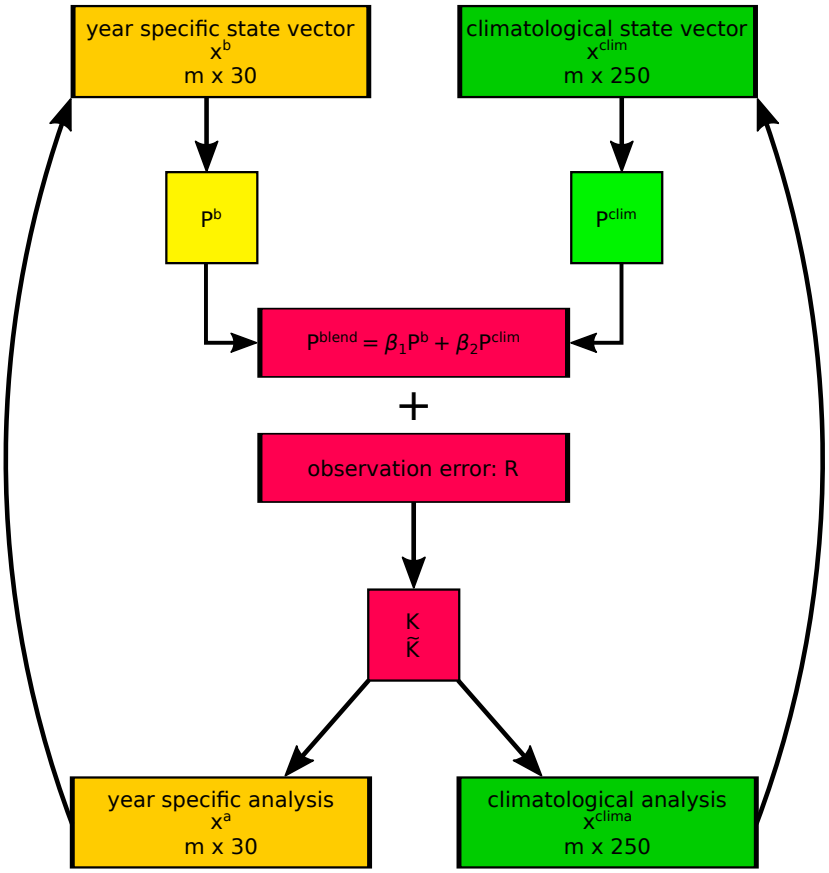

Figure 2. The main steps of the blending experiment in one assimilation window. The blended covariance matrix $\mathbf{P}^{\mathbf{blend}}$ is calculated as a linear combination from the year specific and climatological covariance matrices. The calculation of the Kalman gain ($\mathbf{K}$) and reduced Kalman gain ($\tilde{\mathbf{K}}$) matrices is the same as in Eq.3 and Eq. 4 except the covariance matrix is replaced with $\mathbf{P}^{\mathbf{blend}}$. The observation is assimilated to both state vectors and these analysis become to the starting point for assimilating the next observation.

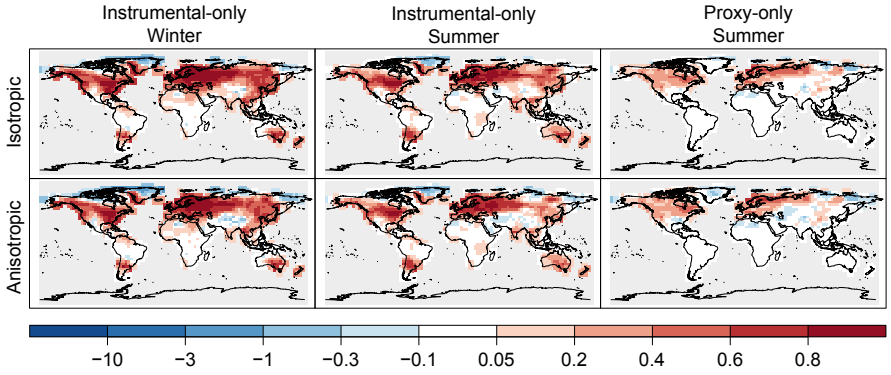

Figure 3. Spatial skill of temperature reconstruction presented by RE values, assimilating only instrumental data (~~a,b,d,e~~left and middle columns) and only proxy records (~~e,f~~right column). Comparing the skill of the reconstruction using isotropic localization function (~~a,b,e~~top row) versus an anisotropic localization function (~~d,e,f~~bottom row). ~~Panel a and d show the skill~~ Skill in the winter season ~~, while panel b, c, e,~~ (left column) and ~~f illustrate the skill~~ in the summer season (middle and right columns) are shown.

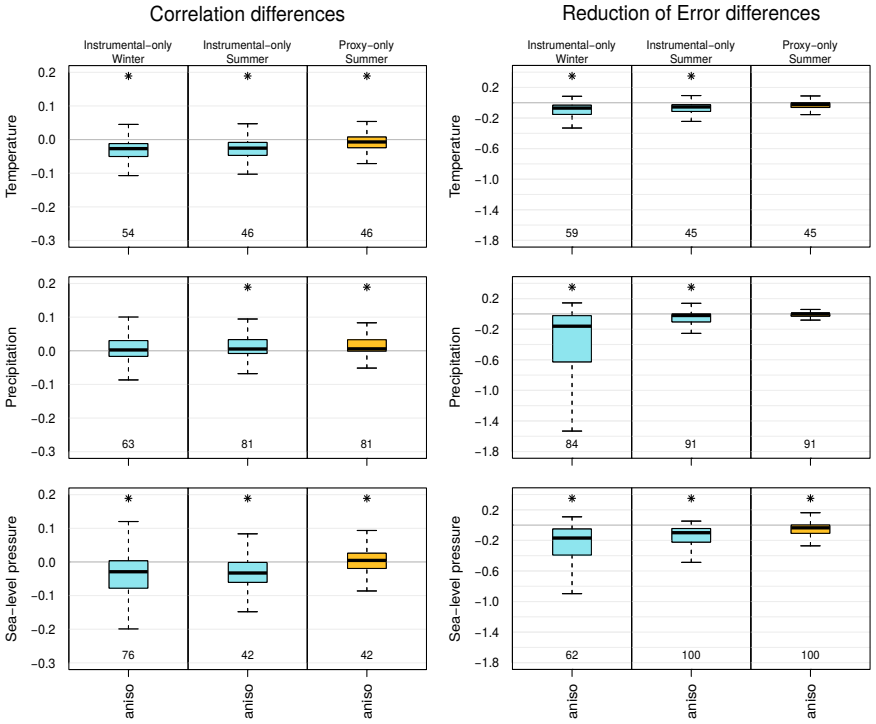

Figure 4. ~~Distribution of~~ Difference between the ~~correlation coefficient values~~ aniso experiment and the original setup in terms of skill scores over the ENH region~~in~~ . Distributions of correlation values and of RE values are on the ~~winter (~~left ~~column)~~ and ~~summer (middle and~~ right ~~columns) half-years~~ figures, respectively. Distribution of temperature (~~a,b,c~~top row), precipitation (~~d,e,f~~middle row), and sea-level pressure (~~g,h,i~~bottom row) are shown. Blue ~~is~~ colour indicates the instrumental-only experiment and yellow ~~is~~ indicates the proxy-only experiment. The midline of the box is the median. The lower (upper) border of the box is the first (third) quartile. The whiskers extend up to 1.5 times the interquartile range; beyond these distances the number of outliers are given under the box plots. The grid boxes were not area-weighted. The asterisk above the box indicates significant differences between the median of the experiment and the original setup.

~~Distribution of RE values, as in Fig. 4.~~

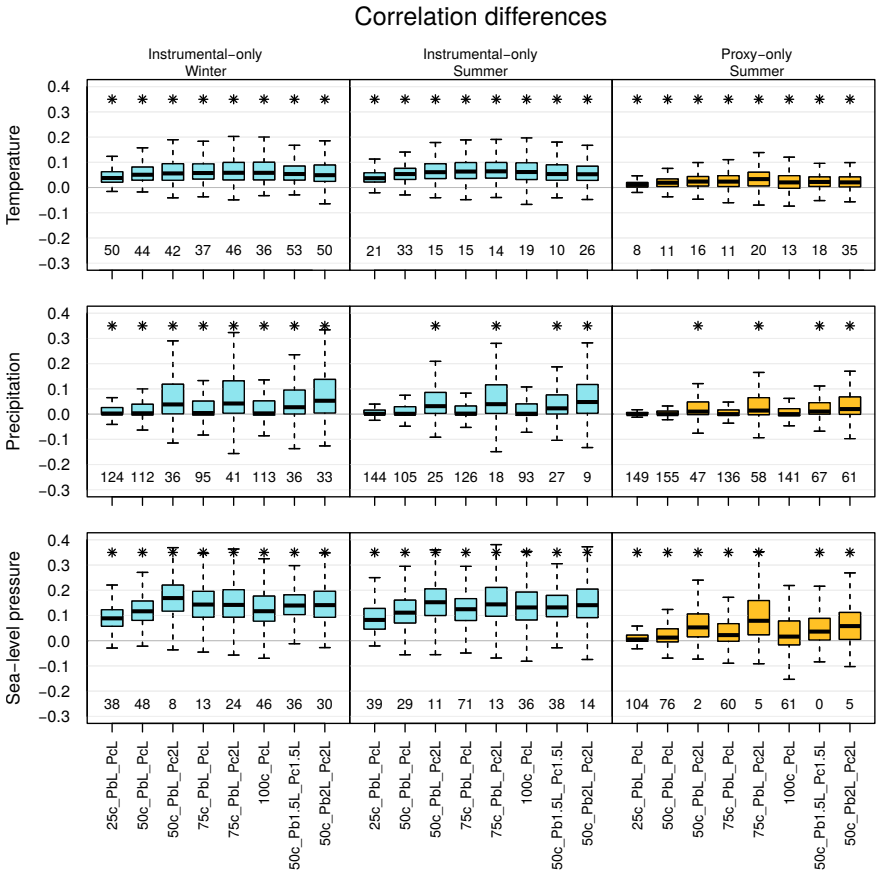

Figure 5. Distribution of correlation coefficients ~~in~~ differences between the ~~different~~ mixed background-error covariance matrix experiments ~~in~~ and the original setup over the ENH region. The left column shows the skill of the reconstruction in the winter seasons, while the middle and right columns in the summer season. The labels on the x-axis indicating the experiments. Box plot, colour, number ~~on the panels~~ and ~~colors~~ asterisk represent the same as in Fig. 4.

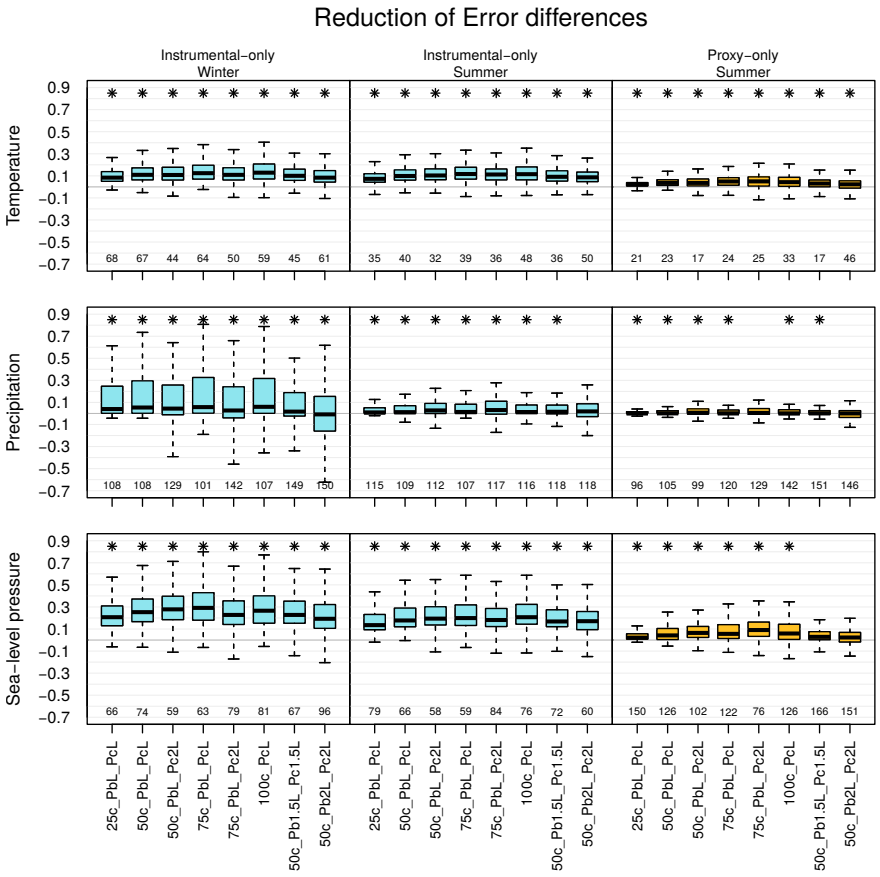

Figure 6. Distribution of RE ~~values in~~ value differences between the ~~different~~ mixed background-error covariance matrix experiments ~~in~~ and the original setup over the ENH region~~,~~; otherwise same as in Fig. 5.

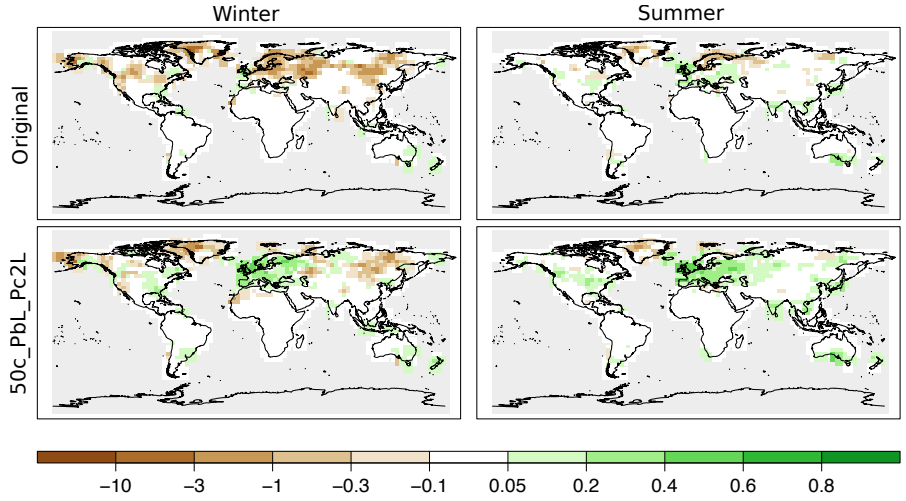

Figure 7. Spatial reconstruction skill of precipitation in terms of RE values, assimilating only instrumental data. ~~Panel a and b show~~ Top row shows the skill ~~using~~ of the original setup, and ~~panel c and d show~~ bottom row shows the result of the ~~75c~~50c_PbL_~~PcL2~~ Pc2L experiment. The skill in the winter season presented in the ~~a~~ left column and ~~c panel and~~ for the summer ~~on~~ season in the ~~b and d panels~~right column.

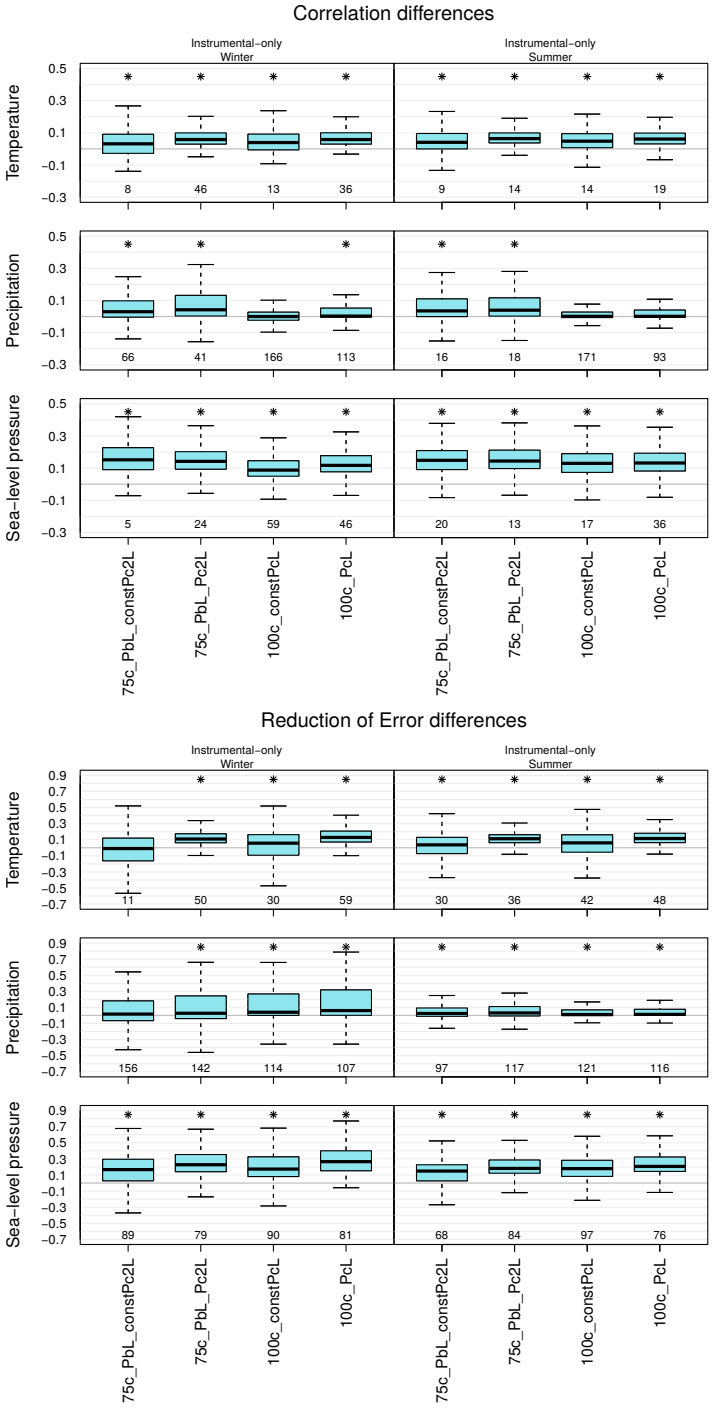

Figure 8. Distribution of skill scores over the ENH region. The skill of the original setup is compared with experiment 75c_PbL_constPc2L, 75c_PbL_Pc2L, 100c_constPcL, and 100c_PcL. Distribution of correlation coefficients in the winter (left column) and in the summer (right column) seasons. Distribution of RE values in the winter (left column) and in the summer (right column) seasons.

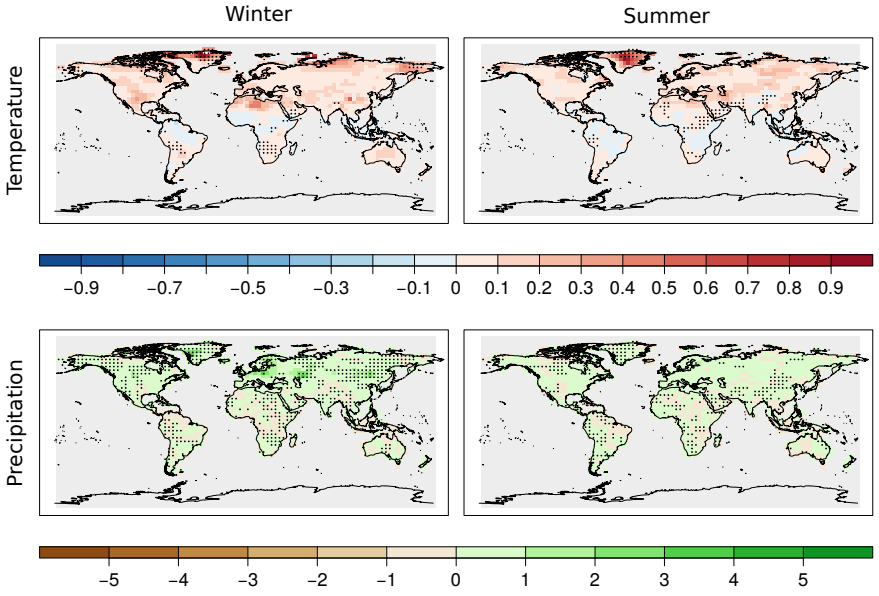

Figure 9. Difference of the RE skill between the temporally localized experiment and the original setup~~: temperature~~ , when only instrumental data are assimilated. Temperature (~~a~~top row) ~~in winter~~ and ~~(b) in summer;~~ precipitation (~~c~~bottom row) differences are shown in the winter ~~and~~ (~~d~~left column) and in the summer (right column) seasons. The black dots indicate the negative RE values in the temporally localized experiment.

Table 1. Defined localization length scale parameters

Variable	Localization length scale (km)
Temperature (2m)	1500
Precipitation	450
Sea-level pressure	2700

Table 2. Summary of the experiments carried out in this study. The name of the experiments indicate which settings were used in the assimilation. Localization refers to the shape of the localization function applied on P^b. γ is the multiplicative inflation factor. x^{clim} indicate from how many ensemble members the climatological state vector was constructed. x^{clim}const stands for keeping the climatological part in the blending experiment unchanged in one October–September time window. P^bloc indicates the localization length scale parameter applied for localizing P^b. β_2 refers to the weight given to P^{clim}. P^{clim}loc indicates the localization length scale parameter applied for localizing P^{clim}. i and p stands for instrumental-only and proxy-only observations experiments, respectively.

Name	Localization	γ	Blending					Temporal localization	Obs. type
			x^{clim}	x^{clim}const	P^bloc	β_2 (%)	P^{clim}loc		
original	iso	no						no	i,p
aniso	aniso	no						no	i,p
mul1.02	iso	1.02						no	i
mul1.12	iso	1.12						no	i
25c_PbL_PcL	iso	no	250	no	L	25	L	no	i,p
50c_PbL_PcnoL	iso	no	250	no	L	50	no	no	i
50c_PbL_PcL	iso	no	250	no	L	50	L	no	i,p
50c_PbL_Pc2L_100m	iso	no	100	no	L	50	2L	no	i
50c_PbL_Pc2L	iso	no	250	no	L	50	2L	no	i,p
50c_PbL_Pc2L_500m	iso	no	500	no	L	50	2L	no	i
50c_Pb1.5L_Pc1.5L	iso	no	250	no	1.5L	50	1.5L	no	i,p
50c_Pb2L_Pc2L	iso	no	250	no	2L	50	2L	no	i,p
75c_PbL_PcL	iso	no	250	no	L	75	L	no	i,p
75c_PbL_Pc2L	iso	no	250	no	L	75	2L	no	i,p
75c_PbL_constPc2L	iso	no	250	yes	L	75	2L	no	i
100c_PcL	iso	no	250	no		100	L	no	i,p
100c_constPcL	iso	no	250	yes		100	L	no	i
temp_loc	iso	no						yes	i