Reviewer Responses and Manuscript Revisions

Below are the authors' responses to the reviewers. We've included our specific manuscript revisions as blue bullet points below the responses. The bullet point page and line distinctions refer to the marked-up version of the manuscript, which is also found below.

Reviewer #1 Response

I find that this article does not belong to the journal of Climate of the Past. Thus, I reject the publication. However, I find the idea novel and interesting and I would suggest to submit a revised version to a more theoretical journal.

Here is why it doesn't belong to Climate of the Past: The idea of using LIM as a substitute for otherwise expensive online data assimilation sounds wonderful, and would prove to be it if one didn't need to introduce a parameter a. But as authors showed with a being equal to 1, the results of linear online DA are worse than of offline DA. Then the question I pose to authors is how to choose an optimal a? What would be the criteria for optimal a?

We thank the referee for commenting on this paper. On this first point, we disagree with the assessment that this paper does not belong in the journal of Climate of the Past (CP), and our reply to that can be found below. On the need for a tuning parameter (such as a) in the usage of ensemble data assimilation methods, it is common due to the affect such parameters have on ensemble variance. As in Hamill and Snyder (2001), we present results for a range of the blending coefficient (a), but show that there is a general range of blending that gives improved validation metrics compared to the offline case. We do not give an absolute determination of an optimal (a) because, as we show in the paper, the choice of metric for determining the best value for (a) is subjective. The use of correlation, the coefficient of efficiency, and the continuous ranked probability score (CRPS) target different aspects of comparison between the reconstructed and reference data (though CE and CRPS are very similar). The details of the differences in different skill metrics are described in Section 3.1. Altogether, the determination of which metric is best depends on what the end user prioritizes.

- To clarify the goals of this study, we revised the final paragraph in section 3 to reflect that we are searching for reconstruction improvements and how we go about judging those improvements. (pp 6, lines 18-19)

There is yet another manifestation of more theoretical work to be done, mainly in Sec. 4.2, where authors find inconsistency between the best model 20CR in terms of a scalar skill (CE, r, and CRPS) but worse in terms of spatial reconstruction. They

propose that it could be due to a short time scale but this could be checked. And again, does this mean that α = 0.9 is not optimal for 20CR?

We agree that this result can be more thoroughly discussed. The 20CR dataset is a reanalysis performed using surface pressure observations from 1850 to present (Compo et al. 2011). The observational coverage of the southern oceans, especially during the early portion of the record is very low (see Fig. 3 in Woodruff et al. 2010). This is a large region of high variability due to the southern hemisphere storm track. As Fig. 7 in our paper shows, the primary mode of spatial variability in the annually averaged 20CR dataset is quite different in character from the other 3 datasets used. The pattern of variability is focused on regions in the southern oceans, which is where we see a lot of the CE skill degradation. This coupled with the knowledge that it is a region of high variability with few observations leads us to speculate that the features of the 20CR LIM may be influenced by artifacts of the 20CR dataset. Another instrumental reanalysis product spanning 1900 – 2012 (ERA-20C; Poli et al. 2016) has a similar first EOF and forecast mode to the 20CR. The discrepancy between the GMT and spatial performance for the 20CR experiment was surprising, but similar results were found in idealized pseudo-proxy experiments (Annan & Hargreaves 2012, Wang et al., 2014). These studies show that spatial averaging can boost the signal-to-noise ratio for large-scale indices resulting in higher index skill despite poor spatial reconstruction performance. Here we have a situation that is slightly different. The spatial results for the offline case are decent, but the degradation of spatial skill by the 20CR LIM reconstruction enhances the skill in the GMT signal. We interpret this as the spatial degradation having a moderating effect on an aspect of the global signal that is overestimated in the offline case such as the interannual variance or the warming trend. We can provide a breakdown of this idea with plots of calculated spatial trends and interannual GMT variance over time.

- We have altered Section 4.2, paragraph 2-3 (pp 11-12) to reflect this information.

Again, (a=0.9) is the optimal blending when considering CE skill of GMT. Changing the blending parameter will not change the behavior of the forecasts, only how much information is used from the forecast. If a user was inclined to use the 20CR LIM for a reconstruction, they could optimize between spatial skill and GMT skill, but the use of 20CR LIM forecast information will only reduce spatial skill. From what we show in the paper, there are better options than the 20CR LIM to use for spatial reconstruction performance.

Therefore, what I suggest is to study the methodology in a theoretical framework by revising the article and submitting it to a theoretical journal.

Regarding the larger question of relevance to this journal, we believe that there exists well-established precedent for studies like the current one. The field of paleoclimate data assimilation (PDA) has had a number of recent theoretical advances, many of which were published in Climate of the Past (CP). For example, Crespin et al. (2009) extend the

ensemble selection DA method to perform online forecasts with an intermediate complexity climate model. However, they make no comparison with its offline predecessor. Bhend et al. (2012) use an idealized pseudo-proxy experiment to test an offline ensemble square root filter approach. In this work, they use a parameter known as covariance localization to prevent the collapse of ensemble variance, and they also discuss the extension to a coupled model forecast online method as the next step. Annan and Hargreaves (2012) performs an idealized pseudo-proxy experiment with an ensemble selection method to test different geographic densities of proxy observations. They also try a persistence forecast method as an online case, but find no benefits to the reconstruction from the persistence forecast in all cases. Matsikaris et al. (2015) perform a direct comparison between online and offline methods using the ensemble selection method with a 10-member ensemble forecast from a coupled GCM. They also find no discernible benefits to using the online method compared to the offline experiment. They follow up with a study (Matsikaris et al., 2016) further investigating why the online forecast method does not show improvements to their reconstruction targets. Thus it is clear that there are numerous papers in CP similar to ours; this is why we submitted the paper to CP. We have presented the only study to our knowledge that shows a viable online PDA method that can be used for long timescales with large ensembles while also documenting improvements over the offline alternative. Additionally, we document the performance of the method using real proxy data in real reconstructions. We plan to provide code and sample data along with the revised manuscript. The connection to previous CP literature and our novel results make this study both timely and relevant for this journal.

Other major comments:

LIM is calibrated on model 1 (CCSM4) without any data assimilation over a period 850-1850, on model 2 (MPI) without any data assimilation over a period 850-1850, on model 3 with data assimilation (20CR) over a different period 1850-2012, and on a data set over yet a different time period (1950-2010 I would assume, though it is not mentioned in the paper). Thus, the models are completely different in terms of the time period, use or not of the data, and only being the data. This makes it hard to compare and draw conclusions. Instead LIM should be calibrated on a model without DA, on the same model with DA, and on observations used in that DA.

That would be a nice framework for a future study, but it is clearly well beyond the scope of this paper. In any case, our method has sampled widely different sources for variability on which to calibrate the LIMs. During the instrumental period the separation between forced and natural variability is unclear. However, the last millennium simulations are a good substitute for a measure of long-term natural variability under weaker forcing regimes. As far as the use or not of observations in the calibration data, we will note that for climate models there are no common systems that provide data assimilation for simulations like this. That means a comparison like the one suggested would have to be done using systems oriented for reanalysis, which are very different from climate models. Our results show that reanalysis products may have inherent properties making them less useful for our

application. Additionally, the observations used for assimilation during the instrumental period are not necessarily something that we can use as a LIM forecast model in reconstruction. For example, the 20CR assimilates pressure observations. Other reanalysis products use a suite of different measurements including satellite radiances and upper air measurements. These are not easily useable gridded products and are outside the scope of variables we would use in paleoclimate applications. As we said, these are topics that can be explored in future research.

- We changed the text to explicitly define that the BE dataset covers the years 1960-2014. (pp 6, line 1)

As the prior authors used results of the CCSM4 model, the same model they used for LIM calibration. It appears that linear CCSM4 DA provides good results in terms of both scalar skills and spatial reconstruction. Is it because there is less inconsistency? How would it change if the prior was from another model?

We did check results using the MPI last millennium simulation as our prior and the NOAA merged land-ocean surface temperature analysis (MLOST) as the calibration data for the proxy observation models. The CCSM4 LIM still outperformed the MPI LIM by a similar margin in the spatial results (the CCSM4 spatial average CE was +.02 above the MPI case). The GMT skill results show the two models again at a virtual tie. This suggests that the CCSM4 LIM provides forecasts that generate results more consistent with the GISTEMP reference we use for validation.

In order to provide a fair comparison authors need to include "expensive" online DA (using a nonlinear model instead of LIM).

As stated in our introduction, the current computational costs of performing coupled model forecasts in large ensembles over long time periods make this suggestion impractical. Moreover, other studies have explored this possibility (with smaller ensembles forecasting on decadal timescales) and shown no benefit over offline DA. This aspect and the knowledge that a LIM can be comparable in forecast skill to the coupled models (Newman 2013) is why we chose to try a LIM for online paleoclimate data assimilation. We believe this is a fair baseline comparison showing the potential of an online method compared to an offline method. Despite the simplicity of this approach, we show improvements in reconstruction skill over the offline method, which is the first to our knowledge for any online technique. "Expensive" online PDA options likely won't be feasible with ensembles of the size we use here anytime soon.

Minor Comments:

Page 7, Line 18: Why is there a shift in blending coefficient? This is again related to my comment on how to choose an optimal a.

There is a shift in blending coefficient between CE and CRPS because they are different skill metrics. CE is calculated based on mean squared error properties, while CRPS is

calculated on accumulated mean absolute error. Though they are both sensitive to bias, phase, trend, and amplitude differences, there are no guarantees that they will give the same results. This is especially apparent for detrended GMT skill of the persistence case when (a=1.0).

- To help illustrate the difference between the CE and CRPS, we've included a figure in the supplementary information (Fig. S1) comparing the detrended GMT scores for the persistence (a=1.0) and MPI (a=0.95) cases.

Page 7, Line 31: Why is there improvement compared to offline DA even though the trend is largely underestimated for α = 0.95?

There is improvement because the trend is not the only consideration of the CRPS/CE skill metrics. Other aspects (phase and amplitude) can still be improving while the trend difference is not large enough to decrease the skill measure.

It would be interesting to introduce another metric – bias – in order to check whether the model either underestimates or overestimates the observed values.

The CE metric can be separated to show skill related specifically to bias. We will consider expanding the GMT and spatial breakdown of CE skill between bias and other aspects of the reconstruction.

- We included some discussion of bias, amplitude, and trend controls related to GMT skill in section 4.1, paragraph 1 (pp 8), and a more specific discussion related changes to the GMT skill in the 20CR LIM reconstruction (pp 11, lines 23-33).

I suggest plotting time series of averaged temperature of different models against observations for best a for CE, for best a for r, and for best a for CRPS.

We agree that we should include a figure of the actual reconstructed GMT for selected reconstructions. Thank you for this suggestion.

- We included a plot of reconstructed GMT against GISTEMP (Fig. 2) for each experiment using the best a for CE (considering a > 0)

Reviewer #2 Responses

We thank reviewer #2 for thorough and helpful comments on the manuscript.

General Comments:

The presentation of the work in the paper feels incomplete in several ways. The figures show comparisons of metrics of skill, and comparisons of estimated climate fields to a benchmark estimate, but no visualization of the reconstructions themselves, or comparisons to the actual target (the GISTEMP field and/or GMT time series)

We agree that the paper can be improved by the inclusion of figures detailing reconstructed data against the target data of, for example, GISTEMP. We will use the specific questions posed by you and the other reviewer to guide an expansion of the discussion/interpretation for these results.

- We have included the requested figures (GMT comparison (Fig. 2), skill uncertainties (Fig. 3), and non-differenced measures of skill against GISTEMP (Fig S2 and S3)). We also expanded our discussion of GMT skill (pp. 8, lines 3-7), the North Atlantic/Barents Sea skill changes (pp 10-11, lines 31-3), and the 20CR spatial and GMT skill behavior (pp 11-12, section 4.2, paragraphs 2-3).

The authors have also neglected to describe of tabulate the computational expense associated with their reconstruction exercises. In addition, I wonder if they plan to make code for carrying out any of the reconstructions publicly available

On our desktop workstation (4-core CPU @ 3.4 GHz), the time required for a single iteration for a 151-year reconstruction with 100 ensemble members and using 110 proxies is between 1.5 – 2 minutes. Thus, a full experiment (100 iterations for each of the 12 blending coefficients), 1200 100-member reconstructions per experiment, takes around 40 hours. However, the iterations themselves can be run in parallel on different nodes of a computing cluster. In comparison, the offline method (no forecasting) takes about 1 minute to complete a single iteration. Therefore, the addition of the LIM approximately doubles the computational expense in the worst case, but the total wall-clock time is still quite manageable for large experiments. We have not taken steps to fully optimize the code so there are likely ways to lower the overall computational expense. We plan to archive the version of the code that was used in these experiments, but also plan for a public version of the code hosted on a platform such as Github. This repository and documentation will be coordinated with other Last Millennium Reanalysis projects.

- This information has been placed in supplementary information (Section S1).

Specific Comments:

The authors would like to note that any suggestion not directly addressed was taken into account when revising the text.

Title:

- On the reviewer's suggestion, we've altered the title to: "Reconstructing paleoclimate fields using online data assimilation with a linear inverse model"

Abstract:

LIMs have been shown to have comparable skill to CGCMs in what sense?

LIMs have been shown to have comparable skill to CGCMs for forecasting surface temperature anomalies. We will be more specific in the text.

- Revised (pp 1, lines 5-6)

The last sentence may need to be revised or made more specific, to address the meaning of the "dynamical evolution" to which the authors attribute improvements in skill. When I think of "dynamics," I think of the description of the underlying physical mechanisms driving changes in time, where the term is used in contrast to a "statistical" description. The LIM is purely statistical though, so I think the authors mean the term in the sense of using the model forecast as a prior for each subsequent timestep.

By "dynamical evolution", we imply that the LIM has encoded linear dynamical properties of the system that it is calibrated on. Using a LIM to forecast the next prior imparts those linear dynamical constraints on the forecast field. We will clarify this distinction in the text.

- Revised (pp 1, line 17)

Introduction:

It seems the physical consistency issue due to use of EOFs is also a limitation of the method presented in this paper though, right? Seems a bit disingenuous to list this here as if it's a limitation the present approach will address

We understand the reviewer's point, but EOFs are simply used here as a basis-reduction method, which is standard practice in the LIM literature. We make no claims that the EOFs have a physical basis in isolation (although others would make that claim). Whether the model basis is grid points, spherical harmonics, or EOFs, the goal of representing the dynamics remains the same. EOFs happen to be a compact space that, in total, captures more variance than other approaches with similar degrees of freedom.

- We changed section 2.1, paragraph 2 (pp 4) to highlight that the EOFs are used as a basis, and to re-state the limits from this choice.

The authors might expand upon what they mean by "dynamical" at the first use of the word here, to make the precise nature of their contribution more immediately accessible to a wider audience.

We will define more clearly what we mean by "dynamics" in the context of the reconstruction.

- Added clarification (pp 2-3, lines 36-1)

How many modes are retained in this study? (This detail is sufficiently important to be moved from the appendix to the main paper). What's the justification for the choice based on e-folding times of a year or greater? Are results sensitive to number of retained modes?

We retain 8 EOFs for use in the LIM and will move this detail into the main text. In the appendix, we say that number of modes covers those with e-folding times (EFT) of 1-year or greater meaning that we picked a number of modes that encompasses those EFTs, not necessarily restricts the EFTs to 1-year or greater. For example, the CCSM4 LIM has four

forecast modes with EFTs above 1-year and three modes around 1-year (EFT of 0.7 and 0.85 years respectively). Other LIMs have similar distributions of EFTs. We keep these modes because EFTs much lower than 1-year will not have a large impact when we are forecasting on annual time scales. It is worth noting that Newman (2013) keeps 11 EOFs for SSTs and 6 EOFs for land temperatures since he is using separate observational datasets, but he also states that his results are insensitive to the exact number of EOFs he retains. He also finds that most of the LIM skill on 2-9 year time scale can be reproduced with the first three forecast modes. This gives us confidence that retaining 8 EOFs for all our LIMs is a reasonable choice. We realize the statement in the text on the EOFs can be misleading and will amend it to state the reasoning behind our retained modes more clearly.

- Added the number of modes retained to Section 2.1 (pp 4, lines 16-17) and clarified our reasoning for this choice in Appendix B, list-item 4 (pp 16, lines 7-10)

Data and experimental configuration:

"For the prior, we used . . .. the CCSM4 last-millennium simulation": Do the authors mean this is the model used for the climatological prior used for the blending used to prevent the collapse of the ensemble as described at the end of the previous section? If so: I would expect the EOFs of the prior and the CCSM4-based LIM to be the same, but different for the other CGCM-based LIMs, thereby perhaps giving the CCSM4-based LIM an advantage, or at least somehow controlling the divergence of that ensemble differently than for the three other CGCM-based LIMs?

Yes, we mean that the CCSM4 simulation is used as the static prior. We also considered that the CCSM4 LIM had a slight edge because of the usage of CCSM4 as a prior, but we do not believe this to be the case based on further investigation. We performed the same LIM experiments using the MPI last millennium simulation as a prior and the NOAA Merged Land-Ocean Surface Temperature analysis (MLOST) to calibrate the proxy observation models. The GMT skill between the CCSM4 and MPI-prior experiments is basically the same, but the spatial results show the CCSM4 LIM outperforms the MPI LIM experiment in both cases (the CCSM4 spatial average CE was +0.02 above the MPI case when using the MPI prior).

- Added emphasis that the prior choice is referring to the static ensemble used for blending (pp 6, lines 6-7)

The linear observation models for proxy data" should be described in enough detail to enable reproducibility. Are the proxy data simply linear in temperature of the gridcell containing each proxy location? Or a collection of gridcells representing the regional signal of each record referred to in the next sentence?

The observation models we use are the same as discussed in Hakim et al. (2016). The model is formed by a linear regression against the time series from the nearest grid point in the calibration dataset.

- Added a description of the observation models (pp 6, lines 8-9)

Also, is there any particular justification for the choice of the GISTEMP product for calibrating the proxy models?

The choice of GISTEMP as the calibration data was an arbitrary choice. The observational dataset used to calibrate the proxy models will influence the reconstruction, but this sensitivity is outside the scope of the current study and is discussed in Hakim et al. (2016) (see section 4).

Finally, it's interesting the authors use several proxy types with known differences in their spectral signatures. Is there any difference in the construction of linear observation models for the lower versus higher frequency proxies?

All proxies we use have observations provided at annual resolutions. There is no difference in how we calibrate between them here. This is a current topic of research given that we know that some proxies provide mostly seasonal information (e.g. growing season for tree ring widths/densities).

Optimal in what sense? What is the criterion used to determine the optimum?

We use optimal in the sense that it can provide some improvements over the offline method in our experiments, but we see now how this is a vague usage. In our results, we use the GMT CE skill as the measure to define optimal. We then investigate the spatial results for the blending coefficient that achieved the best CE in all experiments except for the persistence forecast experiment. The persistence forecast did not have a blending coefficient showing an improvement over the offline case, so we used the best performing blending coefficient for the detrended GMT CE.

- To clarify the goals of this study, we revised the final paragraph in section 3 to reflect that we are searching for reconstruction improvements and how we go about judging those improvements. (pp 6, lines 18-19)

Results and Discussion:

be more specific than writing the CRPS and CE results are "generally consistent." Do you mean the rank of models is the same as measured by both statistics? Line 18-20: Similar to preceding comment: it's imprecise to say there are "slight differences in results. . . when comparing CE and CRPS." These are two different metrics that measure different things in the first place. I wonder again if the authors mean to make a statement comparing the rank of models as measured by the two different metrics?

Yes, by generally consistent we mean that the rank of the experiments is the same, and the behavior of the two metrics across different blending coefficients is similar. We realize that the CRPS and CE are different metrics and should not be expected to be exactly the same. We simply want to bring attention to the fact that CE and CRPS are largely giving us the same information for the full GMT results. However, we also show with the detrended GMT how the CRPS can give a very different result from CE (persistence experiment for a=1.0)

where it could be very misleading to a user who cares about interannual variability. We will alter this section to make the language more precise.

- We altered the discussion of the CRPS metrics to clarify the comparison with CE (pp 8, lines 13-19), and added a figure to the supplement (Fig. S1) to show the divergence in skill for the detrended GMT of the persistence experiment (a=1.0).

In all figures, the authors show central estimates across ensembles, but no measures of uncertainty. Once it has been established in the results that the skill varies with the blending coefficient, it might be interesting to show some analysis of estimates and uncertainty across ensemble members for fixed "a" (probably at the value that optimizes one metric or another).

We have done some uncertainty quantification, and can indeed include an expansion of this information in the main text or additionally in the supplementary information.

- We included Fig. 3, which shows bootstrap uncertainty estimates for the CE, correlation, and GMT trend from the best performing blending coefficient (as measured by full GMT CE) for each experiment.

I would speculate that this underestimation of trend in combination with skillful match of phase and amplitude of GMT variability might be interpreted in terms of the paleoclimate proxies as high-frequency bandpasses of the climate signal, that do not tend to preserve the low-frequency signal. This is a well-known feature of many dendrochronologies, for example, although there do exist "standardization" methodologies to prepare tree ring time series to preserve the low-frequency signal. It would be interesting to know whether the proxy time series used in this study have been prepared using methods aiming to preserve low frequency climate variability

We used proxy records contained in the PAGES 2k Consortium (2013) database with no additional processing. The proxies in this database were selected as being representative of annual or warm season temperature variability. We do not believe it is the proxy data controlling the variation in reconstructed trends. Instead, we think the trend changes are related to the LIM forecasts and blending. This is because the same proxies are used in each experiment, and from these same proxy records we find varying trend estimates (including overestimates) that provide better skill than the offline case. The aspect controlling the trends is the relative weighting between the prior and proxy information, which is determined by the variance (uncertainty) of the two sources. The uncertainty of the proxies is fixed, so that means the weighting is being affected by the LIM forecast and the amount of blending between the static and forecast states. We discuss this on page 8 lines 4-12.

Subsection 4.1 is missing figures and reporting of the estimated GMT time series compared to the target GMT time series. Pp. 9, line 21– seems odd not to show some spatial measures of skill against the target, rather than just against the offline case.

Thank you for this suggestion, we will include figures of the actual temperature reconstructions and spatial measures against the GISTEMP target.

- We added a GMT comparison between the reconstruction experiments and GISTEMP (Fig. 2) to the main text, and added the spatial correlation difference maps (Fig. S4) as well as the non-differenced spatial CE and correlation maps (Figs. S2 and S3) to the supplementary information.

Is there climatic significance to this North Atlantic/Barents Sea area that might explain why the LIM forecast- based reconstructions seem to improve skill there compared to the offline case? Or can the authors speculate as to why this region has low skill in the offline case to begin with to explain the near uniform improvements there under forecasting?

The North Atlantic/Barents Sea region is a region near the sea ice edge that seems to be poorly constrained by proxy data alone. We inspected a grid point northeast of Iceland where there are negative CE values to assess the causes of the improvement. We found that the temperature variance of the offline experiment is an order of magnitude larger than the variance in GISTEMP at this location. The CCSM4 LIM reconstruction had approximately 30% as much variance in this location as in the offline case. There was not a significant change in correlation or trend between the offline and LIM experiment, so the change in temperature variance is the primary factor in CE improvement. As far as why we might expect a LIM to help here, a modeling study used a LIM to assess the predictability of the HadCM3 coupled climate model in the North Atlantic (Hawkins and Sutton, 2009), suggesting that there is decent predictability in the North Atlantic/Barents Sea region with annual lead time (see Fig. 6 in their paper). The Newman (2013) study also shows predictability for this region when using their detrended LIM for 2-5 and 6-9 year lead times. Additionally, a preliminary LIM predictability analysis we performed shows some positive anomaly correlation skill for this region in all the LIMs we test in this study.

- We included this as an expanded discussion point in section 4.2, paragraph 1 (pp. 10-11, lines 31-3)

"There is a clear distinction between LIMs calibrated on data from the shorter instrumental era, and the millennium-scale climate simulation data"– This is an interesting point. Remind readers explicitly at this point which are which, so that readers can easily reference what you're talking about in the figures. Also, can you describe the distinction you mean clearly and precisely? Looks to me like the millennial-scale ones have fewer regions of large-amplitude degradation in CE relative to the offline case.

We will rework the sentence after this (pp. 10, line 4-5) to make this point clearer.

- Revised with clarification (pp. 11, lines 15-18)

References:

Annan, J. D., & Hargreaves, J. C. (2012). Identification of climatic state with limited proxy data. *Climate of the Past*, *8*(4), 1141–1151. http://doi.org/10.5194/cp-8-1141-2012

Bhend, J., Franke, J., Folini, D., Wild, M., & Brönnimann, S. (2012). An ensemble-based approach to climate reconstructions. *Climate of the Past*, *8*(3), 963–976. http://doi.org/10.5194/cp-8-963-2012

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., … Worley, S. J. (2011). The Twentieth Century Reanalysis Project. *Quarterly Journal of the Royal Meteorological Society*, *137*(654), 1–28. http://doi.org/10.1002/qj.776

Crespin, E., Goosse, H., Fichefet, T., & Mann, M. E. (2009). The 15th century Arctic warming in coupled model simulations with data assimilation. *Climate of the Past*, *5*(3), 389–401. http://doi.org/10.5194/cp-5-389-2009

Hakim, G. J., Emile-Geay, J., Steig, E. J., Noone, D., Anderson, D. M., Tardif, R., … Perkins, W. A. (2016). The Last Millennium Climate Reanalysis Project: Framework and First Results. *Journal of Geophysical Research: Atmospheres*. http://doi.org/10.1002/2016JD024751

Hamill, T. M., & Snyder, C. (2000). A Hybrid Ensemble Kalman Filter–3D Variational Analysis Scheme. *Monthly Weather Review*, *128*(8), 2905–2919. http://doi.org/10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2

Hawkins, E., & Sutton, R. (2009). Decadal predictability of the Atlantic Ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling. *Journal of Climate*, *22*(14), 3960–3978. http://doi.org/10.1175/2009JCLI2720.1

Matsikaris, A., Widmann, M., & Jungclaus, J. (2015). On-line and off-line data assimilation in palaeoclimatology: a case study. *Climate of the Past*, *11*(1), 81–93. http://doi.org/10.5194/cp-11-81-2015

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

Newman, M. (2013). An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies. *Journal of Climate*, *26*(14), 5260–5269. http://doi.org/10.1175/JCLI-D-12-00590.1

PAGES 2k Consortium. (2013). Continental-scale temperature variability during the past two millennia. *Nature Geoscience*, *6*, 339–346. http://doi.org/10.1038/NGEO1797

Poli, P., Hersbach, H., Dee, D. P., Berrisford, P., Simmons, A. J., Vitart, F., … Fisher, M. (2016). ERA-20C: An atmospheric reanalysis of the twentieth century. *Journal of Climate*, *29*(11), 4083–4097. http://doi.org/10.1175/JCLI-D-15-0556.1

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J. E., & Rajaratnam, B. (2014). Evaluating climate field reconstruction techniques using improved emulations of real-world conditions. *Climate of the Past*, *10*(1), 1–19. http://doi.org/10.5194/cp-10-1-2014

Reconstructing ~~past climate by~~ paleoclimate fields using ~~proxy~~ online data ~~and~~ assimilation with a linear ~~climate~~ inverse model

Walter A. Perkins[1] and Gregory J. Hakim[1]

[1]University of Washington, Seattle, WA, USA

Correspondence to: Walter Perkins (wperkins@uw.edu)

Abstract. We examine the skill of a new approach to climate field reconstructions (CFRs) using an online paleoclimate data assimilation (PDA) method. Several recent studies have foregone climate model forecasts during assimilation due to the computational expense of running coupled global climate models (CGCMs), and the relatively low skill of these forecasts on longer timescales. Here we greatly diminish the computational cost by employing an empirical forecast model (linear inverse model; LIM), which has been shown to have comparable skill to CGCMs for forecasting annual-to-decadal surface temperature anomalies. We reconstruct annual-average 2m air temperature over the instrumental period (1850 - 2000) using proxy records from the Pages 2k Consortium phase 1 database; proxy ~~system~~ models for estimating proxy observations are calibrated on GISTEMP surface temperature analyses. We compare results for LIMs calibrated on observational (Berkeley Earth), reanalysis (20th Century Reanalysis), and CMIP5 climate model (CCSM4 and MPI) data relative to a control offline reconstruction method. Generally, we find that the usage of LIM forecasts for online PDA increases reconstruction agreement with the instrumental record for both spatial fields and global mean temperature (GMT). Specifically, the coefficient of efficiency (CE) skill metric for detrended GMT increases by an average of 57% over the offline benchmark. LIM experiments display a common pattern of skill improvement in the spatial fields over northern hemisphere land areas and in the high-latitude North Atlantic – Barents Sea corridor. Experiments for non-CGCM-calibrated LIMs reveal region-specific reductions in spatial skill compared to the offline control, likely due to aspects of the LIM calibration process. Overall, the CGCM-calibrated LIMs have the best performance when considering both spatial fields and GMT. A comparison with the persistence forecast experiment suggests that improvements are associated with the ~~dynamical evolution~~linear dynamical constraints of the forecast, and not simply persistence of temperature anomalies.

1 Introduction

Climate field reconstructions (CFRs) aim to provide essential information on climate variability beyond the instrumental record. These experiments take noisy and sparse proxies (e.g. tree rings, ice cores, isotope ratio measurements, etc.) and use them to infer a spatial estimate of relevant climate variables. A common approach to CFR uses a statistical regression model calibrated on the instrumental record to project as far into the past as data will allow (e.g. Mann et al., 1998, 2009; Smerdon et al., 2011b, see Smerdon et al., 2011a for a discussion and comparison of methods). ~~This technique provides~~ These techniques provide a useful estimate of past spatial patterns (Wahl and Smerdon, 2012), but ~~it also has~~ also have inherent limitations. For example,

regression-based CFRs assume climate state to be a function of the proxy data, which can lead to an underestimation of past climate anomaly amplitudes (Smerdon et al., 2011b; Wahl and Smerdon, 2012). Furthermore, because regression methods produce past spatial fields through combinations of primary variability modes (i.e. empirical orthogonal functions; EOFs), the resulting field is not guaranteed to be a physically consistent solution.

An alternate method of performing CFRs known as paleoclimate data assimilation (PDA) can circumvent some of the limitations inherent to regression-based methods. PDA broadly characterizes a set of techniques where observational information from proxy data ~~can be optimally~~ is combined with dynamical information from climate models. Recently, the ensemble Kalman filter (EnKF) was adapted for use with time-averaged observations like those used in CFRs (Dirren and Hakim, 2005; Huntley and Hakim, 2010). Studies using the EnKF method and idealized psuedoproxy experiments have shown that it operates well under sparse data availability (Bhend et al., 2012), and outperforms modern statistical CFR methods (Steiger et al., 2014). More recently, EnKF PDA was tested with real proxy data in the Last Millennium Reanalysis project (LMR; Hakim et al., 2016), and shows promising skill in reconstructing robust spatial fields in a computationally efficient manner. Due to the expense of performing coupled global climate model (CGCM) simulations and relatively low forecast skill, the initial EnKF adaptation for PDA does not use a forecast and instead reconstructs each time period independently using climatological data. This is known as an "offline" approach. The EnKF method is traditionally accompanied by forward model forecasts to translate information between analysis time periods (e.g. reanalysis products of the instrumental era). Dynamical constraints from these forecasts can increase physical consistency and reconstruction skill given that the model has sufficient predictability on proxy timescales (e.g. Pendergrass et al., 2012). For CFR applications, predictability on seasonal and longer timescales is required. Ocean memory can be leveraged for inerannual (e.g. El Niño Southern Oscillation; ENSO) to potentially decadal predictability (Branstator et al., 2012). However, at this timescale coupled climate models only seem to capture linearly predictable dynamics (Newman, 2013).

Online assimilation has been attempted using other PDA techniques. Crespin et al. (2009) used forecasts from an earth system model of intermediate complexity (EMIC) in conjunction with the ensemble selection PDA method (see Goosse et al., 2006, 2010) to reconstruct surface temperatures, but did not investigate a comparison with an offline method. Annan and Hargreaves (2012) performed a psuedoproxy experiment using a weighted ensemble selection method and a persistence forecast to reconstruct surface temperatures, but found no benefit compared to their offline experiments. Matsikaris et al. (2015) took a similar approach to Crespin et al. (2009), but used an ensemble of decadal forecasts from a coarse resolution CGCM instead of an EMIC. The authors found that the use of CGCM forecasts had skill, but it was not discernibly superior to the offline method. Possible reasons for the lack of improvement include low skill for regional decadal forecasts of temperature, and issues related to ocean initialization for each decadal interval.

These results suggest that neither the simple persistence forecast, nor a small ensemble of decadal CGCM forecasts add ~~valuable~~ significant information to CFRs. In order to test the viability of a more traditional EnKF method we require the ability to perform annual forecasts for longer time spans (the past millennium) and in large ensembles (~100 members). These requirements rule out the use of a CGCM. Instead, we explore a simple, empirically-based forecast from a linear inverse model (LIM; Penland and Sardeshmukh, 1995). A LIM encodes the linear dynamical properties of a system and produces forecasts

that are subject to the constraints of its derived linear modes. The forecast skill of LIMs is such that they are currently used for operational ENSO forecasts (Newman et al., 2009). Moreover, recent studies ~~found~~ show LIM skill to be comparable to ~~the skill~~ that of CGCMs when performing annual-to-decadal hindcast experiments over the instrumental era (Newman, 2013; Huddart et al., 2016).

~~In this work~~Here, we propose a computationally efficient "online" data assimilation approach for use in paleoclimate field reconstructions. The primary goal is to investigate whether the addition of dynamical constraints with a forecast in the online case can increase reconstruction skill relative to the offline EnKF method, which has no forecasting. We perform a series of reconstruction experiments using annual forecasts from a cost-efficient LIM. Global average and spatial ~~results of the online reconstructions~~ reconstructions from the online experiments are compared to results from both a persistence forecast method and the offline method of Hakim et al. (2016). In ~~Sect.~~ Section 2 we discuss the basics of the EnKF method and define the use of LIM forecasts in reconstruction experiments. Section 3 details the datasets used and the general experimental configuration. Section 4 discusses and compares results between the online and offline reconstructions, followed by conclusions in ~~Sect.~~ Section 5.

2 Online PDA

The Last Millennium Reanalysis (LMR) framework (Hakim et al., 2016) provides a setting to run many computationally efficient realizations of an offline climate reconstruction. Here we begin with it as the basis for our implementation and investigation of online PDA. Central to the LMR framework is the use of the ensemble Kalman filter (EnKF; Kalnay, 2003), which assumes gaussian distributed errors. The EnKF update equation (Eq. 1) describes the calculation of a posterior (analysis) state vector $\mathbf{x}_a$ through the optimal update to a prior (background) state vector $\mathbf{x}_b$ using proxy information,

$$\mathbf{x_a} = \mathbf{x_b} + \mathbf{K}[\mathbf{y} - \mathcal{H}(\mathbf{x_b})]. \quad (1)$$

The innovation, $[\mathbf{y} - \mathcal{H}(\mathbf{x}_b)]$, characterizes new information content as a difference between proxy observations in vector $\mathbf{y}$ and observations estimated from the prior by $\mathcal{H}(\mathbf{x}_b)$ (hereafter denoted as $\mathbf{y}_e$). $\mathcal{H}()$ is a potentially non-linear operator that maps the prior state into observation space. The Kalman gain matrix $\mathbf{K}$, defined by Eq. (2), spreads information into the analysis weighted by prior covariance and the observational error covariance matrix $\mathbf{R}$

$$\mathbf{K} = cov(\mathbf{x}_b, \mathbf{y}_e)[cov(\mathbf{y}_e, \mathbf{y}_e) + \mathbf{R}]^{-1} \quad (2)$$

where $cov(a, b)$ represents a covariance expectation. The LMR framework uses a variant of the EnKF update, known as an ensemble square root filter (EnSRF; Whitaker and Hamill (2002)). This process updates the ensemble mean and perturbations from the mean separately allowing for the serial assimilation of proxy data, and simplification of the update calculations.

Typical implementations of the EnKF method include a ~~forward~~ model forecast between analysis times. As stated earlier, the computational expense and low skill of CGCM forecasts prompted the use of the offline method where each year is

reconstructed independently without forecasting. Here, instead of using static prior ($\mathbf{x}_b$) at the beginning of each reconstruction year, the current year's posterior analysis is forecast forward by one year with a LIM defined by

$$\mathbf{x}_b^f = \mathbf{G}_1 \mathbf{x}_a. \tag{3}$$

The term $\mathbf{G}_1$ is a mapping term calculated from the calibration of a LIM that maps the current state to a forecasted state 1-year later. Details of the EnKF reconstruction algorithm can be found in Appendix A. The formulation of the LIM used here is described in the following section.

2.1 Linear inverse model formulation

The linear inverse model (LIM; e.g. Penland and Sardeshmukh, 1995) used in this study closely follows the implementation described in Newman (2013). The basic equation describes a linearized dynamical system

$$\frac{d\mathbf{x}}{dt} = \mathbf{L}\mathbf{x} + \xi \tag{4}$$

as the tendency of an anomaly state vector $\mathbf{x}$, given by a ~~linear~~ dynamical operator $\mathbf{L}$, which is linearized about a mean state, plus random white noise, ξ. The dynamical operator $\mathbf{L}$ is assumed to be constant in time. After integrating (Eq. 4) in time, the solution is a mapping of $\mathbf{x}$ at time t (in years) to a state at time $t+1$

$$\mathbf{x}(t+1) = \mathbf{G}_1 \mathbf{x}(t) + \sigma(t) \tag{5}$$

where $\mathbf{G}_1$ is equivalent to $e^{\mathbf{L}}$. As in Newman (2013), we choose to empirically estimate $\mathbf{G}_1$ rather than $\mathbf{L}$ due to sampling deficiencies of a few highly damped eigenmodes of $\mathbf{L}$ on an annual timescale. Each LIM is constructed using an EOF basis retaining the leading 8 modes of variability. See Appendix B for a summary of the details associated with the $\mathbf{G}_1$ calculation.

While the simplicity of a LIM makes it well suited for the current application, it also has issues to be considered. First, LIM forecasts are performed ~~in EOF space based on patterns~~ using an EOF basis derived from the calibration data. ~~As such, the LIM makes an assumption of stationarity for the EOFs and encoded dynamics over the entire reconstruction time period. Furthermore, the process of converting data into EOF space truncates spatial anomaly information and reduces ensemble variance. The next section discusses how we handle variance reductions when using a LIM in the LMR framework~~The EOF basis is used to maximize the variance captured by the fewest degrees of freedom. Although the dominant modes of variability can change in time during the reconstruction period, the space spanned by the variability cannot. This means that variability that falls outside the span of the EOFs will not be resolved. A second issue is that the LIMs tend to be damped (modes of L decay in time), which reduces the ensemble variance; we elaborate on this issue in the next section.

2.2 Ensemble calibration

In any ensemble forecast setting, a basic assumption is ~~made~~ that the sample of ensemble members gives a good approximation to the statistics of the full system (Murphy, 1988). Sampling error often results in too-small variance, which can cause "filter divergence" where observational information is underweighted relative to the forecast prior and the ensemble variance collapses toward zero. The online PDA technique presented here is especially vulnerable to filter divergence because all eigenmodes of $\mathbf{G}_1$ are damped (negative real eigenvalues). Moreover, the conversion of the analysis ($\mathbf{x_a}$) into EOF space at each timestep removes any spatial information that does not project upon the retained modes of a given LIM. Consequently, LIM forecasts lose ensemble variance in time.

There are a variety of well-tested methods available to address information loss in the forecast ensemble. Here we use an adaptation of the hybrid ensemble Kalman filter–3D variational scheme (Hamill and Snyder, 2000) to prevent filter divergence and to facilitate comparison with the offline PDA technique. This technique handles the loss of ensemble variance in the forecast ensemble ($\mathbf{x}_b^f$) by blending it with a static source ($\mathbf{x}_b^s$), which is the same climatological prior that ~~would be~~ is used independently for each year in the offline method. As a result, the update equations use a blended prior state $\hat{\mathbf{x}}_b^f$ (Eq. 6) and a blended Kalman gain term $\hat{\mathbf{K}}$ (Eq. 7):

$$\hat{\mathbf{x}}_b^f = a\mathbf{x}_b^f + (1-a)\mathbf{x}_b^s \tag{6}$$

$$\hat{\mathbf{K}} = \frac{(a)cov(\hat{\mathbf{x}}_b^f, \hat{\mathbf{y}}_e^f) + (1-a)cov(\mathbf{x}_b^s, \mathbf{y}_e^s)}{(a)cov(\hat{\mathbf{y}}_e^f, \hat{\mathbf{y}}_e^f) + (1-a)cov(\mathbf{y}_e^s, \mathbf{y}_e^s) + \mathbf{R}}. \tag{7}$$

Appendix A provides details on how this is incorporated into the reconstruction algorithm.

In these hybrid DA equations, the parameter a controls the relative weighting between static and forecast information sources. When $a = 0.0$, reconstructions are identical to the offline case wherein the prior $\hat{\mathbf{x}}_b^f$ is reset to the static prior for every year with no blending. ~~For the opposite case,~~ When $a = 1.0$, only forecast information is used with no contribution from static information.

3 Data and experimental configuration

The relative forecast skill of a LIM is dependent on the data used to empirically derive the mapping term $\mathbf{G}_1$. For this reason, we explore LIMs calibrated on four different data sets. CGCM calibration data are used from two last-millennium climate simulations in the Coupled Model Intercomparison Project phase 5 (CMIP5; Taylor et al., 2012): the Community Climate System Model v4 (CCSM4; Landrum et al., 2013) and the Max Planck Insitute Earth System Model paleo-mode (MPI). These simulations cover a 1000 year pre-industrial (850–1850 C.E.) time period including volcanic forcing events (aerosols and greenhouse gases), solar variability, and human-related land cover changes. The 20th Century Reanalysis (20CR; Compo et al., 2011), a DA synthesis of observations and a weather forecast model, provides over 150 years of reanalysis data spanning the instrumental record (1850–2012). Finally, we use the Berekely Earth surface temperature dataset (BE; Rohde et al., 2013)

~~as an~~ for observational calibration. BE provides a ~~60-year sample~~ 65-year sample (1960–2014) with nearly complete global coverage. The different LIM calibration datasets used here span linear modes of predictability derived from model space to that of observations.

The basic configuration we use for all experiments, including the offline control, involves a choice of data to sample as the ~~prior~~static prior ensemble, an instrumental data source to calibrate proxy observation models, and a proxy record dataset. For the static prior, we use annually-averaged 2m air temperature anomalies from the CCSM4 last-millennium simulation. The ~~linear~~ observation models for proxy data are calibrated against the NASA Goddard Institute for Space Studies surface temperature analysis dataset (GISTEMP; Hansen et al., 2010) by linearly regressing the proxy timeseries against the nearest grid point in the calibration data (for details, see Hakim et al., 2016). All experiments use annually-resolved proxy records from the PAGES 2k Consortium (2013) database. These proxies have been ascertained to covary regionally with temperature and include: tree rings, ice cores, corals, sediment cores, and speleothems. Only proxies with a minimum of 10 years overlapping with the observation model calibration data, and a minimum calibration-fit correlation of 0.2 are used. It should be noted that the correlation threshold is not strictly necessary, but Hakim et al. (2016) found that this threshold did not quantitatively affect the reconstruction results. Here, the reduction in proxies to those with more information helps reduce computational costs, allowing a larger number of reconstruction experiments.

We reconstruct annual-mean 2m air temperature anomalies for the period of 1850–2000 C.E. as in Hakim et al. (2016). Using the four LIM calibrations, we search the parameter space of $0 \leq a \leq 1$ for ~~an optimal~~ a weighting between static and forecast information sources ~~in our hybrid PDA framework~~that improves the reconstruction compared to the offline method. We judge improvements by means of our chosen skill metrics (CE, correlation, and CRPS) for both GMT and spatial fields. Additionally, we perform a persistence experiment for comparison against LIM-based performance where the posterior for year ~~n~~ t is used as the prior for year ~~$n+1$~~ $t+1$. The persistence forecast uses the same hybrid PDA blending scheme as the LIM forecast experiments to mitigate the effects of reductions in ensemble variance from the assimilation process. We account for the sensitivity to the proxy data used in a CFR through random resampling of available proxy data and the static prior ensemble. ~~In a single realization~~ A single realization uses a random sample of 75% of the usable proxy records, and a 100 member sample of anomaly states from the prior~~source, are selected~~. A total of 100 realizations are performed for each LIM calibration and blending coefficient. In order to make the realizations consistent between the experiments using different blending coefficients, we ensure the same sequences of random samples are taken by seeding the random number generator for each a-value. In total, this gives 10^4 reconstructions of the climate state for each experiment. These reconstructions are then averaged to give the final analysis. See Section S1 in the supplementary information for a brief discussion of the computational costs associated with these experiments.

3.1 Skill metrics

The primary skill metrics used are correlation and the coefficient of efficiency (CE; Eq. 8,9; Nash and Sutcliffe, 1970). Correlation gives an overall sense of signal timing (phase), while CE is a stricter metric that is sensitive to signal timing, amplitude, and

bias. Using these metrics, we compare the reconstructed ensemble-mean[1] values x against GISTEMP ~~verification~~ validation values v. The value τ represents the number of ~~verification~~ validation times available (in this case representing the ~~period~~ GISTEMP timespan of 1880 - 2000), an overbar (e.g. $\bar{v}$) denotes a temporal average, while σ_x and σ_v are the standard deviations of the respective time series. Skill scores are compared for the reconstructed global mean temperature (GMT) and spatial grid points.

$$corr = \frac{1}{\tau}\sum_{t=1}^{\tau}\frac{(x_t - \bar{x})(v_t - \bar{v})}{\sigma_x \sigma_v} \tag{8}$$

$$CE = 1 - \frac{\sum_{t=1}^{\tau}(v_t - x_t)^2}{\sum_{t=1}^{\tau}(v_t - \bar{v})^2} \tag{9}$$

We also use the continuous ranked probability score (CRPS; Gneiting and Raftery, 2007) as a comparison against CE skill metrics.

$$CRPS = \sum_{t=1}^{\tau}\left(\frac{1}{K}\sum_{i}^{K}\left|x_t^{(i)} - v_t\right| - \frac{1}{2K^2}\sum_{i}^{K}\sum_{j}^{K}\left|x_t^{(i)} - x_t^{(j)}\right|\right) \tag{10}$$

The CRPS is considered to be a 'proper' scoring technique which prevents manipulations of the data from overestimating the reconstruction skill; the measure reflects the mean absolute error and narrowness of the ensemble distribution. We use the CRPS as defined in Tipton et al. (2016) to calculate the CRPS of the reconstructed GMT for each realization (Eq. 10) where $x_t^{(i)}$ denotes a single member of a K-member ensemble. We take the score as the average CRPS over all realizations and use a Kolmogorov-Smirnov test on the resulting distribution to determine whether it is significantly different than the offline case with 95% confidence. As defined here, lower values of CRPS indicate better performance with a limiting case of CRPS = 0 being a perfect ensemble fit (no error and no ensemble spread).

4 Results and discussion

4.1 ~~Verification~~ Validation of global mean temperature

Figure 1 displays global mean 2m air temperature (GMT) results ~~verified~~ validated against GISTEMP for all tested values of the blending parameter a. Every case except for the persistence forecast method ~~yield~~ yields CE values greater than the offline case. Correlations are higher than the offline benchmark for all experiments, including the persistence forecast. Best skill is achieved between the a-values of 0.7 to 0.9 with a steep drop in ~~verification~~ validation skill as a approaches unity (a pure LIM forecast) due to filter divergence (the ensemble variance decreases with time, decreasing the weight on the proxies, so that the reconstructed states diverge from reality). The CCSM4 and 20CR LIM display the best overall CE performance with a 9% improvement over the offline method. These two experiments also display a 2% increase in correlation and have slightly

[1]Ensemble mean represents the average taken over all ensemble members and all realizations.

smaller correlation than the persistence forecast experiment (Table 1). ~~Correlations for the~~ Figure 2 shows the reconstructed GMT from each experiment at the best blending coefficient (with respect to CE) compared to GISTEMP and the offline case. As evidenced by the high skill scores, the reconstructions capture the variability in the global temperature signal of GISTEMP quite well. Compared to the offline experiment, the forecasting experiments tend to decrease the amplitude of the interannual variability, which is at times largely overestimated by the offline experiment, and they tend to change the overall warming trend. There is no apparent systematic bias in the reconstructed GMTs, so skill changes for the different experiments are likely controlled by these two factors. Figure 3 displays the bootstrap estimate of 95% confidence bounds for reconstructed GMT CE, correlation, and trends at the same blending coefficients as chosen in Figure 2. For these blending choices, the CE scores for all LIM experiments show significant increases compared to the offline CE, while the persistence, 20CR, CCSM4, and MPI experiments ~~are significantly different (with~~ show significant increases in correlation. Additionally, reconstructed GMT trends fall within the 95% confidence ~~) than the offline reference based on bootstrap results for skill metric error bounds. The CE changes for all LIM experiments are also significant~~interval of the GISTEMP trend, except for the persistence case. The LIM experiments show trends much closer to the mean estimated GISTEMP trend than the offline case. CRPS values ~~for the GMT results~~ across blending coefficients (Fig. 4) ~~are generally consistent with those for the CE skill metric .~~[2] show similar skill behavior as the CE metric (albeit mirrored in the vertical[2]). Specifically, GMT ~~verification~~ validation with CRPS shows that all LIM ~~forecasting~~ experiments outperform the offline method, with the CCSM4 and 20CR LIMs again having the best performace (18% better than the offline case). All LIM experiments' best CRPS scores are also significantly better than the offline case with 95% confidence. ~~There~~ By and large, CRPS is giving nearly equivalent information as CE about the skill of the GMT reconstruction. However, as CRPS is a different metric, there are slight differences in results for the MPI and BE LIM cases when comparing CE and CRPS: the blending coefficient achieving the best score shifts to the next highest a-value in both cases, and the BE LIM outperforms the MPI LIM when considering CRPS (Table 1).

Both the CE and CRPS measures for different blending coefficients are affected by the degree of fit to the warming trend in the GISTEMP reference. The trends for all experiments are shown in Fig. 5. The trend of the offline case ($a = 0$) is 0.62 K/100 yrs, about 0.07 K/100 yrs above the GISTEMP trend and near the edge of the GISTEMP trend 95% confidence interval (Figure 3). However, for the MPI and CCSM4 LIM experiments, as well as the persistence forecast experiment, the reconstructed trend increases as the blending parameter a increases. This increase in trend away from the GISTEMP trend for the MPI and CCSM4 experiments is reflected in the lowered CE (Fig. 1) and CRPS (Fig. 4) for the a-values from approximately 0.0 to 0.6. The reconstructed trend from the two CGCM-based LIM experiments begins to decrease around $a = 0.6$ where CE also shows a significant increase towards maximum values. The persistence forecast trend has the largest disagreement, increasing to approximately 0.72 K/100 yrs for $a = 0.9$, which results in the CE and CRPS never surpassing the offline benchmark in this case. The 20CR LIM only increases the reconstructed trend slightly over the GISTEMP trend for middle a-values. The BE experiment has a decreasing trend for increasing a, and drops to a very low trend of 0.38 K/100 yrs when $a = 0.9$. Interestingly, though the trends for the MPI, CCSM4, and 20CR experiments are below that of the GISTEMP trend for $a = 0.95$, their skill

[2] ~~Note that best results for CRPS occur at minimum values instead of the maximum.~~

[2] Best results for CRPS occur at minimum values instead of the maximum.

still outperforms the offline case in all three metrics. Despite the mismatch in the overall trend, these online forecasting methods still produce better matches of phase and amplitude of GMT variability for the reconstructed anomalies compared to the offline case.

The trend results also illustrate the relative amount of proxy data utilization between these different experiments. Given that every experiment uses the same prescribed list of seeds to generate proxy record samples, the differences in reconstructed trend can only arise from differences in weighting of the proxies or the LIM forecasts. Since LIMs are calibrated on detrended data and their forecast modes are damped, the forecast contribution to a long-term ~~trends~~ global mean trend is likely small; ~~these trends are~~ the trend is instead governed by utilization of proxy information. For the EnKF PDA method, the weighting of information is controlled by the prior ensemble variance and proxy error variance. The proxy error variance is fixed for all experiments we perform, so the changes in the reconstructed trend are a result of how the LIM forecasts affect the ensemble variance. In all forecast experiments, skill and the reconstructed trends drop off severely as a approaches 1.0. When using only forecast information ($a = 1.0$), the ensemble variance collapses due to the damped properties of the LIMs, which results in filter divergence. The BE LIM case reaches its maximum CRPS and CE values at smaller a and also has the lowest reconstructed trends of the LIM experiments. This suggests the BE forecast produces less ensemble variance than the other LIMs, possibly due to forecast mode damping or poor projection of the posterior analysis into the LIM EOF space. The eigenvalues of the BE LIM's leading two forecast modes have e-folding times of 5.4 and 1.5 years, respectively. This is in the same range of the leading forecast modes of the CGCM-calibrated LIMs (e.g. 3.7 and 1.2 year e-folding times for the MPI LIM). Consequently, a poor projection of the analysis ensemble onto the forecast modes of the BE LIM is likely the cause of the reduced ensemble variance. The persistence forecast displays an interesting disparity between the skill metrics; overall, it performs the best in correlation, but the worst in CE and CRPS. Having the largest reconstructed trend suggests that the persistence case has the highest weighting of proxy data. With a persistence forecast there is no damping of reconstructed spatial anomalies or truncation of the ensemble variance from projection into EOF space. The resulting higher proxy weighting may explain why the persistence case correlation is better than the other forecasting methods. The linear observation models ~~used to estimate observations~~ in each case are based on a calibration against GISTEMP~~. Proxies that have~~ , so proxies with a better calibration fit (higher correlation) with GISTEMP have less error variance, and therefore, have more influence on the posterior analysis. The persistence case allows more information from the influential (well-correlated) proxies into the analysis because the ~~ensemble~~ prior variance is larger. However, from the CE and CRPS values, which are sensitive to more than just signal phase matching, it is clear the general trend mismatch degrades the quality of the persistence ~~forecast~~ reconstruction compared to the offline benchmark.

Removing the linear trend from each case allows for an examination of how well the reconstructions capture variability not associated with the warming trend (i.e. interannual and decadal variability; evident in Fig. 6). Generally, the performance increases for the detrended data over the offline case are much larger than for the full time series. Compared to ~~verification~~ validation with the full time series, the correlation of the detrended offline case drops from 0.9 to 0.67, and the CE drops from 0.77 to 0.29 (Table 2). With respect to CE, all experiments (including the persistence method) improve upon the offline benchmark. The 20CR LIM achieves the best improvement over the offline case (72% increase), while the persistence case

shows the least improvement (35% increase). Except in the BE LIM experiment, detrended correlation metrics again increase slightly over the offline case. The BE LIM hovers around the benchmark correlation for $0.0 < a < 0.7$ and then drops below it. A CE improvement with no change in correlation implies that the BE LIM improves the detrended anomaly amplitudes and bias, but does not improve the signal timing compared to the offline case.

The CE and correlation skill metrics of the offline experiment both decrease when calculated on the detrended GMT. In contrast, the CRPS improves by 7% (minimizing from 12.5 to 11.6). CRPS rewards reduction in mean absolute error, and 'narrowness' of the forecast ensemble, whereas CE and correlation depend more on variance properties of the reconstructed time series. In removing the linear warming trend, we remove a large degree of the time series' variance and subsequently lose the associated skill in CE and correlation. In the case of CRPS, the linear trend is only a source of mean error when it does not closely match the reference trend. ~~Since~~ When we remove the trend, ~~while not affecting the ensemble spread, the~~ the mean errors decrease (the ensemble spread is unaffected) and the CRPS metric improves. Figure 7 shows CRPS for all blending coefficients with detrended data. The behavior is quite similar to the full GMT CRPS, ~~but~~ and as the detrended CE reflects, even the persistence forecast shows improvement over the offline method. An aspect that stands out with detrended CRPS is that the persistence forecast achieves the best value when $a = 1.0$. A cursory examination of the detrended GMT timeseries of the persistence case compared to the detrended GISTEMP GMT timeseries (~~not shown~~see supplement; Figure S1) reveals that it captures some decadal variability over the instrumental record, but ~~virtually~~ none of the interannual variability; i.e., the $a = 1.0$ persistence reconstruction gives a smoothed representation of the GMT. This again highlights a difference between the two metrics of CE and CRPS. The CRPS metric, which generalizes to the mean absolute error of the ensemble summed over time, does not penalize the smoothed GMT signal approximately bisecting the interannual signal for $a = 1.0$. The CE metric, which sums the squared errors of the ensemble mean and then normalizes by the climatological variance, does penalize this behavior.

4.2 ~~Verification~~ Validation of spatial fields

Here we ~~examine~~ compare the skill of the spatial fields ~~against~~ with the offline case using correlation and CE; CRPS is omitted because the full spatial field ensembles are too large to store. ~~The offline caseshows positive skill over most of the globe except in~~ For ease of visualization, we provide CE difference maps to highlight changes relative to the offline case, but full spatial skill maps can be found in the supplementary material (Figures S2 and S3). Over the globe, skill is mostly positive, but there are a few regions with highly negative skill departures such as the Southern Hemisphere oceans and ~~in~~ the high-latitude North Atlantic to Barents Sea corridor (Fig. 8). In the high-latitude Atlantic region, the proximity to the sea ice edge seems to negatively affect the ability to constrain the temperature field when using only proxies. All LIM-forecasting cases show improvements to CE ~~, most notably~~ in the same North Atlantic to Barents Sea area, and across northern Europe into Asia; there are also smaller skill increases across western North America. ~~The~~ Studies of LIM predictability have shown the North Atlantic/Barents sea region can have forecast skill on annual and longer timescales (e.g. Hawkins and Sutton, 2009; Newman, 2013). This may be one reason why skill increases are common in this area for all LIM experiments. An inspection of grid point temperature values northeast of Iceland (see supplement; Figure S5) shows that the reconstructed temperature variance is largely overestimated in

the offline case, and that there is a slight trend in the reconstruction that is not present in GISTEMP data. The CE increase for the CCSM4 LIM experiment (increasing from -7.2 to -1.7) relates to a temperature variance reduction by approximately 70% compared to the offline case. The CCSM4, MPI, and BE LIMs generally show large CE increases in the high-latitude southern ocean. In contrast, the reconstruction with the 20CR LIM does not improve the southern ocean at all and has large deficiencies over many ocean areas. The persistence case generally shows decreases in CE across large areas of the globe. Of the global mean CE for each grid, the CCSM4 LIM gives the best performance, increasing the global mean CE by 0.09, followed by the MPI LIM with an improvement of 0.06. The 20CR and persistence cases show decreases in average spatial skill across the grid, with the 20CR being worst with a global mean CE change of -0.18. The BE LIM, while showing improvements over North Hemisphere land areas, has compensating decreases in ~~ocean skill~~ skill over the ocean that make the global mean CE nearly equivalent to the offline case. All global mean CE values, except in the BE LIM case, are significantly different (at 95% confidence) from the offline case when comparing grid point skill distributions using a Student's t-test. Changes in spatial correlation (~~not shown~~see supplement; Figure S4) are generally small in regions where ~~the CE increased suggesting that~~ CE increases, which suggests improvements are not related to signal phasing~~; however~~. However, some of the large decreases in CE for the 20CR, BE, and persistence experiments do coincide with areas of correlation decreases.

In the spatial results, there is a clear distinction between LIMs calibrated on data from the shorter instrumental era (20CR and BE), and the millennium-scale climate simulation data (CCSM4 and MPI). Compared to the offline spatial skill, large areas of CE skill degradation are apparent for both the 20CR and BE LIM reconstructions. The 20CR LIM CE skill degredations are large in amplitude ($\Delta CE < -1$) and mostly over ocean regions. It is surprising that the 20CR LIM has the worst spatial skill given that it has the best GMT timeseries CE skill. ~~The leading EOF for each LIM calibration reveals a difference in spatial structure that forms part of the forecast basis (Fig.~~ However, previous CFR studies show similar results (Annan and Hargreaves, 2012; Wang the spatial averaging over a poorly reconstructed field can boost the signal-to-noise ratio for large-scale indices enough to result in positive index skill. The situation is somewhat different in this study. In the offline case, spatial skill is reasonably positive, but when adding 20CR LIM forecasts, the spatial CE skill decreases and the GMT CE skill increases. A possible interpretation for this behavior is that the degradation in spatial field fidelity is compensating for aspects of the GMT that are overestimated in other experiments. For example, the offline reconstructed GMT overestimates interannual variability during certain time periods when compared to GISTEMP, but the usage of LIM forecasts mitigates this effect. Compared to the CCSM4 LIM experiment, the reconstructed GMT for the 20CR LIM experiment reduces the sum of the squared error (numerator term in CE) by approximately 5%. The bias accounts for only about 1% of the total sum squared error term, which leaves the trend and anomaly amplitude agreement as primary candidates for the error reduction. As shown earlier, the 20CR GMT trends tend to track lower than the two CGCM LIM experiments and closer to the GISTEMP trend (Figure 5). The trend reduction in the 20CR experiment appears to be caused by large areas of negative trends in the Southern Hemisphere (see supplement; Figure S6).

The reason behind the poor spatial performance of the 20CR LIM appears to be linked to its EOF basis (Figure 9). The leading EOF of the 20CR experiment lacks the ENSO/PDO-like pattern of the other LIMs and instead focuses on variability structures ~~rooted~~ in the southern ocean~~where relatively~~. This is a region of high variability due to its proximity to

the storm track, but there are also fewer pressure observations ~~are~~ available for assimilation ~~; many of the~~ by the 20CR (see Figure 3 in Woodruff et al., 2011), especially during the early portion of the record. Many of the large decreases in CE are located in these same regions~~. The BE LIM displays a leading mode more similar to the CGCM-based LIMs, but still has skill problems over large ocean areas. One reason the skill of the~~ , which leads us to speculate that the features of the 20CR LIM may be influenced by artifacts in the 20CR dataset. Another consideration for the lower performance of both instrumental era LIMs ~~may be different~~ is that they are based on shorter records that coincide with variability related to anthropogenic forcing. The BE LIM displays a primary EOF more similar to the CGCM-based LIMs, but still has skill problems over large ocean areas. Separating the global warming trend from the LIM by means of linear detrending is bound to leave residual signals that affect LIM forecast modes. A LIM based on a shorter record may not have enough of a sample to properly characterize representative modes of variability over a longer time span. While the BE LIM is based on ~~only 60~~ 65 years of data, it produces much less spatial skill degradation than the 150 years of 20CR data. This suggests there may be ~~inconsistencies in variability caused by observational coverage and the data~~ confounding factors influencing the skill in the 20CR experiment between the length of record, linear detrending, and the assimilation method used ~~in creating~~ to create the 20CR ~~dataset~~data.

5 Conclusions

We have outlined and tested a new method for performing online paleoclimate data assimilation (PDA) for climate field reconstructions (CFRs) using linear inverse models (LIMs). We tested four different LIMs empirically derived from surface temperature data from the following data sets: Berkeley Earth (BE), the 20th Century Reanalysis (20CR), and two last-millennium climate simulations (CCSM4 and MPI) from the Coupled Model Intercomparison Project phase 5 (CMIP5). We also performed a persistence forecast experiment for comparison. In general, we find that LIM-enabled online assimilation improves upon the offline results for both the global average and spatial field of 2m air temperature.

~~With respect to GMT verification,~~ Broadly speaking, the LIM experiments show ~~large improvements~~ good ability to reconstruct many aspects of the GISTEMP GMT data including interannual and low freqency variability. The largest skill improvements occur for skill metrics calculated on the detrended GMT, which suggests the LIM forecasts are adding useful information at interannual timescales. The coefficient of efficiency (CE) ~~values~~ for detrended metrics show an average increase around 57%, while correlations increase around 4%. The continuous ranked probability score (CRPS) metrics increase by an average of 15% across all LIM experiments. Skill metrics tend to maximize for blending coefficients with a higher weighting ~~of flow-dependent~~ on LIM forecast information ($0.7 < a < 0.95$). Spatial skill reveals that the addition of LIM forecasting provides spatial information in regions where the offline method performs poorly— including North Hemisphere land areas, and the North Atlantic to Barents Sea corridor. Large skill improvements seen in the North Atlantic through Barents Sea region are primarily a result of better constraining the temperature variance at these locations. The two LIMs calibrated on instrumental era data (20CR and BE) display large regions over the ocean where the skill degrades compared to the offline case. Even with the large areas of improvement, the 20CR LIM decreases the area-weighted average CE (-0.18), and the BE LIM area-weighted average breaks even. In contrast, the two CGCM-based LIM experiments show area-weighted average CE increases of 0.09 (CCSM4) and

0.06 (MPI), respectively. When considering both GMT and spatial skill results, the CGCM-based LIMs have the best overall performance with the CCSM4 LIM slightly outperforming the MPI LIM. The persistence forecast fails to improve the more stringent GMT skill metrics (CE and CRPS) as well as general spatial skill, but does well in GMT timeseries correlation. Subsequently, this suggests that the improvements of online reconstructions when using a LIM are due to forecast information and not simply the addition of temporal persistence.

Though we are reconstructing instrumental era surface temperatures, it is an interesting result that CGCM-calibrated LIMs based on last millennium (850-1850) simulations have the best overall performance. This could mean that having long-running samples of variability that do not contend with ~~major sources of variability related to~~ influences from anthropogenic forcing are beneficial for reconstruction purposes. The LIMs used in these experiments were all calibrated on data with the least-squares linear fit trend removed. In order for LIMs based on observational data sources to achieve similar results, it may be necessary to employ a more sophisticated method of filtering out the global warming signal. However, one benefit of using CGCM-based LIMs is that it enables forecasts for a much larger set of climate-related quantities than are available from observations alone.

In this work, we have shown that we can improve both GMT and spatial field skill over the offline EnKF PDA method through the inclusion of a simple forecast model. A previous comparison of offline and online PDA using a CGCM as a forecast model found no discernible difference in reconstruction skill (Matsikaris et al., 2015), and earlier studies of the EnKF PDA method forewent the usage of forward models citing insufficient model skill to justify the expense (Bhend et al., 2012; Steiger et al., 2014). Our results show that an online method can increase reconstruction fidelity, and more importantly that it can be done using an empirical forecasting method that is nearly as computationally efficient as the offline approach. As such, this method provides a useful foundation for further investigation of incorporating dynamical constraints of a forecast model into climate field reconstructions.

Appendix A: Online assimilation algorithm

This section details the data assimilation equations used to perform paleoclimate field reconstructions, and the algorithm steps for a single realization of an online climate reconstruction.

A1 Ensemble square root filter (EnSRF)

The EnSRF approach (Whitaker and Hamill, 2002) uses an ensemble sampling approach to solve the Kalman filter equations by separating the ensemble mean ($\bar{\mathbf{z}}_b$) and ensemble perturbations about that mean ($\mathbf{z}'_b = \mathbf{z}_b - \bar{\mathbf{z}}_b$). Note that $\mathbf{z}_b$ represents an augmented state vector, $\mathbf{z}_b = \begin{bmatrix} \mathbf{x}_b \\ \mathbf{y}_e \end{bmatrix}$, combining the prior state ($\mathbf{x}_b$) and the estimated observations ($\mathbf{y}_e$). The EnSRF method allows for the serial assimilation of proxy observations using the equations

$$\bar{\mathbf{z}}_a = \bar{\mathbf{z}}_b + \mathbf{K}[y_i - \bar{\mathbf{y}}_{ei}] \tag{A1}$$

$$\mathbf{z}'_a = \mathbf{z}'_b + \tilde{\mathbf{K}}\mathbf{y}'_{ei} \tag{A2}$$

for each proxy $i = 1, ..., p$. The mean state $\bar{\mathbf{z}}_b$ is an $m \times 1$ column vector, the ensemble perturbations $\mathbf{z}'_b$ is an $m \times n$ matrix, the mean estimated observation for proxy i, $\bar{\mathbf{y}}_{ei}$, is a scalar value, and perturbations about the mean estimated observation $\mathbf{y}'_{ei}$ is a $1 \times n$ row vector. Note that when observations are serially assimilated, the Kalman gain (Eq. 2) ~~is simplified~~ simplifies to

$$\mathbf{K} = \frac{cov(\mathbf{z}'_b, \mathbf{y}'_{ei})}{var(\mathbf{y}'_{ei}) + \sigma_i^2} \tag{A3}$$

where σ_i^2 is the observational error for proxy y_i and the denominator is now a scalar value . The perturbation update equation $\tilde{\mathbf{K}}$ is given by

$$\tilde{\mathbf{K}} = \left[1 + \sqrt{\frac{\sigma_i^2}{var(\mathbf{y}'_{ei}) + \sigma_i^2}}\right]^{-1} \mathbf{K} \tag{A4}$$

Finally, to adapt the hybrid assimilation scheme into the EnSRF method, we incorporate the data source blending as shown in Eq. (7). At this point, the blended state ($\hat{\mathbf{z}}_b^f$) contains both flow dependent and static ~~(climatological)~~ information. After incorporation, the Kalman gain (Eq. A3) becomes

$$\hat{\mathbf{K}} = \frac{(a)cov(\hat{\mathbf{z}}_b^{f\prime}, \hat{\mathbf{y}}_{ei}^{f\prime}) + (1-a)cov(\mathbf{z}_b^{s\prime}, \mathbf{y}_{ei}^{s\prime})}{(a)var(\hat{\mathbf{y}}_{ei}^{f\prime}) + (1-a)var(\mathbf{y}_{ei}^{s\prime}) + \sigma_i^2}, \tag{A5}$$

the perturbation Kalman gain (Eq. A4) becomes

$$\tilde{\hat{\mathbf{K}}} = \left[1 + \sqrt{\frac{\sigma_i^2}{(a)var(\hat{\mathbf{y}}_{ei}^{f\prime}) + (1-a)\mathbf{y}_{ei}^{s\prime} + \sigma_i^2}}\right]^{-1} \mathbf{K}, \tag{A6}$$

and the mean and perturbation updates from Eq. (A1) and Eq. (A2) become

$$\bar{\mathbf{z}}_a = \bar{\tilde{\mathbf{z}}}_b^f + \hat{\mathbf{K}}[y_i - \bar{\hat{\mathbf{y}}}_{ei}^f] \tag{A7}$$

$$\mathbf{z}'_a = \hat{\mathbf{z}}_b^{f\prime} + \tilde{\hat{\mathbf{K}}}\hat{\mathbf{y}}_{ei}^{f\prime}. \tag{A8}$$

A2 Assimilation algorithm

1. Choose a static ensemble prior ($\mathbf{x}_b^s$) of n members, and group of p proxies ($\mathbf{y}$) to assimilate.

2. Calibrate observation models for each proxy record by applying a univariate linear fit against co-located instrumental data.

3. Create an estimated observation ensemble ($\mathbf{y}_e^s$) for each proxy record using their corresponding observation model and augment the prior ensemble to form the static state ensemble, $\mathbf{z}_b^s = \begin{bmatrix} \mathbf{x}_b^s \\ \mathbf{y}_e^s \end{bmatrix}$. This an $(m+p) \times n$ state vector that will be updated during assimilation.

4. For each reconstruction year:

   (a) If not the first reconstruction year, then reset $\mathbf{z}_b^s$ back to the original static prior and form a blended prior ($\hat{\mathbf{z}}_b^f$) using Eq. (6).

   (b) For each proxy y_i from $\mathbf{y} = [y_1, y_2, ..., y_p]$:

      i. If the current proxy has no observations for the current year, then skip to the next proxy

      ii. Else, select the matching estimated observations ($\hat{\mathbf{y}}_{ei}^f$, $\mathbf{y}_{ei}^s$) corresponding to the current proxy y_i from the augmented states ($\hat{\mathbf{z}}_b^f$, $\mathbf{z}_b^s$).

      iii. Calculate the mean and perturbation of the estimated observations and augmented state vectors for use in the serial ensemble square root filter method.

      iv. Use the perturbations of $\hat{\mathbf{z}}_b^f$, $\mathbf{z}_b^s$, $\hat{\mathbf{y}}_{ei}^f$, $\mathbf{y}_{ei}^s$ to form a blended Kalman gain ~~terms~~ as shown in Eq. (A5) and Eq. (A6).

      v. Update the mean and perturbations of $\hat{\mathbf{z}}_b^f$ and $\mathbf{z}_b^s$ by using the blended gain matrices from ~~the previous step~~ step iv in Eq. (A7) and Eq. (A8).

      vi. Reassemble $\mathbf{z}_a$ and $\mathbf{z}_a^s$ by adding the ensemble mean back into the perturbations. These will be used for the next proxy assimilated as $\hat{\mathbf{z}}_b^f$ and $\mathbf{z}_b^s$.

   (c) After all proxies have been assimilated, extract the climate field $\mathbf{x}_a$ from the augmented analysis state $\mathbf{z}_a$.

   (d) Perform a LIM forecast on $\mathbf{x}_a$ using Eq. (3) resulting in $\mathbf{x}_b^f$.

   (e) Recalculate the estimated observations $\mathbf{y}_e^f$ from $\mathbf{x}_b^f$ and augment the state to form $\mathbf{z}_b^f$.

   (f) Return to step 4.(a).

5. ~~After all years complete, we have our reconstruction of climate states $\mathbf{x}_a$ from all years.~~

Appendix B: LIM calibration

The following steps are performed to empirically derive a LIM from a given data source. The steps detail how we find the mapping term $\mathbf{G}_1$ shown in Eq. (5).

1. If the calibration data contains seasonal signals (e.g., monthly data), then they are removed by smoothing data with a 1-year running mean.

2. The data is converted into anomaly format by removing the climatolical mean for each individual month.

3. The resulting anomaly is detrended. This removes a large degree of the skill found by Newman (2013) when using instrumental data, but we are focused on forecasting modes of interannual variability, not the secular warming trend.

4. The detrended anomaly data is projected into EOF space where the leading 8 modes of variability are retained. The number of modes retained encompasses those with e-folding times (decorrelation time scales) near 1-year or greater based on analysis of $\mathbf{G}_1$ (as calulated in the following step) in these experiments. The near-1-year threshold was chosen because forecast modes with e-folding times much less than 1-year have a small impact on annual forecasts due to the relatively quick signal decay.

5. Finally, we determine $\mathbf{G}_1$ based on the lag-covariance statistics of the calibration data. Specifically, we solve $\mathbf{C}(1) = \mathbf{G}_1\mathbf{C}(0)$ is solved for $\mathbf{G}_1$ $\mathbf{C}(1) = \left\langle \mathbf{x}(t+1)\mathbf{x}(t)^T \right\rangle$.

Acknowledgements. This paper represents a portion of the first-author's Master's Thesis at the University of Washington. This research has been supported by awards from the National Science Foundation (awards AGS-1304263 and AGS-1602223), and the National Oceanic and Atmospheric Administration (NA14OAR4310176). ~~This material is also based upon work~~ The first author has also been supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1256082. The authors would like to thank Matthew Newman for his input on constructing linear inverse models, and Robert Tardif for his large contributions and many thoughtful discussions related to software development for this project.

References

Annan, J. D. and Hargreaves, J. C.: Identification of climatic state with limited proxy data, Climate of the Past, 8, 1141–1151, doi:10.5194/cp-8-1141-2012, 2012.

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

Branstator, G., Teng, H., Meehl, G. A., Kimoto, M., Knight, J. R., Latif, M., and Rosati, A.: Systematic estimates of initial-value decadal predictability for six AOGCMs, Journal of Climate, 25, 1827–1846, doi:10.1175/JCLI-D-11-00227.1, 2012.

Compo, G. P., Whitaker, J. S., Sardeshmukh, P. D., Matsui, N., Allan, R. J., Yin, X., Gleason, B. E., Vose, R. S., Rutledge, G., Bessemoulin, P., Brönnimann, S., Brunet, M., Crouthamel, R. I., Grant, A. N., Groisman, P. Y., Jones, P. D., Kruk, M. C., Kruger, A. C., Marshall, G. J., Maugeri, M., Mok, H. Y., Nordli, Ø., Ross, T. F., Trigo, R. M., Wang, X. L., Woodruff, S. D., and Worley, S. J.: The Twentieth Century Reanalysis Project, Quarterly Journal of the Royal Meteorological Society, 137, 1–28, doi:10.1002/qj.776, http://doi.wiley.com/10.1002/qj.776, 2011.

Crespin, E., Goosse, H., Fichefet, T., and Mann, M. E.: The 15th century Arctic warming in coupled model simulations with data assimilation, Climate of the Past, 5, 389–401, doi:10.5194/cp-5-389-2009, 2009.

Dirren, S. and Hakim, G. J.: Toward the assimilation of time-averaged observations, Geophysical Research Letters, 32, 1–5, doi:10.1029/2004GL021444, 2005.

Gneiting, T. and Raftery, A. E.: Strictly Proper Scoring Rules, Prediction, and Estimation, Journal of the American Statistical Association, 102, 359–378, doi:10.1198/016214506000001437, 2007.

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, doi:10.1007/s00382-006-0128-6, 2006.

Goosse, H., Crespin, E., de Montety, A., Mann, M. E., 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, 115, 1–17, doi: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, doi:10.1002/2016JD024751, http://doi.wiley.com/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, doi:10.1175/1520-0493(2000)128<2905:AHEKFV>2.0.CO;2, 2000.

Hansen, J., Ruedy, R., Sato, M., and Lo, K.: Global surface temperature change, Rev. Geophys., 48, RG4004, doi:10.1029/2010RG000345.1.INTRODUCTION, http://dx.doi.org/10.1029/2010RG000345, 2010.

Hawkins, E. and Sutton, R.: Decadal predictability of the Atlantic Ocean in a coupled GCM: Forecast skill and optimal perturbations using linear inverse modeling, Journal of Climate, 22, 3960–3978, doi:10.1175/2009JCLI2720.1, 2009.

Huddart, B., Subramanian, A., Zanna, L., and Palmer, T.: Seasonal and decadal forecasts of Atlantic Sea surface temperatures using a linear inverse model, Climate Dynamics, pp. 1–13, doi:10.1007/s00382-016-3375-1, http://link.springer.com/10.1007/s00382-016-3375-1, 2016.

Huntley, H. S. and Hakim, G. J.: Assimilation of time-averaged observations in a quasi-geostrophic atmospheric jet model, Climate Dynamics, 35, 995–1009, doi:10.1007/s00382-009-0714-5, http://link.springer.com/10.1007/s00382-009-0714-5, 2010.

Kalnay, E.: Atmospheric modeling, data assimilation, and predictability, vol. 54, doi:10.1256/003590003606835l1, 2003.

Landrum, L., Otto-Bliesner, B. L., Wahl, E. R., Conley, A., Lawrence, P. J., Rosenbloom, N., and Teng, H.: Last Millennium Climate and Its Variability in CCSM4, Journal of Climate, 26, 1085–1111, doi:10.1175/JCLI-D-11-00326.1, http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-11-00326.1, 2013.

Mann, M. E., Bradley, R. S., and Hughes, M. K.: Global-scale temperature patterns and climate forcing over the past six centuries, Nature, 392, 779–787, 1998.

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, doi:10.1126/science.1177303, http://www.sciencemag.org/cgi/doi/10.1126/science.1177303, 2009.

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

Murphy, J. M.: The impact of ensemble forecasts on predictability, Quarterly Journal of the Royal Meteorological Society, 114, 463–493, doi:10.1002/qj.49711448010, http://doi.wiley.com/10.1002/qj.49711448010, 1988.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I: A discussion of principles, Journal of Hydrology, 10, 282–290, doi:10.1016/0022-1694(70)90255-6, 1970.

Newman, M.: An Empirical Benchmark for Decadal Forecasts of Global Surface Temperature Anomalies, Journal of Climate, 26, 5260–5269, doi:10.1175/JCLI-D-12-00590.1, http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-12-00590.1, 2013.

Newman, M., Sardeshmukh, P. D., and Penland, C.: How important is air-sea coupling in ENSO and MJO evolution?, Journal of Climate, 22, 2958–2977, doi:10.1175/2008JCLI2659.1, http://journals.ametsoc.org/doi/abs/10.1175/2008JCLI2659.1, 2009.

PAGES 2k Consortium: Continental-scale temperature variability during the past two millennia, Nature Geoscience, 6, 339–346, doi:10.1038/NGEO1797, www.nature.com/naturegeoscience, 2013.

Pendergrass, A. G., Hakim, G. J., Battisti, D. S., and Roe, G.: Coupled Air–Mixed Layer Temperature Predictability for Climate Reconstruction, Journal of Climate, 25, 459–472, doi:10.1175/2011JCLI4094.1, http://journals.ametsoc.org/doi/abs/10.1175/2011JCLI4094.1, 2012.

Penland, C. and Sardeshmukh, P.: The optimal growth of tropical sea surface temperature anomalies, Journal of climate, http://journals.ametsoc.org/doi/abs/10.1175/1520-0442(1995)008{\T1\textless}1999:TOGOTS{\T1\textgreater}2.0.CO;2, 1995.

Rohde, R., Muller, R., Jacobsen, R., Perlmutter, S., Rosenfeld, A., Wurtele, J., Curry, J., Wickham, C., and Mosher, S.: Berkeley earth temperature averaging process, Geoinfor Geostat: An Overview, 1, 1–13, doi:10.4172/2327-4581.1000103, 2013.

Smerdon, J. E., Kaplan, A., Chang, D., and Evans, M. N.: A Pseudoproxy Evaluation of the CCA and RegEM Methods for Reconstructing Climate Fields of the Last Millennium*, Journal of Climate, 24, 1284–1309, doi:10.1175/2010JCLI4110.1, http://journals.ametsoc.org/doi/abs/10.1175/2010JCLI4110.1, 2011a.

Smerdon, J. E., Kaplan, A., Zorita, E., González-Rouco, J. F., and Evans, M. N.: Spatial performance of four climate field reconstruction methods targeting the Common Era, Geophysical Research Letters, 38, n/a–n/a, doi:10.1029/2011GL047372, http://dx.doi.org/10.1029/2011GL047372, l11705, 2011b.

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, doi:10.1175/JCLI-D-12-00693.1, http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-12-00693.1, 2014.

Taylor, K. E., Stouffer, R. J., and Meehl, G. A.: An overview of CMIP5 and the experiment design, Bulletin of the American Meteorological Society, 93, 485–498, doi:10.1175/BAMS-D-11-00094.1, 2012.

Tipton, J., Hooten, M., Pederson, N., Tingley, M., and Bishop, D.: Reconstruction of late Holocene climate based on tree growth and mechanistic hierarchical models, Environmetrics, 27, 42–54, doi:10.1002/env.2368, http://doi.wiley.com/10.1002/env.2368, 2016.

Wahl, E. R. and Smerdon, J. E.: Comparative performance of paleoclimate field and index reconstructions derived from climate proxies and noise-only predictors, Geophysical Research Letters, 39, n/a–n/a, doi:10.1029/2012GL051086, http://doi.wiley.com/10.1029/2012GL051086, 2012.

Wang, J., Emile-Geay, J., Guillot, D., Smerdon, J. E., and Rajaratnam, B.: Evaluating climate field reconstruction techniques using improved emulations of real-world conditions, Climate of the Past, 10, 1–19, doi:10.5194/cp-10-1-2014, 2014.

Whitaker, J. and Hamill, T.: Ensemble data assimilation without perturbed observations, Monthly Weather Review, pp. 1913–1924, http://journals.ametsoc.org/doi/abs/10.1175/1520-0493(2002)130{\T1\textless}1913:EDAWPO{\T1\textgreater}2.0.CO;2, 2002.

Woodruff, S. D., Worley, S. J., Lubker, S. J., Ji, Z., Eric Freeman, J., Berry, D. I., Brohan, P., Kent, E. C., Reynolds, R. W., Smith, S. R., et al.: ICOADS Release 2.5: extensions and enhancements to the surface marine meteorological archive, International journal of climatology, 31, 951–967, doi:10.1002/joc.2103, 2011.

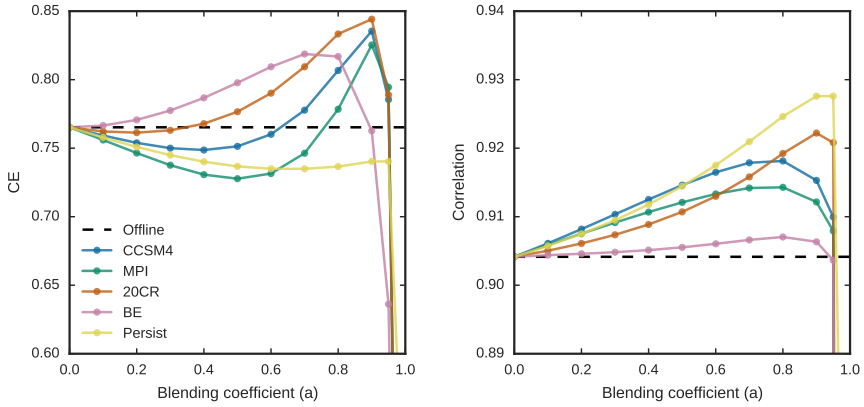

Figure 1. Comparison of global mean 2m air temperature coefficient of efficiency (CE; left) and correlation (right) metrics for different blending coefficients. The colored lines represent the different LIM calibration experiments using data from the Community Climate System Model v4 (CCSM4), NOAA 20th Century Reanalysis v2 (20CR), Max Plank Institute Earth System Model (MPI), Berkeley Earth Surface Temperatures (BE); or the persistence forecast case (Persist). The offline benchmark is depicted as the horizontal dashed black line.

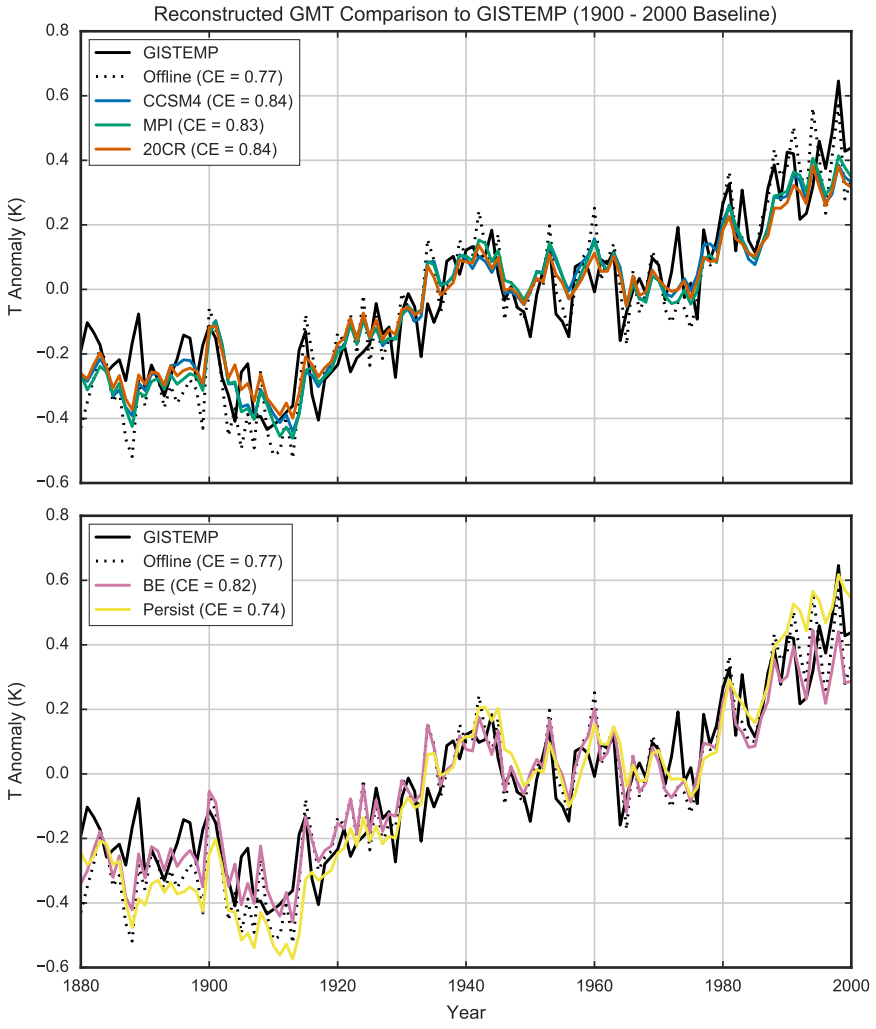

Figure 2. Reconstructed global mean 2m air temperature compared to GISTEMP (solid black) and the offline experiment (dotted black) for each online experiment. The GMT plotted for each forecasting experiment is from the blending coefficient that achieves the highest GMT CE skill (CCSM4: $a = 0.9$, MPI: $a = 0.9$, 20CR: $a = 0.9$, BE: $a = 0.7$) except for the persistence case where $a = 0.9$ was used. The top row shows the three forecasting experiments with the highest GMT CE score, while the bottom row shows the other two experiments.

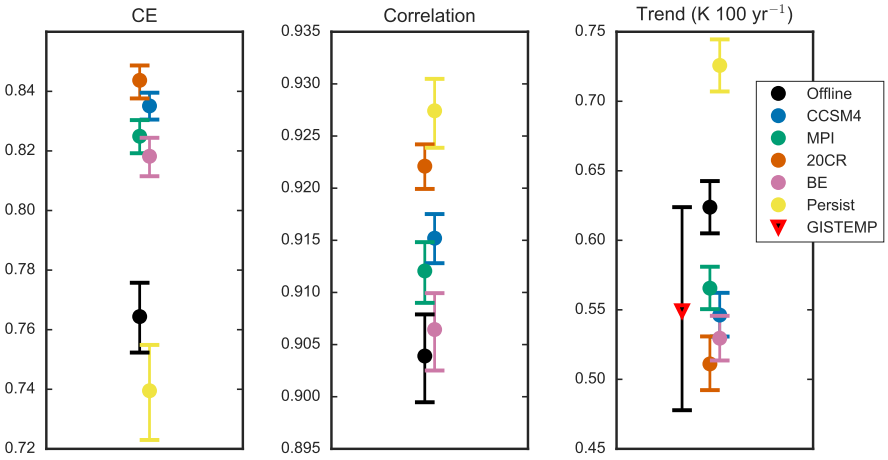

Figure 3. Bootstrap uncertainty estimates (95% confidence interval) for CE scores (left), correlation (middle), and GMT trends (right). The blending coefficient that achieves the highest GMT CE skill is shown for each experiment (CCSM4: $a = 0.9$, MPI: $a = 0.9$, 20CR: $a = 0.9$, BE: $a = 0.7$) except for the persistence case where $a = 0.9$ was used.

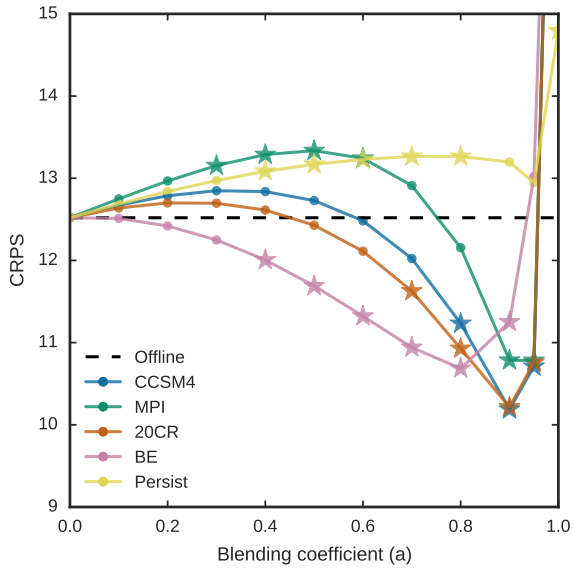

Figure 4. Comparison of global mean 2m air temperature continuous ranked probability score (CRPS) for different blending coefficients. The colored lines represent the different forecasting experiments, while the offline benchmark is depicted as the horizontal dashed black line. Starred points indicate a statistically significant (95% confidence) difference between the offline benchmark and online experiment.

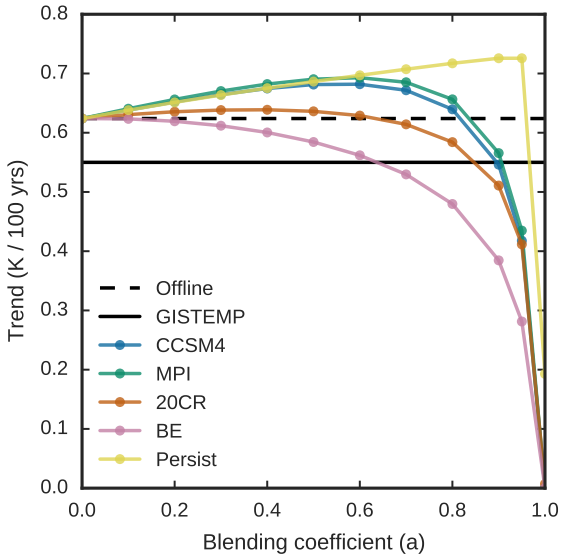

Figure 5. Calculated trends from a least squares fit against the reconstructed global mean 2m air temperature (1880 - 2000). Colored lines depict the calculated trends for each LIM experiment across a range of blending coefficients, while the black lines represent the benchmark trends calculated from the offline reconstruction (dashed) and GISTEMP data (solid).

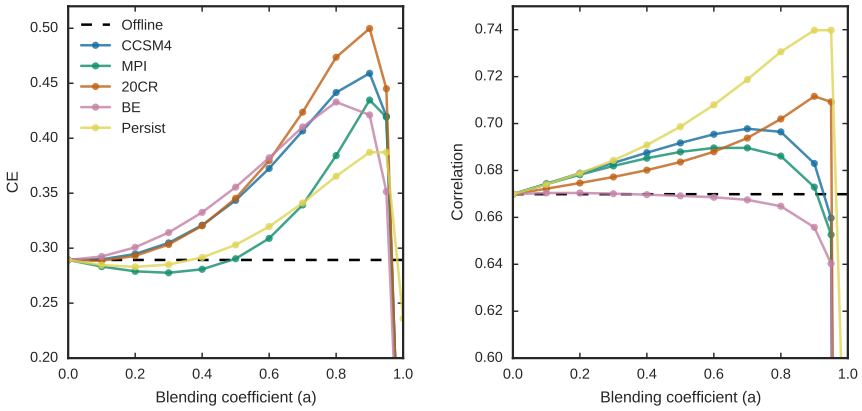

Figure 6. ~~Same as in Figure 1 but with the linear trend removed from the~~ Comparison of detrended global mean 2m air temperature ~~data before calculation of skill~~ CE (left) and correlation (right) metrics across different blending coefficients for all experiments. The colored lines represent the different forecasting experiments, while the offline benchmark is depicted as the horizontal dashed black line.

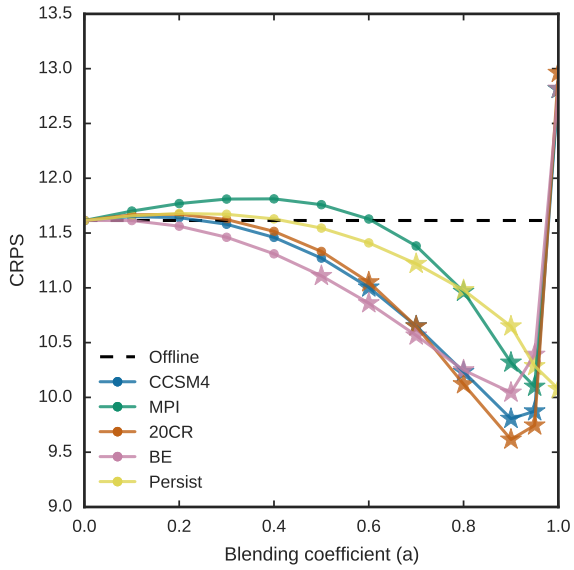

Figure 7. Comparison of detrended global mean 2m air temperature continuous ranked probability score (CRPS) for different blending coefficients. The colored lines represent the different forecasting experiments, while the offline benchmark is depicted as the horizontal dashed black line. Starred points indicate statistical significance (95% confidence) between the offline benchmark and online experiment.

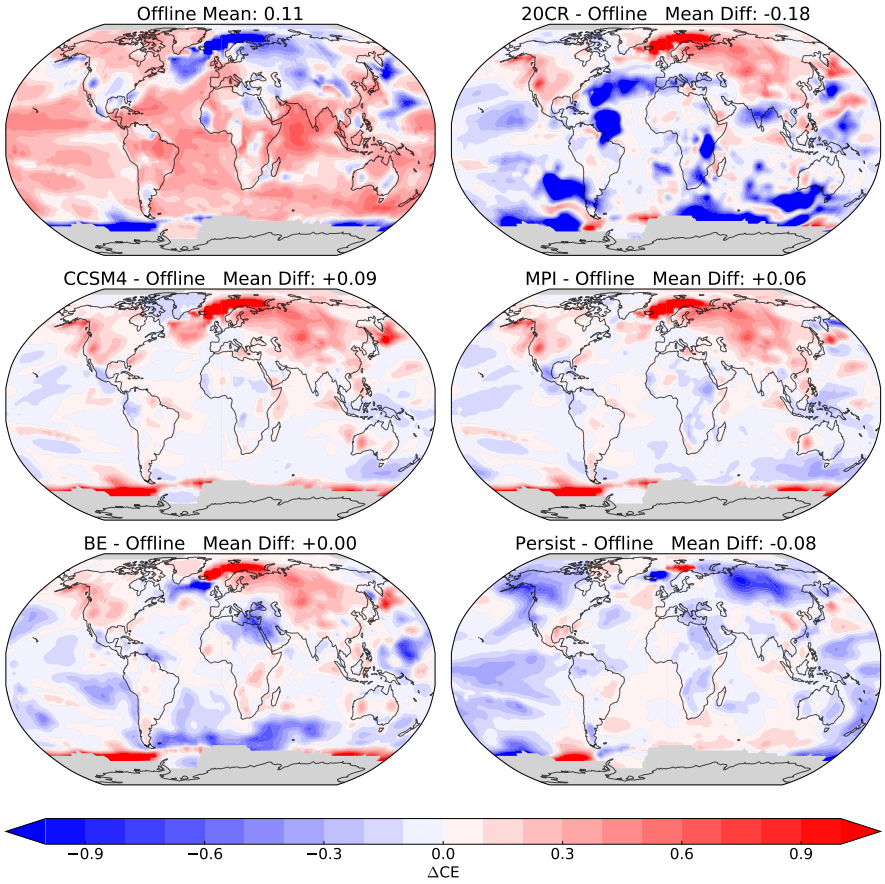

Figure 8. Spatial maps displaying the difference in coefficient of efficiency (CE) from the offline case. Difference maps are displayed for each forecasting experiment using the blending coefficient that achieves the highest full GMT CE skill (CCSM4: $a = 0.9$, MPI: $a = 0.9$, 20CR: $a = 0.9$, BE: $a = 0.7$) except for the persistence case where $a = 0.9$ was used. ~~(See Table 1 for the list of corresponding blending coefficients.)~~ The reference CE of the offline case is shown in the upper left, and uses the same color scale as the difference maps. Area-weighted global average differences are given in the title of each panel. All global mean differences except in the BE LIM case are significantly different than the offline benchmark with 95% confidence.

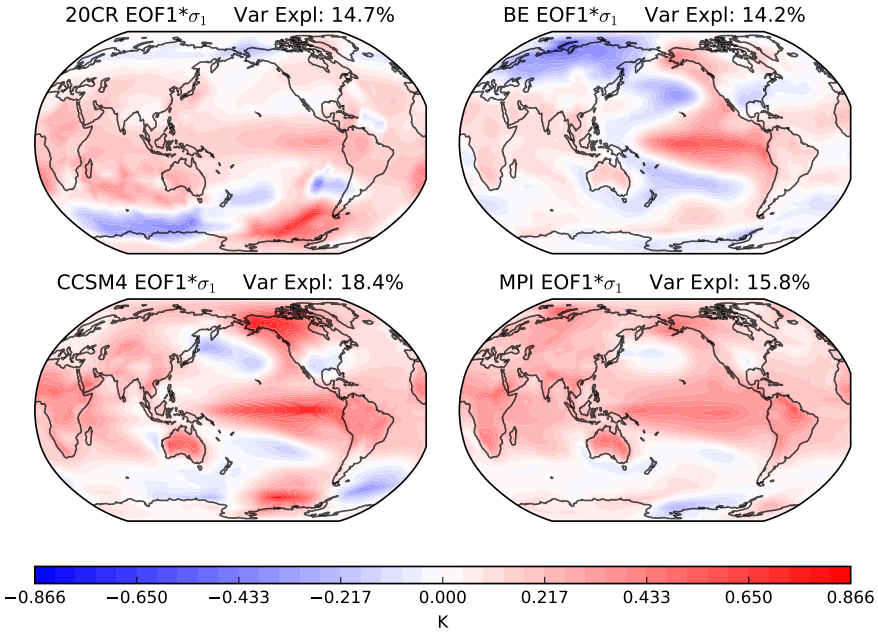

Figure 9. Leading empirical orthogonal function (EOF) from the basis for each LIM calibration. The total fraction of the variance explained is given in the title of each panel. All EOFs have been multiplied by their corresponding singular value.

Table 1. Best value of the coefficient of efficiency (CE), correlation (r), and continuous ranked probability score (CRPS) ~~verification~~ validation metrics for global mean 2m air temperature. For each experiment, best values are given with the corresponding blending coefficients (a) that achieved it and the percentage change compared to the offline case. A (*) indicates which experiment achieved the best performance in a given metric. Offline ~~verification~~ validation metrics are given for reference.

Full GMT	**Max CE**	**%ΔCE**	**CE a-value**	**Max r**	**%Δr**	**r a-value**	**Min CRPS**	**%ΔCRPS**	**CRPS a-value**
Offline	0.77			0.90			12.5		
Persist	0.77	0	0.0	*0.93	3	0.9	12.5	0	0.0
BE	0.82	7	0.7	0.91	1	0.8	10.7	-14	0.8
CCSM4	*0.84	9	0.9	0.92	2	0.8	*10.2	-18	0.9
20CR	*0.84	9	0.9	0.92	2	0.9	*10.2	-18	0.9
MPI	0.83	8	0.9	0.91	1	0.8	10.8	-14	0.95

Table 2. Best value of the CE, correlation, and CRPS ~~verification~~ validation metrics for detrended global mean 2m air temperature. For each experiment, best values are given with the corresponding blending coefficients (a) that achieved it and the percentage change compared to the offline case. A (*) indicates which experiment achieved the best performance in a given metric. Offline ~~verification~~ validation metrics are given for reference.

Detrended GMT	Max CE	%ΔCE	CE a-value	Max r	%Δr	r a-value	Min CRPS	%ΔCRPS	CRPS a-value
Offline	0.29			0.67			11.6		
Persist	0.39	35	0.9	*0.74	11	0.9	10.1	-13	1.0
BE	0.43	48	0.8	0.67	0	0.1	10.0	-14	0.9
CCSM4	0.46	59	0.9	0.70	5	0.7	9.8	-16	0.9
20CR	*0.50	72	0.9	0.71	6	0.9	*9.6	-17	0.9
MPI	0.43	48	0.9	0.69	3	0.7	10.1	-13	0.95