Clim. Past Discuss.,
doi:10.5194/cp-2016-129-AC2, 2017

Climate of the Past Discussions

Interactive comment on "Reconstructing past climate by using proxy data and a linear climate model" *by* Walter A. Perkins and Gregory J. Hakim

Walter A. Perkins and Gregory J. Hakim

wperkins@uw.edu

Received and published: 2 March 2017

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

1 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.

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 with 1200 100-member reconstructions (100 iterations for each of the 12 blending coefficients), 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.

2 Specific Comments:

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

2.1 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.

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.

2.2 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 for forecasting, 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.

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.

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.

2.3 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 (with a maximum CE of 0.82), 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).

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.

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.

2.4 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.

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.

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.

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.

"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 (pp10 line 4-5) to make this point clearer.

3 References:

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

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

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

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

Interactive comment on Clim. Past Discuss., doi:10.5194/cp-2016-129, 2016.