Reply to Referee Comments #1 and #2

We thank both referees for their time and insights. In the following, we address each of the referees' comments and, in most cases, provide an edited section of the paper.

Referee #1

General Comment

"The authors present a new reconstruction of temperature and precipitation over Greenland covering the past 20 000 years using for the first time over such a long period a data assimilation technique successfully applied recently over the past millennia. The paper is very clear, justify nearly all the choices in a very rigorous way and provides comprehensive estimates of the uncertainties. I have thus no doubt that, in addition to the new reconstruction that can be used for instance to drive ice sheet models, this study opens new fields of application of data assimilation of multi-millennial timescales.

However, I consider that the impact of the choice of the prior is not enough discussed and this issue must be addressed before publication. If I understand well, the prior ensemble is made of 100 states obtained by averaging 50 years of model data. Those states are selected randomly over the full length of the simulation (line 175). This method is reasonable if the climate variations are weak, such as during the past millennia, but is it valid for very large changes as observed during the glacial interglacial periods? I may have missed something but, if I am right, a state obtained in the model in the late Holocene can be used to reconstruct the last glacial climate, which may be hard to justify. For instance, the authors argue that it is important to take into account the changes in seasonality of precipitation (e.g. line 233) but I wonder how this could be achieved by selecting model states that are coming from very different periods. I would suggest using as prior only years that are close to the period that is reconstructed so that only glacial states are used to reconstruct glacial climate for instance."

Reply to the General Comment

The reviewer's general comment broadly concerns our selection of the prior ensemble from the entire TraCE-21ka simulation rather than from specific time periods that align with the reconstruction time. This is an excellent point, which we have thought about carefully, but had not elaborated upon in the paper. In reply, we will elaborate on this in the paper and the supplementary information, as follows below.

The reviewer states that, "If I understand well, the prior ensemble is made of 100 states obtained by averaging 50 years of model data. Those states are selected randomly over the full length of the simulation (line 175)." If we are understanding each other correctly, then yes, one state in the 100-member prior ensemble is an average over 50 years of the model data; these 100 states are selected randomly from the full length of the simulation. This implies that both glacial and Holocene states are likely to be contained within the same prior ensemble that is used to reconstruct all time steps over the last 20,000 years. To be clear, a prior ensemble could in principle contain only Holocene states; however, this is not the case for any of the ten prior ensembles we use in the paper. Thus, in reconstructing a time step in the glacial, for example, both glacial and Holocene states are part of the prior ensemble.

We agree with the reviewer that conditionally chosen prior ensembles would be preferable, but for the timescale under consideration this is not yet feasible. To explain the reasoning behind our choices in the paper, we elaborate on the pros and cons of four methods we considered before deciding on the one that we use in this study (#4).

1. For offline data assimilation (i.e., no information passed between assimilation time steps), a justifiable method for choosing the prior ensemble would be to use a 100-member ensemble of 20,000-year climate simulations. These climate simulations would be TraCE-21ka-like (i.e., results from fully-coupled GCMs at T31 resolution or higher), and have varied initial conditions, boundary conditions, and model physics. The prior ensemble for any assimilation time step would be taken from the

same time step in the climate simulations, which would lead to a prior ensemble that varies smoothly in time and is a justifiable initial guess for the climate evolution over the past 20,000 years. Though this option is simple, it is not feasible because the computational cost of running even one TraCE-21ka-like simulation remains near computational limits.

2. Given that there is only one TraCE-21ka-like simulation, another method would be to select states from TraCE-21ka that are closest in time to the reconstruction time step. For example, if we were reconstructing the 50-year average centered on the year 5,000 CE, then we would select the 100 states from TraCE-21ka that are closest in time to 5,000 CE. Given that we are working with 50-year averages, this means we would select all the states between 7,500 and 2,500 CE. This method, which we call the "running-window" method, provides a prior that varies smoothly in time and is a justifiable initial estimate for the climate evolution.

For the running-window method, the variance of the prior ensemble would tend to be small. A prior with small variance would lead to underweighting of the proxy records during assimilation. To avoid this issue, we could use the well-accepted approach of inflating the prior variance (Anderson and Anderson, 1999). However, the use of inflation adds an additional tunable parameter; in this case, it would add an additional parameter per time step. Although inflation can, in principle, be constrained using the ensemble calibration ratio (computed for excluded proxies), we have too few proxy records to meaningfully constrain this parameter without overfitting.

In addition to estimating numerous inflation factors, the running-window method limits us to one estimate of the spatial covariance structure per time step. Thus, we have no way to quantify the uncertainty associated with the prior covariance structure. This could be fixed by expanding the running window and randomly selecting multiple prior ensembles; however, if the running window is expanded enough to create meaningfully different prior ensembles, then Holocene states will leak into glacial prior ensembles (and *vice versa*) and the method essentially becomes the method we use in the paper.

3. To reduce the number of inflation parameters, we could split TraCE-21ka into several distinct time periods. From these time periods, we would randomly select prior ensembles that are only used for the reconstruction of associated assimilation time steps. For example, if we split TraCE-21ka into glacial, transitional, and Holocene periods, then we'd make a glacial prior ensemble that is only used to reconstruct the glacial, a transitional prior ensemble that is only used to reconstruct the transition, and a Holocene prior ensemble that is only used to reconstruct the Holocene. This reduces the number of inflation factors we must estimate to a total of three. A disadvantage, however, is that this makes the prior discontinuous in time, which frequently leads to a discontinuous reconstruction. To adjust the reconstruction and make it continuous requires another source of information. Such post-processing adds an extra layer of complexity.

4. The method used in our study ensures a continuous reconstruction and removes the need for inflation factors. This method uses the same prior ensemble for all time steps (thus it is continuous) and the includes both glacial and Holocene states, which provides enough variance to appropriately weight the proxy records (thus no inflation is needed). Though the time-invariant prior is a poor estimate of the climate evolution over the last 20,000 years, the proxy records are given enough weight to result in a posterior that captures the large climate changes. In addition, we can quantify the uncertainty associated with the spatial covariance pattern by producing multiple posterior ensembles that each stem from a different prior ensemble. In the paper, we use ten different prior ensembles to quantify this uncertainty. Overall, this method is both feasible and simple, thus providing a first step in developing paleoclimate data assimilation for applications on glacial-interglacial timescales.

As mentioned above, we will include a discussion of these points in the revised paper.

We would also like to specifically address the following comment: "the authors argue that it is important to take into account the changes in seasonality of precipitation (e.g. line 233) but I wonder how this could be achieved by selecting model states that are coming from very different periods." We agree with the reviewer that the best way to account for changes in precipitation seasonality is to do it in a time-varying manner, but we use a time-invariant prior for reasons given above. This means that the mean precipitation seasonality is constant in time and determined by the states in our prior ensemble. What is new in our

paper is that for each reconstruction, we use ten different prior ensembles, which gives us ten different estimates of the mean precipitation seasonality. In addition, our approach accounts for spatial variations in precipitation seasonality.

We do not wish to mislead readers into thinking that we account for time-varying precipitation seasonality. Instead, we account for a mean precipitation seasonality, which is determined by the states in our prior ensemble. To clarify this in the paper, we will edit lines 242-244 in the paper to say the following:

> With T^*_{site} in our PSM, we find that the δ^{18}O-T_{site} slope is spatially variable. Using our prior ensembles, which are selected randomly from the full TraCE-21ka simulation, the average slope around Greenland is about 75% of modern, or about 0.5 $‰^{\circ}C^{-1}$. Thus, in using this prior ensemble, we capture the mean of the modern (i.e. high-frequency) relationship and that of the glacial-interglacial (i.e. low-frequency) effective temporal slope. Due to our use of TraCE-21ka, this mean δ^{18}O-T_{site} relationship is relevant both for our reconstruction period (the last 20,000 years) and for our reconstruction method, which uses a mean spatial covariance pattern from TraCE-21ka for all time steps.

Specific Comment #1

"More specifically, still related to the prior, the authors explain (line 135) that 'For paleoclimate data assimilation, it is important that the climate simulation capture a range of possible climate states over the time period of interest.' They should thus first discuss the results of the TraCE-21ka simulation as it seems from Figure 12 that it underestimates the magnitude of the changes. More generally, the authors do not discuss at all the biases of the climate model. They correct for biases in the modern state by using anomalies compared to 1850-2000 (line 146) but this seems to be a small change compared to the signal during the whole simulation (line 366). Besides, the response to forcing is very different between different models as illustrated by the Paleoclimate Model Intercomparison Project. How this model behavior, which can also bias results for distant past, is influencing the results? Another way to phrase this point is that the model biases are not constant over time while the proposed correction assumes the stationarity of the biases."

Reply to Specific Comment #1

The reviewer makes a good point that we should expand our discussion of the TraCE-21ka simulation, especially with respect to model biases.

Model-bias corrections rely on observations. In the modern, there are numerous observations from ground-based and satellite systems. From the past, there are relatively few observations, which are from proxy records. Given an assumption of stationary model bias (unchanging in time), we can use the modern observations to compute the bias correction; however, given an assumption of non-stationary model bias, we must rely on paleoclimate proxy records as well. In our paper, we assume a stationary model bias and apply the delta-change method (Teutschbein and Seibert, 2012). This leaves the proxy records available for data assimilation. Ideally, we would subsample the proxy records and use one subsample for data assimilation, another to correct for model bias *a priori*, and another to assess the influence of model bias on our reconstruction *a posteriori*. This, however, is not possible with a small number of proxy records. Therefore, we have chosen to reserve all proxy records for data assimilation and to assume a stationary bias correction.

In addition to a mean bias in the model, there may also be biases in the variance. The reviewer specifically points out this issue with TraCE-21ka: it has a small glacial-Holocene climate change relative to the ice-core records. This is a bias that is best addressed with information from proxy records, which we have reserved as independent observations for data assimilation.

As the reviewer alluded to, another opportunity to assess the influence of model bias is to simply select our prior ensemble from a different model. By examining how the results are affected by a variety of different model simulations, we could assess the sensitivity of our results to different models (and thus model biases). Doing this analysis in a rigorous manner is not yet possible because TraCE-21ka is the only-available continuous 20,000-year simulation completed with a fully-couple GCM at

a T31 resolution or higher.

We will include this discussion as a paragraph in section 2.2 (lines 132-147):

> To correct for model bias, we assume it is stationary in time and apply the delta-change method (Teutschbein and Seibert, 2012) by taking the anomaly of temperature and the fraction of precipitation relative to the mean of our reference period (1850-2000 CE). Our assumption of a stationary model bias is required because, with a small number of proxy records, we cannot afford to subsample them for the purposes of bias correction, data assimilation, and evaluation.

Specific Comment #2

"Estimating the skill of the reconstruction compared to a constant prior (line 204) is a too low target for me. If the reconstruction was only showing a warming between the glacial period and the Holocene, it would already be skillful compared to this initial estimate and this does not require a very sophisticated technique. The skill of the reconstruction should be evaluated against the transient TraCE-21ka simulation to see if the data assimilation brings some skill compared to the simulation not constrained by data."

Reply to Specific Comment #2

We agree with the reviewer that we should compare our reconstruction skill against that of other 20,000-year reconstructions or simulations. The skill metrics are most comparable if there is either a δ^{18}O or T* variable available in the reconstruction or simulation. Thus, this comparison is straightforward with TraCE-21ka, but not for other reconstructions like that of Buizert et al. (2018). We have computed the correlation coefficient, coefficient of efficiency (CE), and root mean square error (RMSE) for TraCE-21ka, and will add this information both as text and figures.

We will add the following text to line 209 (the end of a paragraph) to explain our evaluation against TraCE-21ka:

> We additionally compute the correlation coefficient, CE, and RMSE on the TraCE-21ka simulation as compared to the proxy records.

We will also edit section 4.1, which discusses the results of the evaluation. Please see our reply to specific comment #6 for the edited text (lines 284 to 340 of this document) and additional figures.

Specific Comment #3

"The authors explain at the end of the conclusion (line 485) that using a model that directly simulates isotopes would likely improve their results. It would be interesting to discuss that earlier because, for instance, they mention a different relationship between reconstructed precipitation and temperature at different time scales (line 352) but what is the potential role of a different relationship between temperature and δ^{18}O on this conclusion?"

Reply to Specific Comment #3

We agree with the reviewer that the isotope-temperature relationship is an interesting one to explore, a topic which we touch on with our sensitivity experiments described in section 2.3.1 of the paper (lines 245-255). We tested the sensitivity of our results to different isotope-temperature relationships, primarily the magnitude of the slope in a linear relationship, but also the spatial pattern of that slope. The reconstructions resulting from these experiments are discussed in section 4.2 and shown in Fig. 10.

An isotope-enabled model, as we mention in the conclusion, would provide another estimate of the magnitude and spatial pattern of the isotope-temperature relationship. The advantage of an isotope-enabled model is that it incorporates the variety of processes that can affect water isotope ratios, whereas, with our experiments, we focus on one primary process: precipi-

tation seasonality. Isotope-enabled models may have biases, and the experiments we perform are important for assessing the sensitivity of reconstructions to the assumed isotope-temperature relationships. As the reviewer suggests, we will bring up this discussion earlier in the paper by adding the following sentences to the last paragraph in section 2.3.1 (lines 245-255):

> These sensitivity tests are equivalent to testing different assumptions about the δ^{18}O-temperature relationship. The availability of a 20,000 year-long isotope-enabled climate simulation would allow us to determine this relationship from model physics, which incorporate a variety of processes that can affect water isotopes, including precipitation seasonality.

The reviewer also astutely suggests that we test the effect of the isotope-temperature relationship on the precipitation-temperature relationship. We have analyzed the temperature-precipitation relationship between all possible combinations of our main reanalysis and sensitivity reanalyses and will add the following paragraph to the paper at the end of section 4.2 (line 413). Table S1 is included at the end of this document.

> Finally, we test how these sensitivity results affect the scaling factor (β) in the precipitation-temperature relationship. We pair the five temperature reconstructions (main, S1-S4) and three precipitation reconstructions (low, moderate, and high) into fifteen possible combinations and conduct the same analysis as described above in Sect. 3. Across these fifteen combinations, we find that the spatial pattern of β is robust (Fig. 6a). The exact magnitude depends primarily on the temperature reconstruction and how cold it is in the glacial, with colder temperatures giving lower β values. To a lesser degree, the magnitude also depends on the precipitation reconstruction, with wetter scenarios giving lower β values. As previously, we find that the low-pass filtered datasets have the same or nearly the same β value as the unfiltered dataset, while the high-pass filtered datasets have consistently lower β values. As an example of this consistency, Table S1 shows the β value found for the Kangerlussuaq region for all fifteen combinations and three filtering options.

Specific Comment #4

"If I am right, when the technique described in section 2.3 is applied for the past millennia, records related to both the temperature and hydrology are assimilated together, as the covariance between the variables can bring interesting information and reduces the uncertainties. Here, it is claimed that having independent temperature and precipitation reconstructions is an advantage. This also means that precipitation and temperature changes could not be dynamically consistent in the proposed reconstruction? Maybe the authors do not want to rely on the covariance between those two variables as simulated by the climate model but they should explain why and, in that case, explain in a bit more detail the added value brought by the assimilation using this model results as prior."

Reply to Specific Comment #4

The reviewer is correct that previous work (e.g., Hakim et al., 2016; Tardif et al., 2019) has used a similar method to assimilate both temperature and precipitation-sensitive proxy records into a reconstruction of multiple climate variables over the last millennium. We agree with the reviewer that our choice to separate the proxy records and reconstruct temperature and precipitation independently requires more explanation in the paper.

We have chosen to independently reconstruct temperature and precipitation because the relationship between these two variables is highly non-linear and non-stationary over the last 20,000 years. Precipitation generally follows the expected thermodynamic relationship on glacial-interglacial timescales (Robin, 1977); however, there have been times in the last 20,000 years, as shown by ice-core records, that precipitation has deviated significantly (even having opposite sign) from the thermodynamic expectation (Cuffey and Clow, 1997). For this reason, we choose not to impose the climate-model-derived mean temperature-precipitation relationship on our reconstruction.

We would also like to address the reviewer's concern that the temperature and precipitation may not be dynamically consistent given our method of reconstructing them independently. The reviewer is correct that our reconstructions may not achieve dynamic consistency through the modeled relationships. However, the reconstructions are dynamically consistent insofar as the empirical δ^{18}O and accumulation records from the ice-core records are dynamically consistent, as they must be.

We will make edits to the paragraph at lines 175-183 to reflect this discussion. Note that some edits to this section are in reply to the general comment #2 from Referee #2 (lines 409 to 410 of this document). New text is in *italics*.

> The prior ensemble is an initial estimate of possible climate states, which we form using 100 randomly-chosen 50-year averages from the TraCE-21ka simulation. *States from both the glacial and the Holocene make up a prior ensemble. The same prior is used for all time steps in the reconstruction, leading to a prior that is constant in time.* Proxy records are assimilated into the prior using Eq. 1, which produces the posterior ensemble, a new estimate of possible climate states. We assimilate δ^{18}O to reconstruct temperature and separately assimilate accumulation to reconstruct precipitation. *This approach maintains independence between temperature and precipitation, which avoids imposing linearity and stationarity on the relationship between these two variables. As Cuffey and Clow (1997) show, not only is this relationship non-linear on long timescales but it is also not well-approximated by simple thermodynamic expectations. Separating these variables ensures that the relationship between temperature and precipitation is consistent with the empirical relationship between $\delta^{18}O$ and accumulation from ice cores, rather than being derived from the climate model.*
>
> *We repeat the data assimilation process* over multiple iterations, with each iteration using one of ten different 100-member prior ensembles and excluding one proxy record. *Each of the ten prior ensembles is made up of a different random selection of 50-year averages from TraCE-21ka. Thus, each prior ensemble has a different variance and spatial covariance structure.* Each proxy record is excluded from a total of ten iterations, where each of these iterations uses a different one of the ten prior options. *Every iteration is uniquely identifiable by which prior ensemble is used and which proxy record is excluded.* For a reanalysis, the total number of iterations is thus ten times the number of proxy records, such that for temperature we have 80 iterations and for precipitation we have 50 iterations. A reanalysis is a compilation of the *100-member* posterior ensembles from these iterations, resulting in a temperature reanalysis having 8,000 ensemble members and a precipitation reanalysis having 5,000 ensemble members.

Specific Comment #5

"Line 153, it is said that 'The offline method is appropriate when characteristic memory in the system is significantly shorter than the time step' (here 50 years). Is this valid here, for Bølling-Allerød and Younger Dryas events for instance?"

Reply to Specific Comment #5

The reviewer raises an excellent point that periods with strong forcing, such as the Bølling-Allerød and Younger Dryas, have a longer characteristic memory than periods with weaker forcing, such as the late Holocene. In general, models have little predictive skill on decadal or longer timescales, except perhaps for areas strongly influenced by the thermohaline circulation (Latif and Keenlyside, 2011, and references therein). During periods of strong forcing, model predictive skill may increase if both the forcing and the response are appropriately represented by the model; however, the predictive skill may also decrease if model uncertainty is large (Hawkins and Sutton, 2009). Rather than assuming our method is appropriate for all times in the last 20,000 years, we would ideally make use of an ensemble of long climate simulations or online data assimilation; however, both alternative approaches are not feasible at this time due to computational cost.

To clarify this point in the paper, we will edit the first paragraph of section 2.3 (lines 149 to 153) to say the following (new text is in *italics*):

To combine the ice-core data and climate-model data, we use an offline data assimilation method similar to that described in Hakim et al. (2016). "Offline" refers to the absence of a forecast model that evolves the climate state between *assimilation* time steps, such that in offline data assimilation the same initial climate state is used for every time step. The offline method is appropriate when model predictive-skill is significantly shorter than the *assimilation* time step (Hakim et al., 2016, and references therein). *Model predictive-skill is generally poor on decadal to longer timescales (Latif and Keenlyside, 2011, and references therein) except possibly during times of strong forcing, such as the Bølling-Allerød (14.7-12.7 ka) and the Younger Dryas (12.7-11.7 ka) (Hawkins and Sutton, 2009). Because each of our time steps is an average over 50 years, as dictated by the resolution to which we average the proxy records, the offline method is appropriate except possibly during these large-forcing events. For these events, an online method would be most appropriate (assuming that the models correctly capture both the forcing and the response); however, online data assimilation over glacial-interglacial cycles is not feasible at this time given the computational cost.*

Specific Comment #6

"Line 375. The reanalysis skill over the full period is clear compared to a constant climate but this would be informative to quantify it more precisely for the two selected 5000-year periods. Stating that it is lower than for the full period is not enough I think. From the figures 7 and 8, it seems that the CE is negative for nearly all the points. Stating line 377 that 'the reanalysis shows overall improvement over the prior ensemble' is also a weak conclusion as discussed in point 2."

Reply to Specific Comment #6

We agree with the reviewer that the evaluation over the two 5,000-year periods warrants more discussion. In section 4.1, we will insert this discussion as well as the comparison of our reconstruction skill to that of TraCE-21ka (in reply to specific comment #2, lines 146 to 165). New text is in *italics* and new figures, S5-S8, are at the end of this document.

4.1 Independent proxy evaluation

Here we evaluate our results against proxy records that are excluded from ten of the iterations that make up the temperature and precipitation reanalyses. For this evaluation, we use the raw results (without a mean-bias correction). We find, however, that the mean biases are small relative to climate changes over the last 20,000 years; there is little difference between our bias-corrected and uncorrected results and it is unlikely that the mean bias has a large effect on our evaluation.

Evaluation against independent proxy data shows that our reanalysis captures the timing and magnitude of low-frequency climate changes (Figs. 7 and 8) and is an improvement over both the prior ensemble *and TraCE-21ka* (Figs. S3-S4 *and S7-S8*). Evaluation over the full 20,000 years of the temperature and precipitation results shows high, positive correlation coefficients (ranging from a minimum of 0.97 to maximum of 0.99), which indicate that the reanalysis captures both the timing and sign of climate events, while high CE (0.87-0.98) and low RMSE values (0.62-1.2 ‰for δ^{18}O and 0.04-0.08 for accumulation) indicate that the reanalysis captures the magnitude of these events. Our skill during this longest evaluation period is primarily due to the presence of low-frequency climate changes, which tend to be coherent across Greenland, such that evaluation over this full 20,000-year period shows more skill than evaluation over the full Holocene, which shows more skill than evaluation over just 5,000 years in the Holocene (or 5,000 years in the glacial) (Figs. 7 and 8).

Our posterior ensemble consistently shows improvement over the uninformed, constant prior ensemble during the 20,000-year evaluation period (Figs. S3 and S4). TraCE-21ka provides a better comparison, because the model simulation is also uninformed by the ice core data, but it is transient and generally captures glacial to Holocene changes. Over the 20,000-year evaluation period, relative to the reconstructions, we find that TraCE-21ka has consistently lower correlation coefficients (0.86-0.96), lower CE values (0.50-0.86), and higher RMSE values (1.9-2.8 ‰for δ^{18}O and 0.11-0.15

for accumulation) (Figs. S5-S8). This comparison suggests that our reconstruction captures the timing and magnitude of the glacial to Holocene transition better than TraCE-21ka.

Even for the shorter evaluation periods, which are dominated by high-frequency, spatially-incoherent noise, the reanalysis shows improvement over both the prior ensemble and TraCE-21ka (Figs. S3-S4 and S7-S8); however, the improvement is not as consistent as for the 20,000-year evaluation period. For our reconstruction, correlation is positive except for three locations in the temperature reconstruction, with Holocene precipitation showing the largest correlation values (up to 0.60) and the total range being -0.30 to 0.60. The prior correlation is zero for these shorter evaluation periods and locations; however, TraCE-21ka shows correlation values ranging from -0.29 to 0.69, with eight negative correlations (more than we find for our reconstruction), but generally higher correlations in the Holocene than our reconstruction.

For the shorter evaluation periods, the reconstruction CE is generally negative (ranging from -82 to 0.17) with a few exceptions; however, the reconstruction may still be skillful (e.g., Cook et al., 1999). The skill of the reconstruction is better measured by the difference between prior and posterior CE due to the strong influence that bias can have on CE (Hakim et al., 2016). For this difference metric, we find consistent improvement of the posterior CE over that of the prior (ranging from an increase of 4.7 to 3200) and over that of TraCE-21ke (ranging from an decrease of -6.9 to an increase of 230). RMSE is the most consistent of the skill metrics for these shorter evaluation periods, with our reconstruction showing improvement over both the prior and TraCE-21ka, the one exception being that TraCE-21ka has greater skill at the Agassiz ice-core site in the Holocene. For the reconstruction, RMSE values range from 0.24 to 1.8 ‰^{2} for temperature and 0.025 to 0.10 for precipitation.

For all evaluation periods and both variables, the ensemble calibration ratio (ECR) for the prior is skewed towards values greater than unity (0.66-8.7), which suggests that the prior ensemble tends to have too little spread. Conversely, for the posterior, the *ECR* is generally less than unity (0.10-1.7) (Figs. 7 and 8), suggesting that the posterior ensemble *has more spread than* the error in the ensemble mean (as compared to the proxy records). This result indicates that the reanalysis ensemble encompasses the climate as recorded by the proxy records for most times and locations over the last 20,000 years.

Referee #2

"The authors present new reconstructions of temperature and precipitation over the last 20,000 years over Greenland. For this, they apply a data assimilation technique on δ^{18}O and accumulation records from Greenland ice cores and use the temperature and precipitation outputs from the TraCE-21ka simulation to extend the information to all the continent. The paper is in general clear enough for that people not having skills in data assimilation can read and understand quite easily the methodology presented in this manuscript. In my knowledge, the technique presented here is innovative for such a long period, and the different assumptions are presented and tested in a very rigorous way. This manuscript is worthy for publication in Climate of the Past, after having considered the comments below."

General Comment #1

"Compared to the ice cores part, which is well discussed in the Methods section and in the Supplementary Material, the results of the TraCE-21ka simulation are not discussed enough in my opinion. More details about the similarities/differences with other PMIP simulations and/or climate reconstructions for PI and LGM could be discussed for example. How do the last 1000/2000 years fit well with last millennium simulations or reconstructions from isotopic proxies? Moreover, the rapid climate transitions are not so well captured by the TraCE-21ka, especially the Younger Dryas. This point should be discussed in terms of potential consequences for the reconstructions by data assimilation. The spatial resolution (T31 and 26 atmospheric vertical levels for the atmosphere if I am not wrong) should be clearly stated, and the uncertainties related to this aspect could be discussed (even if

it is mentioned later). For example, is the limited number of grid points over Greenland a problem for the paleo DA technique? The ice sheet boundary conditions are also of major importance in this type of simulation. The expected differences if a more recent ice-sheet reconstruction would be prescribed could be discussed. Last point: what about the precipitation seasonality in TraCE-21ka over Greenland? Is it consistent with observations? Is the seasonality different for Holocene and glacial periods? I guess it could have an important impact on the δ18O PSM and the final temperature reconstruction. . . "

Reply to General Comment #1

We agree with the referee that our discussion of the TraCE-21ka simulation is limited compared to our discussion of the ice-core records. This is primarily because TraCE-21ka is described extensively in the literature (e.g., Liu et al., 2009, 2012; He, 2011; He et al., 2013), whereas a number of aspects of the ice-core network, particularly the accumulation records, are novel or have been discussed little in previous work. In section 2.2 of the paper, we do discuss attributes of TraCE-21ka that are especially relevant to our method, including, the glacial to Holocene mean state changes, temperature and precipitation seasonality, and the ice-sheet boundary conditions. We agree with the referee that we are missing a statement about the spatial resolution of TraCE-21ka (indeed it is T31) and more details concerning seasonality and the ice sheets. We will include these details in the revised paper.

We want to emphasize that our paper represents the first attempt to use the ensemble Kalman filter approach to reconstruct climate over glacial-interglacial timescales. Ideally, there would be more 20,000-year (or longer) TraCE-21ka-like simulations (i.e., from fully-coupled GCMs at T31 resolution or higher). With other simulations from different models, we could address the influence of model bias, spatial resolution, boundary conditions, and initial conditions (for a discussion of model bias, see our reply to Referee #1's specific comment #1, lines 116 to 145 of this document). If we had other simulations, then we agree with the referee that it would warrant a comparison between simulations and a discussion of how the differences affect our results. With only one TraCE-21ka-like simulation available, we cannot conduct a meaningful comparison with how other climate-model simulations would affect our results. Our focus instead is on demonstrating that the method is feasible and skillful. Thus, we do not see it as productive to elaborate on how TraCE-21ka compares with PMIP or other simulations. In addition, the previous literature has extensively interrogated the results of the TraCE-21ka simulation (e.g., He, 2011; Buizert et al., 2014; Pedro et al., 2016; Zhang et al., 2017, 2018; Marsicek et al., 2018). We will add new text to refer the reader to this previous work.

The referee additionally suggests that we include a discussion of how well TraCE-21ka captures rapid climate transitions, such as the Younger Dryas. We agree that this would be necessary if we were to choose our prior ensemble exclusively from time periods that are close to our reconstruction time (for examples of this, see our reply to Referee #1's general comment, lines 21 to 105 of this document); however, our prior ensemble is time-invariant and thus our method is insensitive to the temporal evolution of TraCE-21ka (as long as the variance and spatial covariance structures of TraCE-21ka are preserved).

General Comment #2

"The way how the prior ensemble is made should be clarified. To avoid misunderstanding, the authors should state at line 175 that the prior is constant in time (and not later at line 206). If I understand clearly, 10 different 100-member prior ensembles are made (it should be said directly at the beginning). To form a prior ensemble, do you then randomly pick up 100 snapshots from the resampled TraCE-21ka temperature and precipitation outputs (see my minor comment for the line 146)? Or do you take randomly from the yearly TraCE-21ka outputs 50 consecutive years of data that you average in time for a member, other 50 consecutive years of model outputs for another time period for the second member, and so on. . .? Or another way? Anyway, it needs clarification (in the way of the section 2.3 of Hakim et al. 2016 for example). I have the same remark as the first referee: what would be the difference if you would use, for instance, a "glacial prior" to reconstruct climate variables from glacial period and a "Holocene prior" for the warmer period instead of a constant prior for all the 20,000 years? In link with my first major remark, what would be the impact on the seasonality of precipitation, that influences the reconstructions of the authors?"

Reply to General Comment #2

We thank the referee for calling our attention to these points of confusion. We will clearly state earlier in the paper that the prior is constant in time. We will also clarify how we average TraCE-21ka in line 146 of the paper by making the following edits (new text is in *italics*):

> We then average to a 50-year resolution, as for the ice-core records. *This averaging results in 401 time steps spaced 50 years apart.*

We will also reorganize and edit the paragraph starting at line 175 to clarify how we form our ten different 100-member prior ensembles. Please see lines 232 to 251 in this document for the edits.

The referee additionally raises concerns over our random selection of the prior ensemble from the full TraCE-21ka simulation and the effect this has on the seasonality of precipitation and our reconstructions. For this we refer to our reply to the general comments from Referee #1 (lines 21 to 105 in this document).

Minor Comments

We thank the referee for these specific edits and the effort to make the paper clearer and more concise.

Line 14: "requires understanding its sensitivity to changes. . ."
Thank you. This edit will be made in the revised paper.

Line 18: I would put into brackets the terms "and arid" and "and wet".
We think that it is important to equally emphasize the thermal and hydrologic differences between glacial and interglacial periods. For this reason, we'd rather not make the suggested revision.

Line 31: the spatial resolution, especially for paleoclimate simulations, brings also uncertainties.
We will edit this to say: "In contrast, climate-model simulations are spatially-complete estimates of past climate, but they are subject to uncertainty due to model dynamics, boundary conditions, and spatial resolution."

Line 37 and passim: I think this is TraCE-21ka and not TraCE21ka.
We will make this change throughout the paper.

Paragraph lines 106-116: Does the matching of δ^{18}O from Dye3 to the δ^{18}O record from NGRIP bring a dependency when evaluating the posterior against Dye3 δ^{18}O record?
While dating uncertainty is non-zero, it is clear from multiple independent lines of evidence that the major δ^{18}O variations in all Greenland cores are nearly synchronous. Most important is the variance, rather than the timing, of δ^{18}O, and this is not affected by the minor changes to the timescales imposed here. (Note that our matching of the δ^{18}O from Dye3 to the δ^{18}O record from NGRIP is similar to the methods used to create the GICC05 depth-age scale and apply it to other ice cores.)

Section 2.2: I understand when you use the term "transient ice-sheet" that the prescribed ice-sheet is changed over time. But some people can misunderstand and think that it is done dynamically with a coupled ice-sheet model. I would use an expression like "prescribed transient ice-sheet boundary conditions" for example.
We will change the sentence over lines 141-142 to say: "TraCE-21ka also includes prescribed transient ice-sheets as a boundary condition, the transient nature of which is important for capturing the influence of elevation change on the ice-core records."

Line 146: What is the initial temporal resolution of TraCE-21ka outputs? Monthly mean? When you talk about "average of 50-year resolution", do you mean "resampling" every 50 model years? If you take the last 20,000 years, it

makes something like 400 time steps, right?
We thank the referee for calling our attention to this point of confusion. We will edit section 2.2 to clarify the initial temporal resolution of TraCE-21ka (which is monthly) and what we mean by averaging it to a 50-year resolution. The referee is correct that we end up with about 400 time steps after averaging. Our method is to take 50 consecutive years (600 months) of TraCE-21ka and to average them. No year (or month) is used in more than one 50-year average.

Lines 230-232: Other model studies like Gierz et al. 2017 (JAMES) for the LIG and Cauquoin et al. 2019 (CP) for 6k-PI climates have shown that the seasonality of precipitation affects the δ^{18}O-temperature relationship over Greenland.
We thank the referee for this comment. We did not mean to imply that Werner et al. (2000) is the only modeling study that has shown that precipitation seasonality affects the δ^{18}O-temperature relationship in Greenland. We will add the recommended citations.

Line 286: "that the proxy y and prior estimate of the proxy H(xb)".
Thank you. This edit will be made in the revised paper.

Lines 311-316: What does it give compared to the TraCE-21ka results?
We agree with the referee that it's important to compare our results to that of TraCE-21ka; however, we have made a point of saving comparisons with TraCE-21ka and Buizert et al. (2018) for the discussion (section 5).

Line 326: "from nearly +2 °C in northern. . ."
Thank you. This edit will be made in the revised paper.

Line 367: "has a large effect on our evaluation."
Thank you. This edit will be made in the revised paper.

Line 380: "the ECR is. . ."
Thank you. This edit will be made in the revised paper.

Line 404: the slower warming trends are hard to see. Make a zoom in the figure or give numbers.
Thank you for the suggestion. We have edited Fig. 10 to show a zoom-in on the Younger Dryas to Holocene transition. Please see the edited version at the end of this document.

Line 428: For S4 and high P cases, say clearly that it refers to the "sensitivity" curves on figure 12.
Thank you, we will clarify that "sensitivity" in the Fig. 12 legend label refers to S4 and high P.

Line 486: you can add the reference Okazaki and Yoshimura 2017 (CP).
We agree that this is a relevant citation and will add it to the revised paper.

Line 488: Add the references Cauquoin et al. 2019 (CP) and Okazaki and Yoshimura 2019 (JGR Atmos).
We agree that these are relevant citations and will add them to the revised paper.

Figure 4: add maybe contours for more clarity. And change the scale for the precipitation fraction at the peak warmth in the Holocene (panel b).
Thank you for the suggestion. We have added contours for clarification and changed the color scale in panel (b). Please see the edited version at the end of this document.

Figures 7, 8, S3, and S4: quite normal that the correlation is improved for the full period compared to the constant prior climate state.
Yes, we agree and have stated this in the paper (lines 205-209). If the referee thinks it is necessary, we can state it again later

in the text or in the figure captions. We will additionally be comparing the skill of our results to TraCE-21ka, as suggested by Referee #1.

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, 1999.

Buizert, C., Gkinis, V., Severinghaus, J. P., He, F., Lecavalier, B. S., Kindler, P., Leuenberger, M., Carlson, A. E., Vinther, B., Masson-Delmotte, V., et al.: Greenland temperature response to climate forcing during the last deglaciation, Science, 345, 1177–1180, 2014.

Buizert, C., Keisling, B., Box, J., He, F., Carlson, A., Sinclair, G., and DeConto, R.: Greenland-Wide Seasonal Temperatures During the Last Deglaciation, Geophysical Research Letters, 45, 1905–1914, 2018.

Cook, E. R., Meko, D. M., Stahle, D. W., and Cleaveland, M. K.: Drought reconstructions for the continental United States, Journal of Climate, 12, 1145–1162, 1999.

Cuffey, K. M. and Clow, G. D.: Temperature, accumulation, and ice sheet elevation in central Greenland through the last deglacial transition, Journal of Geophysical Research: Oceans, 102, 26 383–26 396, 1997.

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, 2016.

Hawkins, E. and Sutton, R.: The potential to narrow uncertainty in regional climate predictions, Bulletin of the American Meteorological Society, 90, 1095–1108, 2009.

He, F.: Simulating transient climate evolution of the last deglaciation with CCSM 3, vol. 72, 2011.

He, F., Shakun, J. D., Clark, P. U., Carlson, A. E., Liu, Z., Otto-Bliesner, B. L., and Kutzbach, J. E.: Northern Hemisphere forcing of Southern Hemisphere climate during the last deglaciation, Nature, 494, 81, 2013.

Latif, M. and Keenlyside, N. S.: A perspective on decadal climate variability and predictability, Deep Sea Research Part II: Topical Studies in Oceanography, 58, 1880–1894, 2011.

Liu, Z., Otto-Bliesner, B., He, F., Brady, E., Tomas, R., Clark, P., Carlson, A., Lynch-Stieglitz, J., Curry, W., Brook, E., et al.: Transient simulation of last deglaciation with a new mechanism for Bølling-Allerød warming, Science, 325, 310–314, 2009.

Liu, Z., Carlson, A. E., He, F., Brady, E. C., Otto-Bliesner, B. L., Briegleb, B. P., Wehrenberg, M., Clark, P. U., Wu, S., Cheng, J., et al.: Younger Dryas cooling and the Greenland climate response to CO2, Proceedings of the National Academy of Sciences, 109, 11 101–11 104, 2012.

Marsicek, J., Shuman, B. N., Bartlein, P. J., Shafer, S. L., and Brewer, S.: Reconciling divergent trends and millennial variations in Holocene temperatures, Nature, 554, 92–96, 2018.

Pedro, J. B., Bostock, H. C., Bitz, C. M., He, F., Vandergoes, M. J., Steig, E. J., Chase, B. M., Krause, C. E., Rasmussen, S. O., Markle, B. R., et al.: The spatial extent and dynamics of the Antarctic Cold Reversal, Nature Geoscience, 9, 51–55, 2016.

Robin, G. d. Q.: Ice cores and climatic change, Philosophical Transactions of the Royal Society of London. B, Biological Sciences, 280, 143–168, 1977.

Tardif, R., Hakim, G. J., Perkins, W. A., Horlick, K. A., Erb, M. P., Emile-Geay, J., Anderson, D. M., Steig, E. J., and Noone, D.: Last millennium reanalysis with an expanded proxy database and seasonal proxy modeling, Climate of the Past, 15, 1251–1273, 2019.

Teutschbein, C. and Seibert, J.: Bias correction of regional climate model simulations for hydrological climate-change impact studies: Review and evaluation of different methods, Journal of hydrology, 456, 12–29, 2012.

Zhang, Y., Renssen, H., Seppä, H., and Valdes, P. J.: Holocene temperature evolution in the Northern Hemisphere high latitudes–Model-data comparisons, Quaternary Science Reviews, 173, 101–113, 2017.

Zhang, Y., Renssen, H., Seppä, H., and Valdes, P. J.: Holocene temperature trends in the extratropical Northern Hemisphere based on inter-model comparisons, Journal of Quaternary Science, 33, 464–476, 2018.

Figures

Table S1. Scaling factors (β) for the temperature-precipitation relationship in the Kangerlussuaq region. The results for the main reanalysis are in bold.

Temperature scenario	**Precipitation scenario**	**No Filtering**	**Low-Pass (5,000 years^{-1})**	**High-Pass (5,000 years^{-1})**
Main	Low	0.09	0.09	0.04
Main	**Moderate**	**0.08**	**0.08**	**0.04**
Main	High	0.08	0.08	0.05
S1	Low	0.12	0.11	0.05
S1	Moderate	0.11	0.11	0.06
S1	High	0.11	0.11	0.07
S2	Low	0.09	0.09	0.04
S2	Moderate	0.09	0.08	0.04
S2	High	0.09	0.08	0.05
S3	Low	0.06	0.06	0.03
S3	Moderate	0.06	0.06	0.03
S3	High	0.06	0.06	0.04
S4	Low	0.07	0.07	0.03
S4	Moderate	0.07	0.07	0.03
S4	High	0.07	0.07	0.04

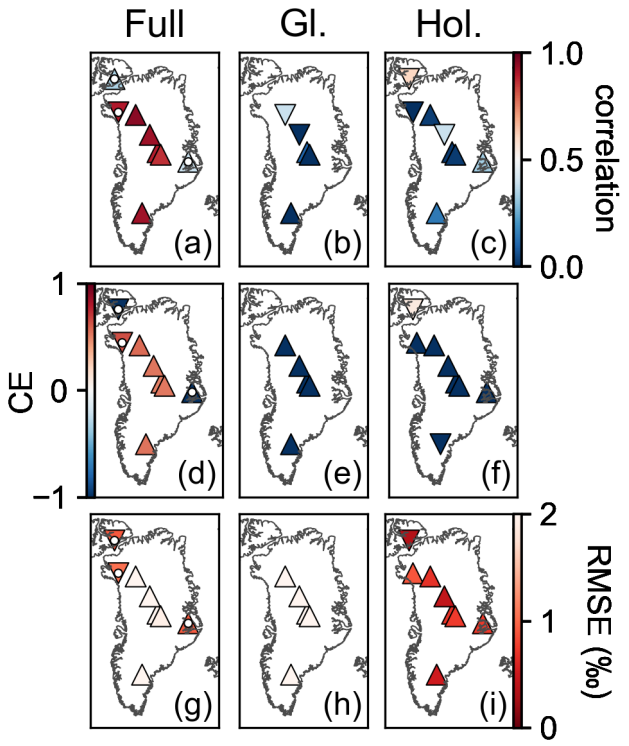

Figure S5. Temperature skill metrics for TraCE-21ka simulation. The first column (panels (a), (d), and (g)) shows the skill metrics for the full overlap (Full) between the proxy record and reanalysis. A white dot indicates evaluation against proxy records that overlap only the Holocene (11.7-0 ka). The middle column (panels (b), (e), and (h)) shows the skill metrics for a period in the glacial (Gl.) (20-15 ka), while the right column (panels (c), (f), and (i)) is for a period in the Holocene (Hol.) (8-3 ka). The first row (panels (a)-(c)) reports the correlation coefficient, the second row (panels (d)-(f)) the coefficient of efficiency (CE), and the third (panels (g)-(i)) the root mean square error (RMSE). Triangle symbols pointing up indicate that the posterior ensemble evaluates better than TraCE-21ka for that location and statistic. Triangle symbols pointing down indicate the opposite. A result is considered improved where the correlation coefficient closer to 1, CE closer to 1, and RMSE closer to 0.

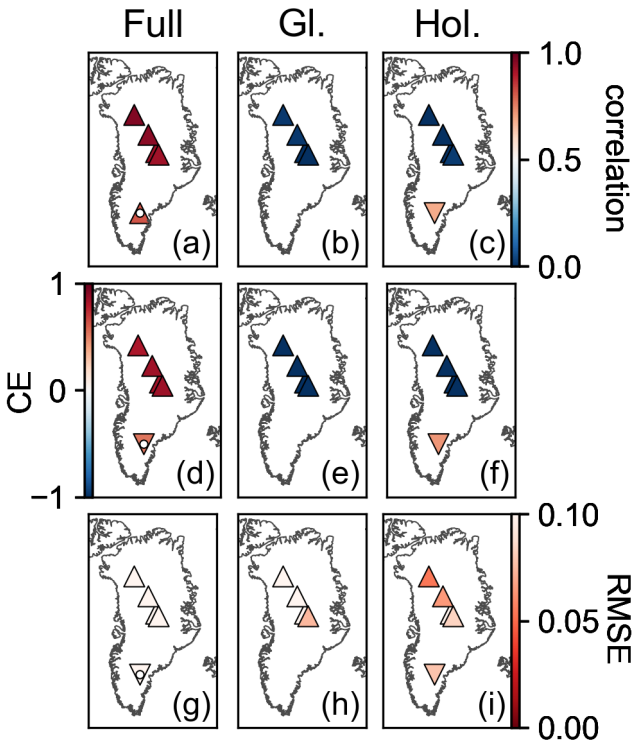

Figure S6. Precipitation skill metrics for TraCE-21ka simulation. The first column (panels (a), (d), and (g)) shows the skill metrics for the full overlap (Full) between the proxy record and reanalysis. A white dot indicates evaluation against proxy records that overlap only the Holocene (11.7-0 ka). The middle column (panels (b), (e), and (h)) shows the skill metrics for a period in the glacial (Gl.) (20-15 ka), while the right column (panels (c), (f), and (i)) is for a period in the Holocene (Hol.) (8-3 ka). The first row (panels (a)-(c)) reports the correlation coefficient, the second row (panels (d)-(f)) the coefficient of efficiency (CE), and the third (panels (g)-(i)) the root mean square error (RMSE). Triangle symbols pointing up indicate that the posterior ensemble evaluates better than TraCE-21ka for that location and statistic. Triangle symbols pointing down indicate the opposite. A result is considered improved where the correlation coefficient closer to 1, CE closer to 1, and RMSE closer to 0.

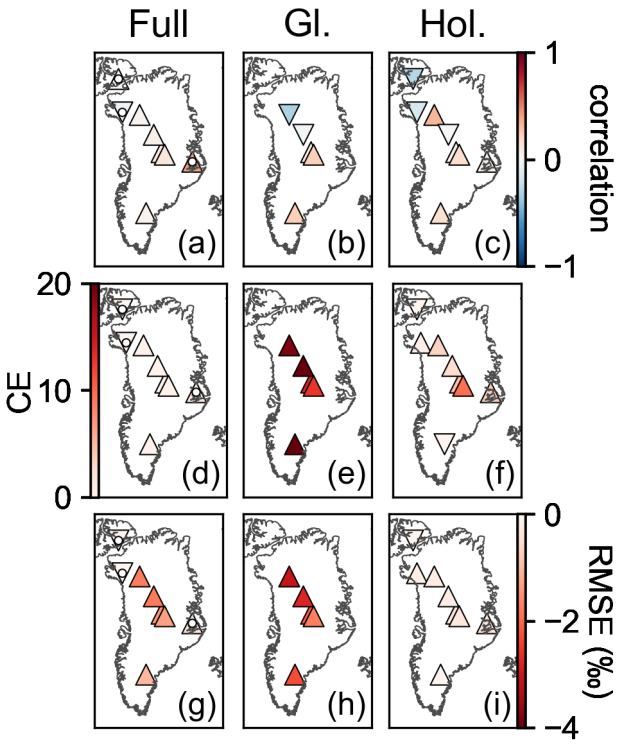

Figure S7. Difference in skill metrics, for temperature, between the posterior ensemble (averaged over iterations and time) and the TraCE-21ka simulation (difference = posterior – TraCE-21ka). Description of individual panels as in Figure S6.

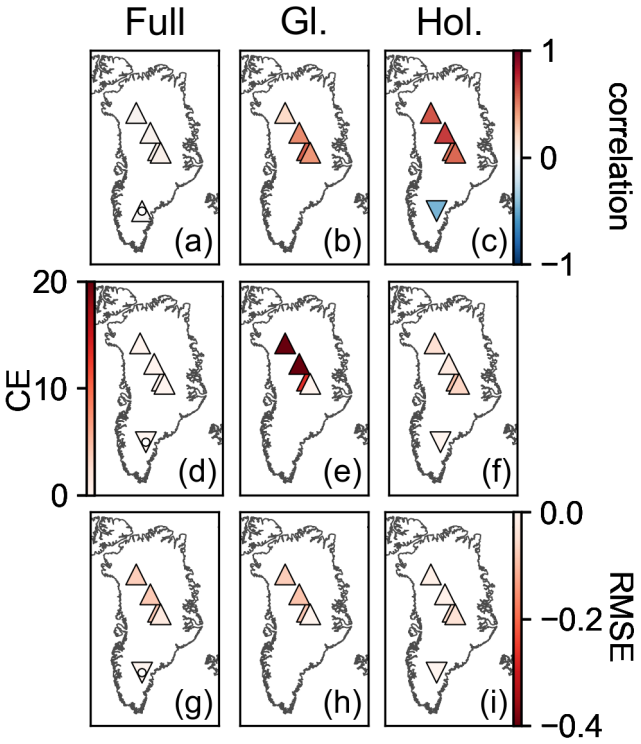

Figure S8. Difference in skill metrics, for precipitation, between the posterior ensemble (averaged over iterations and time) and the TraCE-21ka simulation (difference = posterior – TraCE-21ka). Description of individual panels as in Figure S7.

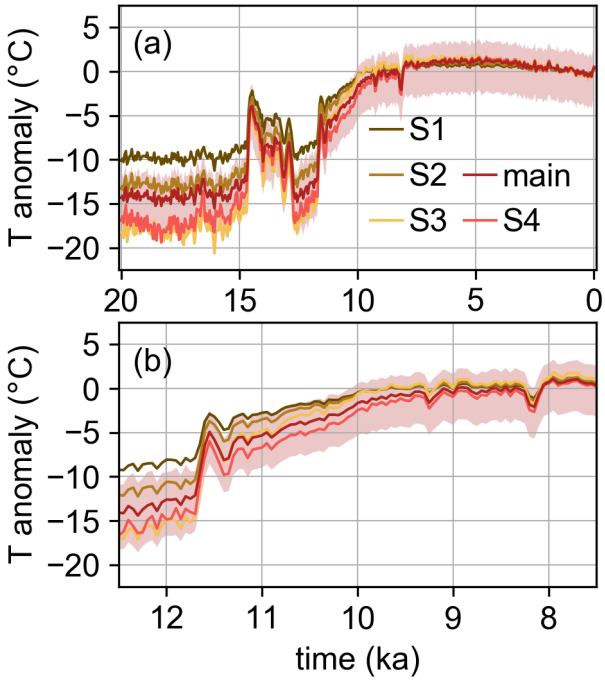

Figure 10. The main temperature (T) reanalysis (ensemble mean and 5^{th} to 95^{th} percentile shading) and ensemble mean for four sensitivity scenarios, S1-S4. Each sensitivity scenario reflects a different assumption about precipitation seasonality, with S1-S3 assuming a spatially-uniform seasonality and S3-S4 assuming stronger seasonality than the main reanalysis. Anomalies are with respect to the mean of 1850-2000 CE. These time series are for the location closest to Summit, which is representative of the results around Greenland. Panel (a) shows the full time period, and panel (b) the Younger-Dryas through the Holocene, showing that S1, S2, and S3 warm more quickly than the main reconstruction and S4.

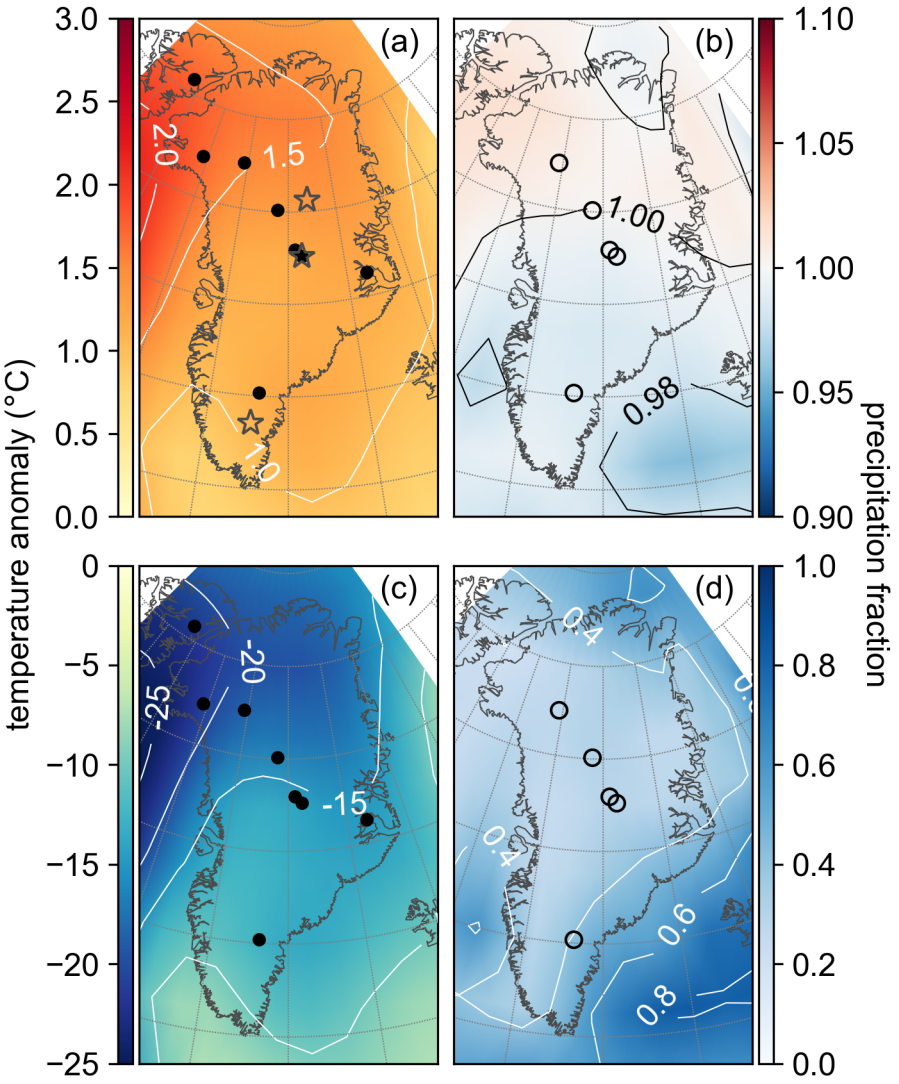

Figure 4. Spatial pattern of the reanalysis mean for temperature (panels (a), (c)) and precipitation (panels (b), (d)) with contours for clarity. (a) and (b) are averaged over 1,000 years around the peak warmth in the Holocene, 5.5-4.5 ka, while (c) and (d) are averaged over 5,000 years in the late glacial, 20-15 ka. Anomalies and fractions are with respect to the mean of 1850-2000 CE. Points show ice-core locations used for each reanalysis with closed circles indicating δ^{18}O records and open circles indicating accumulation records. Grey stars show the locations of the EGRIP ice-core site, Summit, and South Dome, which are referenced in Figs. 5 and 11.